COMPREHENSIVE NEURAL NETWORK FORECASTING SYSTEM FOR GROUND LEVEL OZONE IN MULTIPLE REGIONS

by

GAUTAM RAGHAVENDRA EAPI

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

December 2015

Acknowledgements

I am thankful to my Professors' Dr. Melanie L. Sattler and Dr. Michael T. Manry for their support and valuable guidance throughout my research. To the best of my knowledge, Dr. Sattler is very kind, approachable, and a helpful professor in the Civil Engineering Department and Dr. Manry is known for high quality of research work. The amount of research hours Dr. Manry spends with the students in the IPNNL lab and the long hours he spends correcting and revising the dissertation documents to ensure that the output quality is good, is an example of this. He is my most valued professor at UTA and I owe him a lot.

I am thankful to my committee members Dr. Ghandehari, Dr. Victoria Chen and Dr. Choi for their suggestions in my research work. I had the opportunity to work with Dr. Qasim, Dr. Crosby, Dr. Kruzic and the experience I gained is invaluable. Dr.Qasim, Dr.Ghandehari, Dr. Anand Puppala and Dr. Abolmaali have been a constant source of motivation and encouragement throughout my PhD. I am also thankful to Dr. Fuqiang, Joshua Been, Dr. Kim, and Richard Gaines: all good instructors.

I am thankful to my friends Rohit Rawat, Kanishka Tyagi, Bito Irie, Son Nguyen, Kunal, Jignesh, Aditi, Auddy, Jeshwanth, Parastoo, Yilong and Jugal in the IPNNL lab (Electrical Engineering Department). I am indebted to all of them for a lifetime. Their positive words helped me continue with this research topic that was above my level of understanding during the initial years. Rohit has always been a great support for me all these years and I am blessed to have a friend like him.

I am thankful to all my friends from Civil engineering department. I am grateful to Srinivas Chittoori, Richa and Neelesh for their guidance and moral support during my initial years at UTA. My civil engineering friends Nan, Reza, Zak, Ishtiaq, Zahangir, Prince, Su, Said, Hesham, Sulak, Ketwalee, Roja, Annaprabha, Sahithi, Rukmini, Pinku, Meenakshi, Elahe, Jiaqi, Mahsa, Ujwal, Daniel, Shang, Ariel, Wasui, Fari, Amir, Anna, Binu, Shikshita, Arezoo, Niloofar, Lince, Thiru, Rathan, Maryam, Arezoo, Srinivas Prabakar, Arpita, Aditya, Parthen, Angelique, Shankar, Manjeera, Laxman, Sujit, Aravind, Vijay, Rajni, Aashish, Shilpa, Eshwari, Tejo, Asher, Ujwal, Naga, Madanu, Madhu, Shammi, Vennila, Rakesh, Ranjit, Praveen, Ratna, Spoorti, Ashraf, Ramya, Kiran, Jithendra have been great support. To have friends like Kartik Siddhabathula, Harsha Meka, Chakri, Harsha Raju, Sai Krishna, Avinash, Subbu, Mandar, Snehal, Naveen, Charan, Vinod, Sid, Raja, Sayan, Thejjesh, Shravya, Vikram, Varun, Shriram, Rohit Toom, Balmu, Dileep, Vinay, Venkat, Kartik, Prasad, Harsha, Jitesh, Nema, Soma, Rohit Reddy, Sairam, Srinivas, Teju, Subash, Aditya, Sharmista, Om, Deepak, Pradeep, Raghavendra, Rohit Sawkar, Chaitu, Shreya, Singi, Praveen, Srinivas, Balla, Naval, Sreemanth, and Angela is a gift. I am thankful to my childhood friends and neighbors, Prasad, Madhu, Shyam, Nani, Balu, Rohit, Rabi akka, Jyothi akka, Vijay, and Bhushan. I might have missed a few names but I love you all. I am thankful to all my friends from High School, Bachelors, and Masters.

I am thankful to the Civil Engineering Department, and the SEL staff. Sara, Ava, Lewis, Ginny, Tina, Lynda, Ladan, Kierra, Kristina and Jamie have always been student friendly.

I loved being a TA for the classes Spring 2013, Summer 2013, Fall 2013, Spring 2014, Summer 2014, Fall 2014, Spring 2015, Summer 2015, and Fall 2015. Thank you guys for all your reviews. It was only because of you that I am living a contented life since 2013.

The support I got from my family is priceless. I am blessed to have cousins, Sughani anna, Manga akka, Sujatha, Udai, Ravi, Pedanna, Bala anna, Suma odina,

Manju, Sailu, Sarada akka, Pinku, Chanti, Prakash, Sarada akka, Srinu anna, Vivek, Venu, Vijji, Madhavi, Babloo, Bhargavi, Tinku, Munna, Bannu, Likitha, Sai, Santosh, and Sindhu. My nephews: Arnav, Tanish, Vivaan, Tanav, my nieces: Aishwarya, Manasvini, and Indrani have brought lots of happiness in the family. My brother, Bharat and his wife Latha have been very supportive all these years. Bharat is a gift my mother gave me. My father and mother have spent all their earnings towards our education and I think my doctoral degree is what I can give them for all their love and sacrifice.

Finally, the countless blessings of my grandparents, Vasu pedananna, my cousins Krishna anna, Asha, and Arjun from a better world, helped me from time to time. Thanks for being a part of my life.

November 12, 2015

Dedicated to

To my two mothers I love the most:

Seshu pedamma for retaining the utmost faith

in God despite all the losses in her personal life and

my beautiful mother, Durga Devi for her infinite and selfless love

Abstract

COMPREHENSIVE NEURAL NETWORK FORECASTING SYSTEM FOR GROUND LEVEL OZONE IN MULTIPLE REGIONS

Gautam R. Eapi, PhD

The University of Texas at Arlington, 2015

Supervising Professors: Melanie L. Sattler and Michael T. Manry

A comprehensive neural network daily maximum 8 hour-ozone forecasting model was developed based on five years of data (2010-2014) collected from 50 monitoring sites from the Dallas Fort Worth, Houston-Galveston-Brazoria, Los Angeles, San Joaquin and San Diego regions. This work represents the first neural network developed to forecast ozone in multiple regions, as well as multiple sites in the same region. Previous studies have developed separate neural network models to forecast ozone at each location.

Two stages of feature selection were applied to reduce input vector dimension and redundancy. These are Piecewise Linear Orthonormal Floating Search (PLOFS), and Karhunen - Loève Transform (KLT). Two possible approaches for organizing the data were tried. These are a tall file approach and a median file approach. Results showed better performance of the tall file approach. The Multilayer Perceptron (MLP) neural network used in this study showed better prediction performance compared to other existing MLP neural network approaches.

Table of Contents

List of Illustrations

List of Tables

Chapter 1

Introduction

Air is one of the principal components essential for the existence of life on earth and it should be clean to sustain a healthy atmosphere for present and future generations. Unfortunately, the quality of air has declined to unacceptable levels in many locations due to the activities of humans during the past few decades. After the advent of industrialization, technology, and urbanization, environmentalists, researchers and government bodies have made many efforts to curb air pollution. To deal with air pollution, the Clean Air Act (CAA) of 1970, required the United States Environmental Protection Agency (USEPA) to set up primary and secondary standards for all the six criteria pollutants to protect public health and public property, respectively.

Ozone is one of the six criteria pollutants specified by the USEPA. Ozone has several adverse health impacts that include lung infection, chest pain, and eye and throat irritation. Ozone aggravates asthma and bronchitis. Ozone also causes damage to vegetation and natural ecosystems as described by Seinfeld.[29] The current National Ambient Air Quality Standard (NAAQS) for ozone is 0.075 ppm for both primary and secondary standards effective March-2008, based on an 8–hour averaging time. There were 227 non-attainment counties and 46 non-attainment areas for ozone across the US as of January 30th, 2015, which do not meet the ozone NAAQS.[26] States that do not meet these NAAQS standards have to develop a State Implementation Plan (SIP) for that area. In its SIP, a state government specifies the implementation measures and the proper planning methods it will adopt to reduce emissions in future years.

Considering the harmful effects of ozone, forecasting of ozone is essential to inform the public well in advance about outdoor air quality. Sensitive people can be made aware of ozone episodes beforehand which should lead to fewer hospital visits. On ozone unhealthy days, people can be advised to limit driving and buying fuel before 10 a.m. and stay indoors. Apart from taking care of public concerns, forecasting ozone gives an idea, how well the emission reduction specifications are being implemented. Further guidelines to alleviate ozone levels can be formulated by making use of these forecasting results.

In the US, air quality modeling is carried out by the EPA to predict air quality in such region. Currently, photochemical grid models are used in the development of SIP.[26, 27] Photochemical grid models are deterministic models that simulate meteorological parameters (such as winds that carry pollutants, surface temperature, solar radiation, relative humidity), pollutant emissions (oxides of nitrogen, and volatile organic compounds from sources that participate in ozone formation), and chemistry (complex reactions that result in the formation of ozone). [27] The Comprehensive Air Quality Model with extensions (CAMx) is one such model that simulates air quality over many geographic scales. Urban Airshed Model Variable Grid (UAM-V) is another model that is widely used for modeling ozone episodes.

Even though these deterministic models are appropriate, they involve a great deal of computational effort because of the complexity of ozone chemistry and the meteorology. Also, deterministic models incur more costs and more processing time than other computing techniques. Statistical models are faster than deterministic models and do not consider the physical, chemical and meteorological factors involved in ozone formation.[5] In the past, considerable research has been done in air quality modeling using computing techniques such as neural networks[6,8,9,12, 23, 25, 70,77, 78, 83, 84, 87, 88, 95,110], fuzzy logic[34,35] and well-known regression methods[15, 22, 89].

Among these statistical models, the usage of neural networks is quite popular. Artificial neural networks (ANNs) have wide applications in the field of Civil Engineering.

Apart from air quality modeling, they have been used in water and wastewater treatment[41], geotechnical engineering[38, 39], transportation planning[36, 37], and rainfall runoff modeling[40], to mention a few.

Previous literature studies show that neural networks are reliable and faster alternative to deterministic models that are more laborious and time–intensive in forecasting complex nonlinear atmospheric pollutants. Using neural networks, air pollutants such as NO_X (oxides of nitrogen), particulate matter, and ground level ozone can be predicted. Some of the networks commonly used so far are the multilayer perceptron (MLP), and the radial basis function (RBF). Some of the previous ground–level ozone forecasting studies include MLPs trained using algorithms such as backpropagation (BP),[10, 15, 24] scaled conjugate gradient (SCG),[11, 12, 14, 57] Levenberg–Marquardt (LM).[8, 21] Positive results have been shown for MLPs based on principal components [16], MLPs trained using synergistically coupled Levenberg–Marquardt, a deterministic local optimization algorithm and Particle Swarm Optimization(PSO), and a stochastic global optimization algorithm showed positive results.[15, 21] RBF networks have also been used to predict ozone and oxides of nitrogen.[25, 30]

Even though the choice of modeling tool is problem specific, neural networks can be applied when there is richness in data and theory according to Rumelhart.[18] In this study, a comprehensive ozone forecasting model was developed for multiple cities/regions (Dallas-Fort Worth, Houston, Los Angeles, San Joaquin, and San Diego) using fifty ozone monitoring sites across the US using multilayer perceptron. This work represents the first neural network developed to forecast ozone in multiple regions, as well as at multiple sites in the same region. Previous studies have developed separate neural network models to forecast ozone at each location or few locations.

Chapter 2 reviews ozone chemistry, neural network technology, different types of neural networks, and previous work on air quality modeling. Chapter 3 includes data description. In Chapter 4, the example forecasting system and its problems will be described. In Chapter 5, the possible system approaches to solve the problems mentioned in Chapter 4 will be discussed. In Chapter 6, feature selection is described using subsets and transformation methods. In Chapter 7, Results and Discussion will be described. Chapter 8 includes Final Conclusions and Future Work.

Chapter 2

Literature Review

2.1 Ozone chemistry

Tropospheric ozone, or ground–level ozone or "bad" ozone, is a secondary pollutant formed as a result of the reaction between the primary pollutants namely, oxides of nitrogen (NO_X) and Volatile Organic Compounds (VOCs), in the presence of sunlight [5-9, 11-15, 21-27, 29].

$$NO_X + VOCs \xrightarrow{sunlight} O_3 \qquad (2.1)$$

Equation 2.1 is a simplified version of the actual reaction mechanism, which involves hundreds of reactions.

The formation of ozone can be explained in more detail by the Leighton mechanism 1, 26, 27, 29 that is considered as a backbone of ozone smog formation according to Seinfeld and Pandis.29 When NO2 at wavelengths less than 424nm absorbs the energy of the photon, it can be dissociated to:

$$NO_2 \xrightarrow{h\upsilon} NO + O \qquad (2.2)$$

The oxygen atom is a free radical and is highly reactive. The oxygen atom O, combines with molecular oxygen (O_2) to form ozone (O_3) in the presence of a third generic molecule (M) that absorbs the excess energy and stabilizes ozone molecule formed.

$$O + O_2 + M \rightarrow O_3 + M \qquad (2.3)$$

M is generally O_2 or N_2. Ozone formed reoxidizes to NO_2 according to

$$O_3 + NO \rightarrow NO_2 + O_2 \qquad (2.4)$$

The reaction of (2.2) is a rate limiting reaction for this basic O_3 and NO_x cycle. These three reactions eventually maintain a photo stationary steady state represented by

$$[O_3]_{PSSA} = \frac{k_1}{k_3} \times \frac{[NO_2]}{[NO]} \tag{2.5}$$

where k_1 and k_3 are reaction rate constants for reactions (2.2) and (2.4). Ozone is found to be proportional to the ratio of $[NO_2]$ to $[NO]$ in most conditions if NO_2, NO and O_3 are measured in ambient air. At sunrise and sunset k_1 is small and the above photo stationary steady state assumption does not hold well. 90 to 95% of the NO_X emissions from combustion sources will be in the form of NO and 5 to 10 % in the form NO_2. NO_2 initiates the $NO_X – O_3$ cycle and with such lower concentrations cannot produce actual O_3 concentrations observed in the field. So, the Leighton mechanism alone is not sufficient to explain the ozone formation in urban atmospheres. VOCs play a key role in converting NO to NO_2 with the help of OH^o radical and explain for the actual ozone observations in the field [26,27,29].

2.1.1 Sources and adverse effects of ground level ozone

The presence of ozone in surface air is toxic to humans and also causes damage to vegetation by oxidizing the biological tissue (Jacob, 1999). In densely populated regions, the ozone formation is rapid due to high emissions of oxides of nitrogen and VOC's, thereby making an air pollution problem. The major sources of NOx and VOCs comprise emissions from industrial facilities and electric utilities, motor vehicle exhaust, gasoline vapors and chemical solvents (USEPA). Ozone inhalation initiates health problems, such as chest pain, congestion, throat irritation and lung infection. Ozone can also worsen bronchitis, emphysema, and asthma. Children and senior citizens are more sensitive to ozone pollution (USEPA). Ozone damages vegetation and ecosystems by reducing the agricultural crop and commercial forest yields. “In the US alone, ground level ozone is responsible for an estimated loss of $500 million in reduced crop production each year” (USEPA).

2.2 Neural Network Technology

"A neural network is a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use." [3] Artificial neural networks (ANNs) are empirical, non parametric models and are data driven.

ANNs have several properties that make them powerful computational tools:
(a) Nonlinearity: This property allows ANNs to better fit complex data seen in nature, where linear fits work poorly. For example, the complex relationship between the pollutants in the atmosphere and the meteorology makes a non linear system and the use of ANNs is reasonable.
(b) Non parametric: ANNs are non parametric in nature as they do not assume or have a prior knowledge about the linear or non linear functional relationship between the variables involved.
(c) Learning and self adaptivity: ANNs learn by modifying the weights or connections between the nodes in response to changes in the surrounding environment.
(d) Generalization: The ANN's show good generalization performance when the network is properly trained. Networks with good generalization predict well when fed with unseen or new input data.

Among the many ANNs developed so far, the popular networks are the multilayer perceptron (MLP)[9,12-15,38,] the radial basis function network (RBF)[25,30,36,38,68,75], and the piecewise linear network (PLN).[20, 28, 43, 63, 65] Briefly, these networks are described below.

2.2.1 Multilayer Perceptron

MLPs are the most commonly used ANNs in the field of forecasting. They are considered to have inherent ability to approximate a smooth functional relationship between input and output data and as such they are viewed as universal approximators.[17, 19] This feature of MLPs make them suitable for problems that involve a complex, nonlinear relationship between the input and output variables and where a deterministic solution becomes a laborious task.

MLPs are trained by supervised learning where the learning rule is provided with training data that has both input and target output examples. The weights and thresholds are adjusted in such a way that the error (typically, MSE) between the actual output and the given target output is minimized.

A typical fully connected MLP structure made up of input, hidden and output layers is shown in Figure 2-1. The nodes in the input layer represent input variables, the nodes in the output layer represent output variables and the target values of these are known to the modeler. The weights connect nodes in different layers. The number of nodes in the hidden layer or layers needs to be determined. They deal with the nonlinear part of the network. The determination of the number of hidden units/layers is problem specific and is evaluated generally on trial and error basis or based on modeler's experience. Generally, the hidden units are taken in the range of 1 to 30.[9] The hidden units play a major role in finding the optimum solution for the weights that store information of nonlinear relationship. As such, the network can be over trained if the number of hidden units is chosen to be too large.

Using the same notation as in [19], the MLP can be explained as follows: Let $\{\mathbf{x}_p, \mathbf{t}_p\}$ represent the training data to be used in a MLP network where, $\mathbf{x}_p$ is the p^{th} input vector of dimension N, $\mathbf{t}_p$ represents the p^{th} desired or target output of dimension M and p represents the pattern number that takes values from 1 to N_v. In the (N+1) dimensional augmented input vector $\mathbf{x_p}$, the element $x_p(N+1)$ equals 1 in order to generate network biases. Now, $\mathbf{x_p}$ becomes $[x_p(1), x_p(2), x_p(3),\ldots, x_p(N), x_p(N+1)]^T$.

In a fully connected network, each neuron is connected to the preceding and following layers by connections (represented by arrows in Figure 2-1) with strengths termed as weights. Let $w(k, n)$, $w_{oh}(i, k)$, and $w_{oi}(i, n)$ respectively represent input weight connecting the n^{th} input to the k^{th} hidden unit, hidden weight connecting the k^{th} hidden unit's activation $O_p(k)$ to the i^{th} output $y_p(i)$, and the bypass weight connecting the n^{th} input to the i^{th} output.

Each hidden unit receives data from the input layer and the k^{th} hidden unit's net function for the p^{th} pattern can be expressed as

$$n_p(k) = \sum_{n=1}^{N+1} w(k,n) \bullet x_p(n) \qquad (2.6)$$

The activation, $O_p(k)$ of the k^{th} hidden unit for the p^{th} pattern, is usually a nonlinear transformation of the corresponding net function, such as the sigmoid defined as

$$O_p(k) = f(n_p(k)) = \frac{1}{1+\exp(-n_p(k))} \qquad (2.7)$$

The i^{th} output $y_p(i)$ of the M - dimensional output vector, $\mathbf{y}_p$ for the p^{th} pattern is

$$y_p(i) = \sum_{n=1}^{N+1} w_{oi}(i,n) \bullet x_p(n) + \sum_{k=1}^{N_h} w_{oh} \bullet O_p(k) \qquad (2.8)$$

where N_h denotes the number of hidden units. The weights are obtained by using a training algorithm. Training algorithms can be characterized as one stage, where all weights are updated simultaneously or two stage where input weights are updated separately from output weights. Examples of one stage training algorithms include backpropagation[24], conjugate gradient[11,12,14] and Levenberg Marquardt [8,21].

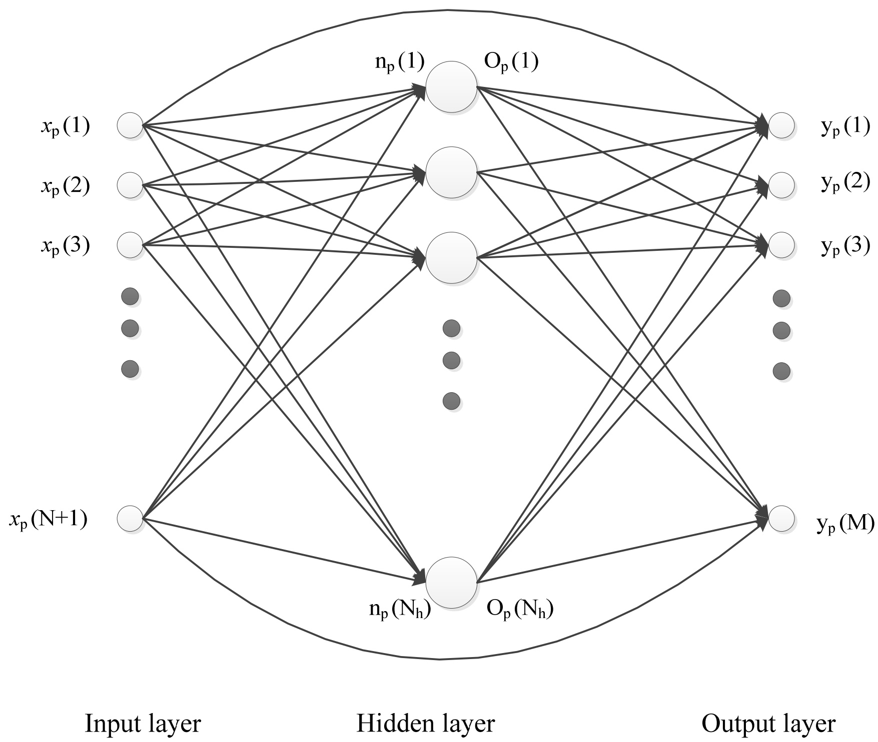


Figure 2-1 Multilayer Perceptron (MLP)[62]

Some two stage training algorithms include OWO – BP [61,62], OIG –OWO[19,62], and MOLF – OWO[19,61,62].

The neural network used in this study is Multilayer Perceptron-Hidden Weight Optimization-Multiple Optimal Learning Factors (MLP–HWOMOLF). More information about the MLP - HWOMOLF algorithm can be found in Rawat et al. (2013). [111]

The cost or error function, typically used in MLP training is the mean squared error (MSE) expressed as:

$$E = \frac{1}{N_V} \sum_{p=1}^{N_V} \sum_{i=1}^{M} [t_p(i) - y_p(i)]^2 \tag{2.9}$$

MLP models suffer from the problem of potential convergence to local minimum along with overtraining. MLP models can be over parameterized and develop memorization characteristics. These over trained models have poor generalization when tested on new unseen data.[13]

Random initialization of hidden weights or input weights is done to avoid the domination of inputs that have large standard deviations. A good practice is to train a lot of MLP models and choose the model that has the best generalization property. In MLP training, the global minimum is generally not obtainable. MLP's with two hidden layers converge faster and escape local minima during training process. [12, 14, 31]

Early stopping is one technique [8,14, 18] used to solve the overtraining problem: Data is divided into three parts, training, validation and testing sets. During the training of the MLP model, validation data is used to check the generalization property. Initially, the generalization is good but after a point the performance of generalization decreases. Training is stopped at this point. Then, performance of the network is tested with the testing data.

Bayesian regularization techniques can also be used to solve overtraining problem where the data is split into two sets, training and testing sets. Here, sensitivity analysis is used to rank the input variables and then the network is trained again considering only the important units. Pruning is used to eliminate insignificant hidden units and input variables. One advantage over early stopping is that a larger training data set is available due to merging of training and validation sets. [8, 12, 14]

The storage capacity of the network is a function of the numbers of weights in the network, the number of outputs and the effectiveness of the training. The lower bound for the pattern storage of the MLP [32, 33, 42, 44] can be expressed as

$$C_L^{MLP} = N + N_h + 1 \tag{2.10}$$

The upper bound for the storage capacity of the MLP[32, 42, 44] when training is effective can be written as

$$C_{U1}^{MLP} = \frac{P_{ab}}{M} \tag{2.11}$$

where P_{ab} denotes the number of absolute free parameters (i.e., the actual weights and thresholds in the network). An alternative and more useful expression for upper bound [32,42,44] would be

$$C_{U2}^{MLP} = \frac{P_{ef}}{M} \tag{2.12}$$

where P_{ef} denotes the effective non-redundant parameters (i.e., the most significant weights) that affect the network performance.

The mapping efficiency which quantifies the efficiency with which the network utilizes weights can be expressed as [32,42,44]

$$E_f = \frac{P_{ef}}{P_{ab}} \tag{2.13}$$

2.2.2 Radial Basis Function Network

An RBF network is a feed forward neural network with a single hidden layer that has the ability to approximate a smooth functional relationship. RBFs are considered a hybrid of a sigma-pi network, an MLP, and Kohonen's Self Organizing Maps (SOM).[3,36,38] The hidden layer nodes have cluster center vectors used as hidden unit input weights. These RBF hidden units have nonlinear activations such as the commonly used Gaussian kernel. A typical radial basis function network is shown in Figure 2-2. Let $\mathbf{m_k}$ represent the mean vector of the k^{th} hidden unit where k has values from 1 to N_h. The net function of the k^{th} hidden unit is defined as

$$net_k = ||\mathbf{x} - \mathbf{m_k}||^2 \tag{2.14}$$

where the norm is generally Euclidean. The Gaussian hidden unit activation function is

$$O\,(net_k) = e^{-net_k/\beta^2} \tag{2.15}$$

where β is a spread parameter. Each output is a linear combination of hidden unit RBF activations, so output weight optimization (OWO) could be used to solve output weights. The i^{th} network output can be written as

$$y_i(x) = b_i + \sum_{k=1}^{N_h} w_{ik}\, O\left(net_k\right) \tag{2.16}$$

where w_{ik} and b_i are the weight from k^{th} hidden unit to the i^{th} output unit and threshold for the i^{th} output unit respectively.

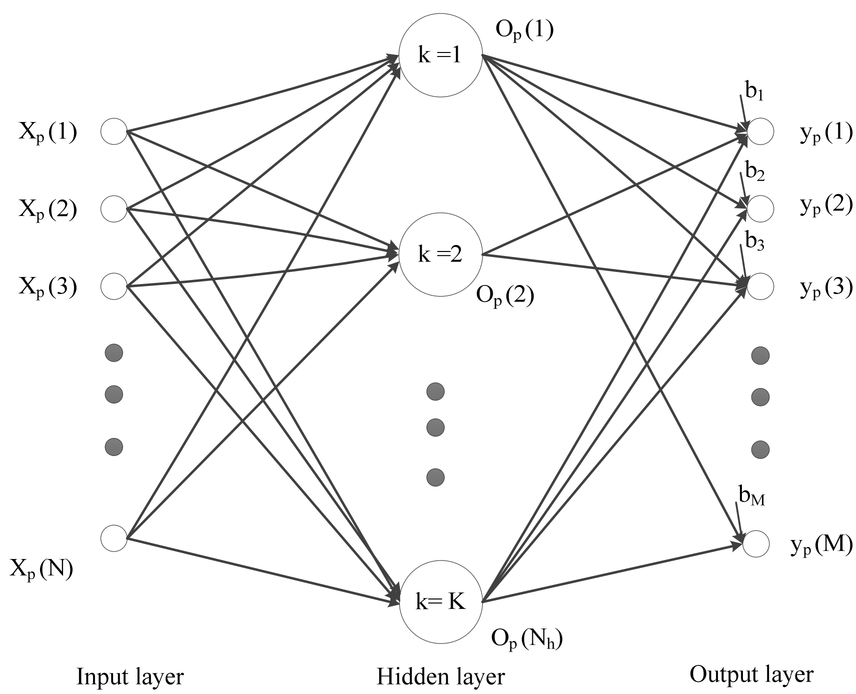


Figure 2-2 Radial Basis Function Network[62]

2.2.3 Piecewise Linear Network

In a piecewise linear network, the N dimensional input space is divided into K clusters and a linear mapping is approximated for patterns within a cluster as shown in Figure 2-3. In this network, the input vector is grouped in the cluster closest to it and the linear network corresponding to that cluster is used to compute the output.[28, 43, 63]

A piecewise linear network [28, 63] is characterized by

(a) K cluster center vectors $\mathbf{m_k}$ each of dimension N, where $1 \le k \le K$.

(b) K weight matrices $\mathbf{W_k}$, $1 \le k \le K$, store the weights of each cluster. Each weight matrix has dimensions M x (N+1).

(c) A weighted Euclidean distance measure to determine the cluster membership as

$k = \arg\min_m \{ d(\mathbf{x}, \mathbf{m_m}) \}$ where

$$d(\mathbf{x},\mathbf{m_m}) = \sum_{n=1}^{N} c(n)[x_p(n) - m_m(n)]^2 \qquad (2.17)$$

Here the weights c(n) are calculated as inverse of the variance ($\frac{1}{\sigma^2(n)}$).

(d) If the k^th^ cluster is chosen in (c), the network output vector is calculated as

$$\mathbf{y_p} = \mathbf{W_k} \cdot \mathbf{x_p} \qquad (2.18)$$

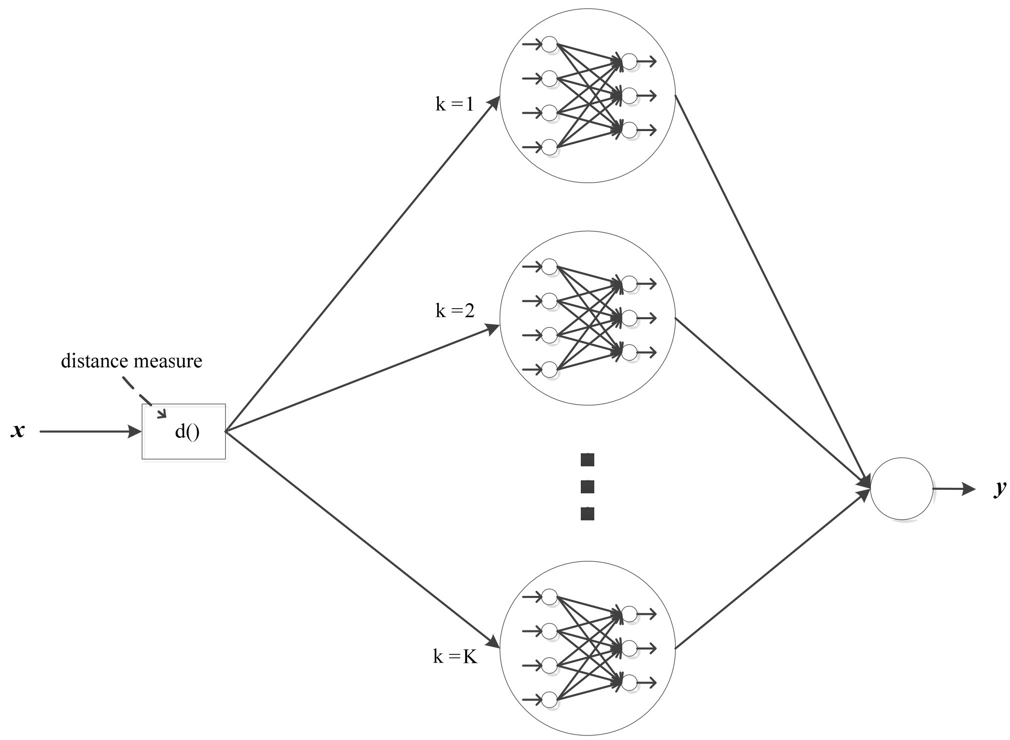


Figure 2-3 Piecewise Linear Network[62]

The number of absolute free parameters for the PLN can be written as [62]

$$P_{ab} = K \cdot N + (N+1) \cdot M \cdot K \qquad (2.19)$$

and the pattern storage can be written as [62]

$$C_{PLN} = P_{ab}/M = (K \cdot N + (N+1) \cdot M \cdot K)/M \qquad (2.20)$$

2.3 Previous work on air quality modeling

The literature contains a large number of research papers on ozone forecasting models. Most of the studies used statistical and artificial intelligence techniques such as multilinear regression (MLR), artificial neural networks (ANN), classification and regression trees (CART), and support vector machines (SVM). In this section, ozone (O_3) forecasting models developed by other researchers are briefly described, starting with the most recent and progressing to the oldest:

Sekar et al. (2015)[103] developed hourly O_3 and oxides of nitrogen (NO_X) prediction models based on Decision Tree algorithms: reduced error pruning tree (REPTree), and M5 P tree , and a multilayer perceptron using Levenberg-Marquardt (MLP-LM) in Delhi, India. A heavy traffic intersection in Delhi for pollutant data, and Safdarjung station for meteorological data corresponding to the years 2008-2010 were chosen for this study. O_3, NO_X, traffic data, atmospheric pressure (P), temperature (OT), wind speed (WS) wind direction (WD), cloud cover (CC), sunshine, rainfall, stability class, mixing height, temporal variables: day of the week and time of the day were used as input variables. MP 5 tree model performed better than MLP-LM and REPTree models.

Biancofiore et al. (2014)[101] applied ELMAN recurrent neural network model and MLR to predict hourly ozone up to 48 hours at Pescara, Central Italy. Hourly O_3, nitrogen dioxide (NO_2), OT, relative humidity (RH), WS, WD and ultraviolet radiation data from the year 2005 were used as input variables. ELMAN network model showed better performance than the MLR model.

Luna et al. (2014)[67] showed the potentiality of ANNs, and SVMs as chemo metric tools by applying these statistical techniques in the prediction of O_3 at Rio de Janeiro city, Brazil. A mobile monitoring station was used to collect hourly data at two locations, namely, Pontifical Catholic University area during July-October 2011, and Rio de Janeiro

State University area during November 2012-March2013. NO_2, NO, carbon monoxide (CO), O_3, OT, scalar wind speed, global solar radiation (SR), moisture content (MC) in the air were used as input variables. The use of principal component analysis (PCA) in dimension reduction was explored. MLP-LM and SVM were trained using the original data sets and the results showed slightly better performance of SVM's compared to that of MLP-LM.

Zahedi et al. (2014)[69] developed an adaptive neuro-fuzzy inference system to predict O_3 around the Shuaiba industrial area in Kuwait. A neuro fuzzy model was developed to predict O_3 using (Sugeno-Takagi-Gang fuzzy inference and hybrid) algorithm around the vicinity of Shuaiba area based on two months (March and April) of data measured every 5 minutes using a mobile station. O_3, WS, WD, RH, OT, SR, methane (CH_4), CO, CO_2, NO, NO_2, SO_2, non-CH_4 hydrocarbons, dust around the industrial area. The results showed that O_3 prediction performance of fuzzy neural network was better than that of a multilayer perceptron trained using back propagation (MLP-BP).

Tamas et al. (2014)[104] developed MLP-LM and persistence models to predict 24 hour ahead O_3 using (2008-2014) data from urban and suburban stations (Canetto, Sposata) in Ajaccio, and (Giraud, Montesoro) in Bastia from the French island of Corsica, France. O_3, NO_2, wind force, SR, OT, precipitation, and temporal variables, hour of the day and weekday number were used as input variables. MLP-LM models performed better than persistence models.

Alkasassbeh (2013)[68] compared the performance of MLP-BP, radial basis function (RBF) network, and SVM on the forecasting of daily mean surface O_3 at Chenbagaramanputhur, Kanyakumari district, India. Based on three months (May 2009 - July 2009) of data (7 readings per day each with 3-hour interval) the mean daily ozone concentration was forecast using RBF, MLP-BP, and SVM with input variables NO_2, mean temperature and RH. It was shown that RBF networks have better prediction capability than SVMs and MLP-BP; SVM's have better prediction capability than MLP-BP.

Arhami et al. (2013)[71] developed hourly prediction models separately for six pollutants for each day of the week using ANN coupled with Monte Carlo simulation. An MLP-BP was used to predict CO, NO_X, NO_2, NO, O_3, and particulate matter of size 10 μm (PM_{10}) based on 2007 hourly data collected from Fatemi station, Iran. Monte Carlo simulation was used to enhance the ANN prediction by reducing the uncertainty and variability involved in the input data by computing and analyzing the prediction interval that serves as an indicator for high degree of uncertainty. The input variables used were air temperature, wet bulb temperature, RH, WS, WD, P, CC, visibility code, and vapor pressure.

Pires et al. (2012)[102] predicted one day ahead hourly average O_3 using MLP-BP and Genetic Algorithms (GA) at Oporto, North Portugal. Hourly CO, NO, NO_2, O_3, OT, RH, SR, WS data were collected during May-August 2004 for this study. The use of GA improved the performance of MLP-BP.

Kandya et al. (2012)[86] studied the suitability of artificial neural networks in forecasting 8-hourly averaged O_3 at a busy traffic junction in Madras, India. Data collected for a period of 19 months (September 2008-March 2010) from a monitoring site located at the Indian Institute of Technology, Madras (IIT-M), Madras was used in developing a MLP to forecast 8 hourly averaged O_3 concentration. Also, comparison studies were made with respect to other O_3 forecasting models developed by Comrie (1997) at Phoenix, Tucson, Boston, Atlanta, and Charlotte. Comparison results showed that model developed by Kandya performed reasonably well. 8-hr average concentration of O_3, NO, NO_2, SO_2, CO, respirable suspended particulate matter, hydrocarbons, WS, WD, solar intensity, and pressure were used as inputs.

Paoli et al. (2011)[72] used neural networks to predict hourly O_3 concentration in Corsica Island, France. MLP - LM was developed to predict 1-hour forecasts of O_3 concentration based on hourly data collected during the period of October 2007 to May 2010 from a suburban station at Sposata, located near Ajaccio on the island of Corsica, France. O_3, NO_2, WS, WD, SR, RH, and hour of the day were used as input variables.

Taormina et al. (2011)[73] predicted daily maximum O_3 concentrations using adaptive neural networks in London. Pollutant data from Harlington station, London Hillington-Harlington (Heathrow airport zone) and meteorological daily data from a monitoring station located in Heathrow airport corresponding to the years 2004 to 2009 was used in developing a MLP-LM model. The optimal network architecture with selected features and proper time lags was saved. The testing results were improved by adaptively changing the weights from the optimal network saved using the back propagation. The input variables used were CO, NO, NO_2, NO_X, O_3, and SR.

Ibarra-Berastegi et al. (2009)[75] used neural networks for short-term prediction of SO_2, CO, NO_2, NO, O_3 pollutants in Bilbao, Spain. An MLP-BP, an MLP trained using two hidden layers, RBF network, and a generalized regression neural network (GRNN) have been shown to have better prognostic capabilities from 1 hour up to 8 hours based on the studies carried out on the two year hourly data (2000, 2001) obtained from six locations in the area of Bilbao. Traffic data, WS, WD, pollutants- SO_2, CO, NO_2, NO, and O_3 were used as input variables.

Salcedo-Sanz et al. (2009)[30] applied RBFs in the spatial regression analysis of NO_x and O_3 concentrations in Madrid, Spain. Hourly measurements of NO_x and O_3 were collected from 27 monitoring stations in Madrid corresponding to 6 years, from 2002 to

2007. In the spatial regression analysis, only quarterly and yearly averages of both the pollutants were considered while training RBFs' having Gaussian kernels and evolutionary based training algorithms. The evolutionary based RBFs' showed better performance and the results obtained from these networks were used as initial points in developing Land-Use Regression models (or Regression Mapping models) with the aid of Geographical Information Systems (GIS). This spatial regression analysis, in general, aids in restructuring the existing air quality monitoring network and statically analyzing the pollutants especially in the cities.

Salazar-Ruiz et al. (2008)[22] developed and compared 12 ozone prediction models based on input data O_3, OT, NO_2, NO, CO, resultant wind speed, and RH collected from Mexicali (Mexico)-Calexico (California, US) border area: using the data collected during the years 1999-2004 (excluding 2001), one day ahead maximum O_3 was predicted based on two different types of data sets i.e., one based on daily means and one based on the mean of the first six hours of the day. A persistence model, multilinear regression model, semi parametric ridge regression model, a MLP-BP model, an ELMAN recurrent neural network model and an SVM model were developed. Prediction performance of the artificial intelligence (AI) based models was better than that of the linear models, and among the AI based models, MLP-BP showed better performance than the ELMAN network and SVM; the ELMAN network performed better than the SVM.

Coman et al. (2008)[58] did comparison studies on a "Static MLP model based on a single MLP" and a "Dynamic model based on a cascade of 24 MLPs" that were developed with data collected during August 2000-July 2001 to predict hourly O_3 for a 24-hour horizon, at Prunay, and Aubervilliers stations, in Paris, France. Limited memory Broyden, Fletcher, Goldfarb, and Sahanno (BFGS) quasi-Newton algorithm and scaled conjugate gradient (SCG) algorithms were used in training the static and dynamic models. Prediction performance based on both the algorithms showed similar results. Also, static models performed slightly better than dynamic and persistence models. Hourly O_3, NO_2, RH, T, SR, sunshine duration, WS, sin (2πh/24), and cos (2πh/24), where h is hour of the day were used as inputs in this study.

Liu (2007)[79] developed a regression with time series (RTSE) model after incorporating principal component variable resulting from PCA to enhance peak daily one hour O_3 concentrations at Ta-Liao in Taiwan. Four different Box-Jenkins time series models were developed to simulate peak daily 1-hr O_3 concentrations in Ta-Liao based on data from the years 1997-2001. RTSE model with PC variable proved to be optimal model compared to ARIMA, RTSE model without PCA and RTSE model with additional PC variables. The input variables used were maximum temperature, dew temperature, WS, sunshine, O_3, and NO_X.

Dutot et al. (2007)[8] used neural network combined with neural classifier in forecasting daily maximum hourly O_3 peaks and European threshold O_3 exceedance level with 24 hour lead time in the city of Orleans, France. One MLP-LM based neural network, and two multilayer perceptrons trained using MLP-LM models with pattern balancing were developed and compared with a linear model, deterministic model and a persistence model based on data collected between April and September during the years (1999-2003) from three monitoring stations namely, Prefecture, La Source and Saint Jean de Braye. This neural network based model now called NEUROZONE is used in real time. Cloudiness, rainfall, WS, WD, temperature gradient, and O_3 were used as input data.

Sousa et al. (2007)[16] developed multiple linear regression models and neural network models based on principal components for the prediction of next day hourly O_3 concentration in Oporto, Northern Portugal. Hourly data from July 2003 was collected from a monitoring site in Oporto, Northern Portugal and four models (MLR, MLP-BP model based on original data, principal component regression and MLP-BP based on principal components) were developed. MLP-BP based on principal components, MLP-BP on original data showed better accuracy prediction compared to the two linear regression models. NO, NO_2, O_3, T, RH, and wind velocity were used as input variables.

Lu et al. (2006)[108] developed two stage neural network models to predict daily maximum O_3 concentrations separately for four air quality stations in Taiwan using five year data corresponding to 1998-2002. The two stage neural network model first utilized an unsupervised self-organizing map neural network (SOM) followed by K- means clustering (two level clustering approach) to delineate the meteorological variables into distinct meteorological regimes and then a supervised multilayer perceptron (MLP) was used to predict O_3 within each meteorological regime. The superior performance of the two stage models developed at four stations separately was shown in comparison with models developed based on multilayer perceptron, multiple linear regression and two level clustering followed by multiple linear regression. Hourly data of O_3, CO, NO_X, SO_2, PM_{10}, WS, WD, OT, average pressure, RH, cloud cover, precipitation, and global radiation were used as input variables.

Wang et al. (2006)[21] forecasted ground level O_3 concentration using a hybrid training algorithm. An MLP model with a single hidden layer was trained using two optimization algorithms coupled synergistically, namely: a Particle Swarm algorithm (PSO) and LM. They used 4 years of data (2000-2003) from two different stations (Tseun Wan, and Tung Chung) in Hong Kong and predicted one day ahead daily maximum 1-hr mean O_3 concentration. Average daily values of NO_2, NO_X, NO, CO, OT, SR, WS and temporal variable (day of the year) were used as input variables.

Pastor-Barcenas et al. (2005)[10] predicted hourly O_3 by employing sensitivity analysis and pruning techniques to artificial neural networks. An MLP-BP was trained using hourly data from April 2002, collected from a rural monitoring station located in “Centre de Capacitacio Agraria de Carcaixent” in Valencia, Spain to predict 24 hour ozone concentration. The input variables used were hourly NO, NO_2, O_3, WS, WD, OT, P, RH and solar irradiance.

Abdul-Wahab et al. (2005)[96] applied PCA and multiple regression in modelling hourly ground level ozone based on the summer data collected (every 5 minutes) during June 1997 using Kuwait University mobile laboratory at Khaldiya, Kuwait. Separate regression analysis was carried out for ozone prediction for day light and night time periods respectively. O_3, NO_2, NO, CO, CO_2, SO_2, non CH_4 hydrocarbons, OT, SR, RH, WS, and WD were used as input variables.

Wirtz et al. (2005)[81] developed a ground level O_3 forecasting neural network model. An MLP-BP was used on data sets from two monitoring stations, Edmonton East monitoring station, and Stony Plain station in Edmonton, Alberta, Canada. Wirtz et al. were successfully able to predict 2 hr O_3 in advance using only the summer data from May to September from the years 1999 to 2003. The input variables used were CO, NO, NO_2, SO_2, total hydrocarbons, mixing height, opacity, RH, WS, WD, and temporal variables (hour of the day, month of the year, day of the week).

Heo et al. (2004) [97] developed methodologies for classifying high-level O_3 episodes within the city of Seoul, Korea by applying cluster and disjoint PCA. Consequently, classified O_3 episodes were used as a database for developing daily maximum O_3 forecasting model using fuzzy expert system, and MLP-BP was used to predict daily maximum hourly O_3. Hourly data with high level O_3 episodes corresponding to four monitoring stations in Seoul during the period of 1989-1999 was used in this study. CO, NO_2, SO_2, O_3, surface wind speed, surface wind direction, upper wind speed, upper wind direction, surface temperature, upper temperature, surface solar radiation, and surface relative humidity were used as inputs.

Zolgadri et al. (2004)[95] developed an integrated operational O_3 warning system at Bordeaux, France based on data collected during 1998-2001 at Bordeaux Grand Parc station. A non linear adaptive state space estimator (NASSE), gain scheduling (defined for modeling threshold exceedance for extreme O_3 concentration) and an (MLP-LM) were used in making the daily maximum O_3 warning monitoring system. Hourly radiation, solar intensity, barometric pressure, WS, WD, RH OT, trend of seasonal variation of O_3, and $[NO_2]/[NO]$ were used as inputs.

Kumar et al. (2004)[89] applied autoregressive integrated moving average (ARIMA) modeling approach to forecast one day ahead daily maximum O_3 in Brunei Darussalam based on O_3 data corresponding to July 1998-March 1999.

Chaloulakou et al. (2003)[109] developed a daily maximum hourly O_3 prediction model using (April - October) 1992-1999 data collected at N. Smirni, Liossia, Maroussi, and Likovrissi stations in Athens, Greece. An MLP-LM model was developed and compared with MLR using WS, SR, RH, surface OT, OT at 850 hPa (850 millibars), WD index and O_3 as input variables. Prediction based on MLP-LM showed better performance than that of MLR.

Rohli et al. (2003)[98] used PCA and multiple regression analysis to forecast daily maximum 8-hour O_3 concentrations in Baton Rouge, Louisiana. Rohli et al. developed regression models for each of the eleven sites chosen based on the data corresponding to the years 1995-2000 and also proposed a decision making tree for short range forecasting of O_3 exceedance at these sites.

Wang et al. (2003)[25] developed an adaptive RBF network to predict daily maximum ozone concentration in Hong Kong. An adaptive RBF that can dynamically determine the number of hidden nodes was used along with the statistical characteristics of ozone to forecast daily maximum O_3 based on (1999-2000) data measured from three monitoring stations, namely, Tsuen Wan, Kwai Chung, and Kwun Tong in Hong Kong. O_3, NO_2, NO, NO_X, SO_2, respirable suspended particles, WS, WD, SR, indoor temperature and OT were used as input variables.

Vautard et al. (2001)[99] developed a simplified hybrid statistical-deterministic chemistry transport model to predict O_3 in real time in Paris during summer, 1999. Weather forecasting data collected from European Center for Medium Range Weather Forecasts (ECMWF) was processed to be used in the chemistry transport model (CHIMERE) for ozone prediction. This model was meant to be suitable to continental cities like Paris only.

Kaprara et al. (2001)[100] predicted daily maximum O_3 concentration levels in Athens using CART technique. Daily maximum and minimum concentrations of pollutant and meteorological data collected during the period of 1990-1999 in Athens area consisting of nine monitoring stations was used in developing a CART based O_3 forecasting model. Results obtained showed better prediction performance of the CART model compared to that of MLR.

Gardner and Dorling (2001, 2000) [13, 14, 83] did extensive research in ground level ozone prediction using neural networks. In their work published in 2001, they described a technique that employed MLP-CG that maximizes the removal of variability in daily maximum O_3 with fluctuations in meteorological conditions and was shown to remove more of the variability than does Kolmogorov-Zurbenko filter and conventional-based technique. In their work published in 2000, they applied MLP- CG, regression trees, and linear models to predict hourly O_3 using the data from the five O_3 monitoring sites, Bristol, Edinburgh, Eskdalemuir, Leeds, and Southampton in UK based on data collected during 1993-1997. MLP-CG preformed better than regression trees and linear models but the regression trees were more readily interpretable.

Cobourn et al. (2000)[85] applied neural networks to predict ground level ozone. An MLP-BP developed based on a data set collected from seven monitoring stations within the Louisville Air Quality Control Region. In their study, Cobourn et al. used ozone season (May to September) data starting from 1993 to 1999 to predict daily maximum 1-hr ozone concentration. Daily 8-hour average of O_3, clear-sky atmospheric transmittance daily minimum temperature, wind speed, cloud cover, and humidity were used as input data.

Prybutok et al. (2000)[23] compared neural network model with ARIMA and multivariate regression model developed to forecast daily maximum O_3 concentration. A MLP-BP, stepwise regression model, and Box Jenkins ARIMA ozone forecasting model were developed using (June-October) 1994 data from a monitoring station in Houston. It was shown that MLP-BP has superior performance compared to ARIMA and regression models. Hourly values of NO, NO_2, O_3, OT, WS, WD, and CO_2 were used as input variables.

Hadjiiski et al. (2000)[6] used sensitivity analysis and neural networks to forecast hourly O_3 concentration in Houston. Using sensitivity analysis, relevant input variables were found and an MLP-BP model was developed based on these selected features to forecast hourly O_3 up to 5 hours with data collected from two monitoring stations namely, Galleria and Clinton, in Houston during the months of June-November, 1993. Fifty three hydrocarbons (C_2-C_{10} compounds), O_3, NO_X, NO, NO_2, ultraviolet radiation, and OT were used as input variables.

Sohn et al. (2000)[11] developed short term and long term O_3 forecasting models using neural networks and spatio-temporal analysis. MLP- CG neural network model was developed in Seoul, South Korea to forecast short term (1- 6 hr) prediction and long term (16-21 hr) prediction using data from the period August-September, 1997. Forecasting results improved when the neural network model was used along with spatio-analysis (distribution of O_3 concentrations) that includes the effects of advection and dispersion. Hourly data of O_3, NO_2, CO, SO_2, OT, WS, sunlight, and humidity were used as input data.

Benvenuto et al. (2000)[82] used neural network to develop short term and medium term forecasting models for O_3, CO, NO_2 in Venice, Italy. I hour, 3 hours and daily maximum concentrations of O_3, CO, NO_2 were predicted using MLP-BP with 1995 data collected from Ente Zona Industriale di Porto Margera and Venice municipality monitoring network areas. Hourly measurements of global radiation, humidity, precipitation, pressure, vehicle flow rate, OT, WS, WD, SO_2, O_3, NO, NO_2, non CH_4 hydrocarbons, and PM_{10} were used as input variables.

Spellman (1999)[15] developed different neural network models for daily maximum O_3 forecasting for five sites in UK. Five sites with different topographical and demographical features (Bloomsbury, Leeds and Birmingham being urban sites; Harwell (Oxfordshire) being rural and Strath Vaich being a remote site) were chosen and (May-September) data corresponding to the years 1993-1996 was used in developing MLP-BP with two hidden layers, and regression models separately for each site. The O_3 prediction accuracy of the ANN models was found to be slightly better than the regression models. Hourly O_3, SO_2, PM_{10}, WS, WD, and OT were used as input variables.

Comrie (1997)[7] compared site specific MLP-BP and MLR models developed for daily maximum O_3 forecasting. Eight monitoring sites from different cities (Atlanta, Boston, Charlotte, Chicago, Phoenix, Pittsburgh, Seattle, and Tucson) in USA were chosen for this study and hourly data was collected during the months of May-September over the five year period 1991-1995. In all the eight study sites, MLP-BP with lagged data performed slightly better than MLP-BP without lagged data, MLR with lagged data, and MLR without lagged data. Daily maximum temperature, average daily wind speed, daily total sunshine, and O_3 were used as input variables.

Yi et al. (1996)[24] developed daily maximum O_3 forecasting model using MLP-BP at a monitoring site in Dallas-Fort Worth (DFW) region, Texas. Based on data corresponding to the months of June to October for the years 1993-94, daily maximum O_3 forecasting models were developed using MLP-BP, multilinear regression model, and Box Jenkins model. MLP-BP showed better prediction results among the three models. Hourly values of NO, NO_2, O_3, CO_2, OT, WS, and WD were used as input variables.

Bloomfield et al. (1996)[112] developed non linear regression model for daily maximum O_3 concentrations in Chicago based on median values of (1981-1991) data collected from 45 monitoring sites.

Ryan (1995)[106] undertook pilot plant studies in Baltimore and developed stepwise regression, subjective or expert analysis and CART O_3 forecasting models. Daily maximum hourly O_3 (up to 72 hours) was predicted using the (1983-1993) data that has sky cover, WS, temperature, pressure, O_3, and dew point temperature as input variables. Results showed that Subjective or Expert Analysis performed better than stepwise MLR and CART in strong O_3 episodes. Stepwise MLR was better than CART.

Clark et al. (1982)[90] developed next day maximum 1-hour O_3 stepwise MLR models for each of the 27 monitoring stations in Northeastern states in the U.S. based on the (June-September) data collected during 1975-1977. 35 prognostic variables including hourly OT, absolute humidity, WS, O_3, NO_X, precipitation, sea level pressure, and altitude were used as input variables.

Karl (1979)[107] predicted 1-hour maximum O_3 concentrations for 1 day ahead and 2-day ahead using data from 25 sampling sites divided into three groups at St.Louis, Missouri. Boundary layer, WS, OT, precipitation, RH, P, O_3, NO_X, dew point, HC's, vertical velocity, sine, and cosine of Julian date were used as input variables.

Wolff et al. (1978)[105] used stepwise regression model to predict maximum afternoon O_3 concentrations based on (April-September) 1976 data from across the northeast quadrant of the U.S. northern New Jersey. Approximately 75 monitoring sites have been considered for this study. A stepwise regression model was calibrated based on New Jersey data and tested on sites at Northeastern Ohio, Marquette, MI, Norfolk, VA, Cook County, IL and Connecticut. OT, absolute humidity, WS, O_3, NO_X and hydrocarbons were used as input parameters.

Chapter 3

Data Description

The data collected for the development of a comprehensive ozone forecasting model for the multiple cities across the U.S. is described in detail as shown in the tables below. For this study, five different regions/air basins in and around the cities that are non attainment for ground level ozone according to United States Environmental Protection Agency (USEPA) have been chosen. They are Dallas Fort Worth region, Houston-Galveston-Brazoria from the state of Texas; Los Angeles (South Coast air basin), San Diego air basin and San Joaquin Valley air basin from the state of California. According to USEPA (2008), Dallas-Fort Worth, Houston-Galveston-Brazoria, Los Angeles, San Diego and San Joaquin have been classified as moderate, marginal, extreme, moderate and extreme 8-hour ozone non attainment areas.

The variables responsible for the formation of ozone include nitrogen dioxide, nitric oxide, volatile organic compounds, temperature, wind, solar radiation, relative humidity, apart from the demographical and topographical characteristics of the area. The choice of input variables considered in this study was based on the availability of data in/near the ozone monitoring sites.

The five year hourly data collected during the period 2010-2014 in all the fifty monitoring stations, in the states of Texas and California includes pollutants ozone (ppb), nitric oxide (ppb), nitrogen dioxide (ppb), and meteorological variables namely, resultant wind speed (miles/hour), resultant wind direction (degrees), temperature (^{o}F), and solar radiation (langleys per minute in Texas Commision on Environmental Quality database, or Watts/m^2 in California Air Resources Board database).

Ozone concentrations are proportional to the ratio $[NO_2]/[NO]$, according to the Leighton mechanism. Thus, the ratio of $[NO_2]/[NO]$ was considered as one of the variables in the initial trials but did not improve the results.

The Appendix A describes the data collected from all the five cities in the US. Thirteen ozone monitoring sites from DFW region, twelve ozone monitoring sites from Houston-Galveston-Brazoria region, ten ozone monitoring sites from Los Angeles county, eight ozone monitoring sites from San Joaquin Valley air basin, and seven ozone monitoring sites from San Diego air basin have been considered in this study.

Chapter 4

Example forecasting system and its problems

The objective of this study is to develop a forecasting system for multiple cities/regions for the prediction of daily maximum ground level ozone. The first step in obtaining a desired forecasting system involves collection of data and preprocessing before this data is fed to the neural network for training. Firstly, the example system that is impractical is described below:

4.1 Example system inputs and outputs

The input variables considered in this study are as follows:

- Temporal variables: Temporal variables include the day of the year (DOY) that has values from 1 to 365 (366 in case of leap year); hour of the day (HOD) that has values from 1 to 24; and day of the week (DOW) that has values say, 1 to 7.
- Spatial variables: Spatial variables include latitude and longitude of the monitoring stations of various cities/regions. The inclusion of these variables makes a neural network distinguish the monitoring stations and allows the network to identify the geographic location the pattern actually belongs to.
- Meteorological variables: Hourly values of meteorological variables, namely, solar radiation (SR), ambient temperature (OT), wind speed (WS) and wind direction (WD).
- Pollutant variables: Hourly concentrations of nitrogen dioxide (ND), nitric oxide (NO) and ozone (Oz).

The hourly data collected is used in generating daily mean, minimum and maximum values of the meteorological and pollutant variables. The output variables include future (one day ahead, two day ahead, and up to three day ahead) daily maximum concentrations of ozone. So the N inputs will include time variables, spatial variables, past mean, minimum and maximum daily values of temperature, solar radiation, wind speed, wind direction, nitrogen dioxide, nitric oxide, and ozone. M outputs will include future daily maximum values of ozone (up to three days).

The input vector, $\mathbf{x}_{\mathbf{city}}$ for each city/region in the example system will have different number of inputs (i.e., N varies for each city) due to different number of monitoring stations as shown below:

$$\begin{aligned}\mathbf{x}_{\mathbf{city}}^{\mathbf{T}} = [\, & \mathbf{x}_{\mathbf{time}}^{\mathbf{T}} \vdots\, \mathbf{x}_{\mathbf{WS(1)}}^{\mathbf{T}} \vdots \mathbf{x}_{\mathbf{WS(2)}}^{\mathbf{T}} \vdots \ldots \mathbf{x}_{\mathbf{WS(NS)}}^{\mathbf{T}} \vdots\, \mathbf{x}_{\mathbf{WD(1)}}^{\mathbf{T}} \vdots \mathbf{x}_{\mathbf{WD(2)}}^{\mathbf{T}} \vdots \ldots \mathbf{x}_{\mathbf{WD(NS)}}^{\mathbf{T}} \vdots\, \mathbf{x}_{\mathbf{OT(1)}}^{\mathbf{T}} \vdots \mathbf{x}_{\mathbf{OT(2)}}^{\mathbf{T}} \vdots \ldots \mathbf{x}_{\mathbf{OT(NS)}}^{\mathbf{T}} \vdots \\ & \mathbf{x}_{\mathbf{SR(1)}}^{\mathbf{T}} \vdots \mathbf{x}_{\mathbf{SR(2)}}^{\mathbf{T}} \vdots \ldots \mathbf{x}_{\mathbf{SR(NS)}}^{\mathbf{T}} \vdots \mathbf{x}_{\mathbf{ND(1)}}^{\mathbf{T}} \vdots \mathbf{x}_{\mathbf{ND(2)}}^{\mathbf{T}} \vdots \ldots \mathbf{x}_{\mathbf{ND(NS)}}^{\mathbf{T}} \vdots\, \mathbf{x}_{\mathbf{NO(1)}}^{\mathbf{T}} \vdots \mathbf{x}_{\mathbf{NO(2)}}^{\mathbf{T}} \vdots \ldots \mathbf{x}_{\mathbf{NO(NS)}}^{\mathbf{T}} \vdots \\ & \mathbf{x}_{\mathbf{POz1}}^{\mathbf{T}} \vdots \mathbf{x}_{\mathbf{POz(2)}}^{\mathbf{T}} \vdots \ldots \mathbf{x}_{\mathbf{POz(NS)}}^{\mathbf{T}} \,]\end{aligned} \tag{4.1}$$

$$\mathbf{x}_{\mathbf{time}}^{\mathbf{T}} = [\mathrm{DOY} \quad \mathrm{DOW}] \tag{4.2}$$

The actual output vector can be represented as

$$\mathbf{y}_{\mathbf{city}}^{\mathbf{T}} = [\mathbf{y}_{\mathbf{Oz(1)}}^{\mathbf{T}} \vdots\, \mathbf{y}_{\mathbf{Oz(2)}}^{\mathbf{T}} \vdots \ldots \mathbf{y}_{\mathbf{Oz(NS)}}^{\mathbf{T}}] \tag{4.3}$$

The desired output vector can be represented as

$$\mathbf{t}_{\mathbf{city}}^{\mathbf{T}} = [\mathbf{t}_{\mathbf{Oz(1)}}^{\mathbf{T}} \vdots\, \mathbf{t}_{\mathbf{Oz(2)}}^{\mathbf{T}} \vdots \ldots \mathbf{t}_{\mathbf{Oz(NS)}}^{\mathbf{T}}] \tag{4.4}$$

where NS = number of monitoring stations; and

$\mathbf{x}_{\mathbf{WS(1)}}^{\mathbf{T}}, \mathbf{x}_{\mathbf{WS(2)}}^{\mathbf{T}}, \mathbf{x}_{\mathbf{WS(NS)}}^{\mathbf{T}}$; $\mathbf{x}_{\mathbf{WD(1)}}^{\mathbf{T}}, \mathbf{x}_{\mathbf{WD(2)}}^{\mathbf{T}}, \mathbf{x}_{\mathbf{WD(NS)}}^{\mathbf{T}}$; $\mathbf{x}_{\mathbf{OT(1)}}^{\mathbf{T}}, \mathbf{x}_{\mathbf{OT(2)}}^{\mathbf{T}}, \mathbf{x}_{\mathbf{OT(NS)}}^{\mathbf{T}}$;

$\mathbf{x}_{\mathbf{SR(1)}}^{\mathbf{T}}, \mathbf{x}_{\mathbf{SR(2)}}^{\mathbf{T}}, \mathbf{x}_{\mathbf{SR(NS)}}^{\mathbf{T}}$; $\mathbf{x}_{\mathbf{ND(1)}}^{\mathbf{T}}, \mathbf{x}_{\mathbf{ND(2)}}^{\mathbf{T}}, \mathbf{x}_{\mathbf{ND(NS)}}^{\mathbf{T}}$; $\mathbf{x}_{\mathbf{NO(1)}}^{\mathbf{T}}, \mathbf{x}_{\mathbf{NO(2)}}^{\mathbf{T}}, \mathbf{x}_{\mathbf{NO(NS)}}^{\mathbf{T}}$;

$\mathbf{x}_{\mathbf{POz\,(1)}}^{\mathbf{T}}, \mathbf{x}_{\mathbf{POz\,(2)}}^{\mathbf{T}}, \mathbf{x}_{\mathbf{POz\,(NS)}}^{\mathbf{T}}$;

represent vectors that have past daily mean, minimum, and maximum values of wind speed, wind direction, temperature, solar radiation, nitrogen dioxide, and nitric oxide, ozone.

$\mathbf{y}^{\mathbf{T}}_{\mathbf{Oz(1)}}, \mathbf{y}^{\mathbf{T}}_{\mathbf{Oz(2)}}, \mathbf{y}^{\mathbf{T}}_{\mathbf{Oz(NS)}};\quad \mathbf{t}^{\mathbf{T}}_{\mathbf{Oz(1)}}, \mathbf{t}^{\mathbf{T}}_{\mathbf{Oz(2)}}, \mathbf{t}^{\mathbf{T}}_{\mathbf{Oz(NS)}};$ represent actual network outputs and desired output vectors that have future daily maximum values (up to three days) of ozone as features at stations 1, 2 up to NS.

4.2 Training/Validation/Testing data in the example system

The data from DFW region, Houston-Galveston-Brazoria region, Los Angeles air basin, San Joaquin air basin and San Diego air basin of years 2010, 2011, 2012, 2013, and 2014 will be collected and the preprocessed data from the years 2010, 2011, 2012, and 2013 will be randomly split so that (3/4) of the data will be used for training and (1/4) of the data for validation. The generated example system pattern files of size (N_v x (N+M)) for each city will be of different size as (N+M) varies due to different number of monitoring stations in each city. Figure 4 -1 shows an impractical example multi-city pollutant forecasting system.

4.3 Problems associated with the example system

4.3.1 Discontinuous inputs

Some inputs can be discontinuous such as the time variables and wind direction. Discontinuity associated with the time inputs: The formation of ozone and its transport is influenced by the seasonal, the diurnal and the hourly changes in the meteorological variables as well as the pollutant variables in the atmosphere. The inclusion of the temporal inputs makes a neural net perform better. The temporal inputs being considered are day of the year (DOY) that accounts for seasonal variations, day of the week (DOW) that indicates vehicle miles travelled per day. The time variables: DOY that has values

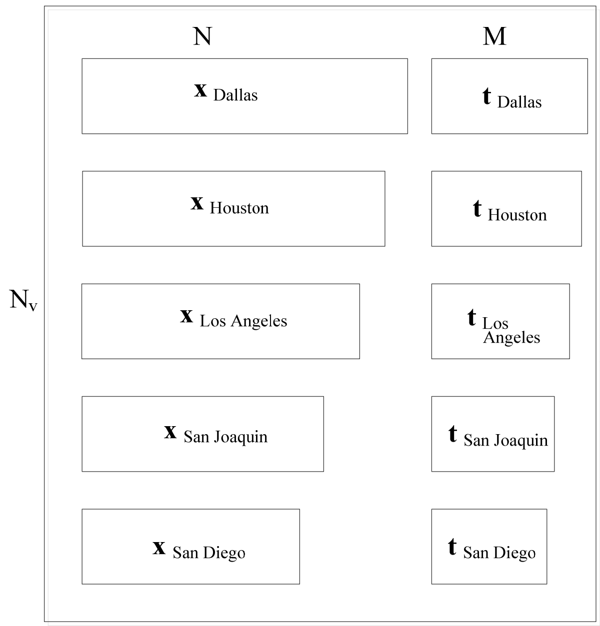


Figure 4-1 An impractical example multi-city pollutant forecasting system

from 1 to 365 in a non leap year is discontinuous since the last day of December (365th day) is followed by first day of January (1St day) of the following year; DOW that has values from 1 to 7 is discontinuous since the last day of a week, Saturday is followed by the first day of a next week, Sunday. One way to represent these input time variables is with binary format. But representing time variables in binary format would increase the number of input features enormously and might affect the network generalization performance.

Discontinuity associated with the wind direction: Wind speed and wind direction are abrupt and uncertain in nature. Wind direction (WD) expressed in radians/degrees is like the phase of a signal and is discontinuous: WD takes values from 0 to 360 degrees but the shift from 0 to 360 degrees is not the same as 360 to 0. The inclusion of these variables without normalizing or transforming might affect the network's ability to learn during training.

4.3.2 Missing data

For many reasons the meteorological data and pollutant data is not always complete or valid. This could be due to the malfunctioning of the equipment operating at various monitoring sites or could be due to power outages. This missing data problem can be handled using linear interpolation. If data was missing for a longer duration (say more than a month), data from nearby station will be used (when nearby station was less than 5 miles); when the nearby stations were at a distance greater than 5 miles, then the average values of the surrounding stations will be used to fill up missing values.

4.3.3 Encoding data from multiple cities

Each city has different number of monitoring stations, N_s. The feature vectors corresponding to each city will have different dimensions as the number of inputs, N will be different due to different number of monitoring stations in each city and arranging these inputs as shown in Figure 4-1 would increase the number of inputs or columns in the data set. Making the pattern files in this way is impractical.

4.3.4 Memorization

In an ideal Bayes estimator, training error decreases monotonically as long as we add more information by increasing N. But, according to Hughes phenomenon[12,14,18,32,33], in real processors (say, MLP), increasing the number of hidden units or the number of inputs, leads to memorization or over-fitting problem or overtraining. In memorization, training error tends to decrease, while testing error increases when the network is fed with new unseen data. This over-fitting problem occurs when the network memorizes the specific input output patterns rather than the relationship between them. To avoid memorization, and achieve better generalization performance we have to make

(a) (N_v/N) large

(b) N_h or N small so that

$$C_{\mathrm{U1}}^{\mathrm{MLP}} = \frac{P_{\mathrm{ab}}}{M} \ll N_{\mathrm{v}} \tag{4.5}$$ [62]

4.3.5 Noisy or dependent data

When the data set contains redundant features or inputs, it is called noisy or dependent data. Redundant inputs increase N and storage capacity without providing useful information. When neural nets like the MLP are trained with noisy data, useless inputs act like noise and this leads to overtraining. Overtrained models show poor generalization performance when fed with unseen data.

Chapter 5

Possible System

5.1 Possible System Approaches

The example forecasting system of Chapter 4 is impractical. To build a better system, the modeler should see that the neural network has the following characteristics to ensure proper training: continuous inputs; training file should be free of the "curse of dimensionality"[43] (The "curse" typically refers to exponential growth in the computational effort as the number of input increases linearly); pattern files should be "thin and tall" with $N_v >> N$; favorably, input features that are linearly independent and are conducive for training; same number of inputs for each pattern; input feature definition same for all the cities.

Two possible approaches that might solve the problems associated with the example system are described below:

5.1.1 Tall file data approach

In the tall file data approach, the pattern files of each monitoring station of all the cities are concatenated one below the other as shown in Figure 5-1. This approach has more patterns and fewer input features (columns); assists in good generalization and prevents memorization. The approach is described below:

(i) Each monitoring station in each city generates one training pattern each day.

(ii) Consider the following input variables for a pattern:

- The latitude and the longitude (expressed in decimal degrees) of each monitoring station in the city. The latitude and longitude of a city's center (say, the average of the latitudes and longitudes of all the monitoring stations in the city). The general sign convention adopted for latitude: North is positive; for longitude: East is positive. The two inputs, i.e., the latitude and longitude tell the system which city the patterns come from.

- For each monitoring station, the two inputs, i.e., the deviations of the latitude and longitude of the monitoring stations from the latitude and longitude of city's center tell the system which monitoring station in the city the data comes from.
- Four temporal variables encoded as Cos$(\frac{2\pi}{365} \times DOY)$, Sin$(\frac{2\pi}{365} \times DOY)$ for non leap year and Cos$(\frac{2\pi}{366} \times DOY)$, Sin$(\frac{2\pi}{366} \times DOY)$ for leap year that represent the season of the year and Cos$(\frac{2\pi}{7} \times DOW)$, Sin$(\frac{2\pi}{7} \times DOW)$ that represent the day of the week. These four variables that account for season, and day of the week are continuous in nature.
- Temperature (x_{OT}), Nitric oxide (x_{NO}), Nitrogen dioxide (x_{ND}) that contribute towards photochemical production.
- Wind speed (x_{WS}), and wind direction (x_{WD}) that relate to ozone transport can be encoded in complex form as ($x_{WS} \cdot$ Cos(x_{WD})), and ($x_{WS} \cdot$ Sin(x_{WD})), to account for the discontinuity associated with the wind direction and uncertainty in the wind speed.
- Previous ozone levels (x_{POz}) that account for ozone accumulation.
- Solar radiation (x_{SR}).

(iii) After preprocessing, the input vector will consist of the daily minimum, mean and maximum values of the above input variables with time delays (up to 3 days). Meteorological variables such as temperature take values that can be expressed as $x_{T(k)} = x_T(-k)$, where $x_T(-k)$ represents ambient temperature at past *k* days. For example, $x_{T(0)} = x_T(0)$ represents ambient temperature at current day, $x_{T(1)} = x_T(-1)$, represents ambient temperature one day ago, $x_{T(2)} = x_T(-2)$, represents ambient temperature 2 days ago. Likewise, pollutant and meteorological variables such as x_{NO}, x_{ND}, x_{Oz} and x_{SR} take values that can be expressed as $x_{NO(k)} = x_{NO}(–k)$, $x_{ND(k)} = x_{ND}(–k)$, $x_{POz(k)} = x_{POz}(–k)$, $x_{SR(k)} = x_{SR}(–k)$, where *k* has values from 0 to 2 respectively. The input vector will now have, say, N = 71 features (i.e., 4 spatial variables, 4 temporal variables, 4x3x3 (= 36) meteorological variables (4 variables × 3 (i.e., mean, minimum, and maximum) x 3 time delays) and 3x3x3 (=27) pollutant variables (i.e., 3 variables × 3 (i.e., mean, minimum, and maximum) x 3 time delays). Another input vector with N = 69 features can also be made by considering only 2 spatial variables (i.e., latitude and longitude of a monitoring station).

(iv)The output feature vector consists of future daily maximum values of ozone that can be expressed as $t_{Oz(j)} = t_{Oz}(j)$, where j has values from 1 to 3. For example, $t_{Oz(1)} = t_{Oz}(1)$, represents ozone concentration one day ahead of time, $t_{Oz(2)} = t_{Oz}(2)$, represents ozone concentration two days ahead of time, and , $t_{Oz(3)} = t_{Oz}(3)$, represents ozone concentration three days ahead of time. The output vector will have M = 3 features.

(v) After preprocessing, the training data file will have (N + M) columns for each monitoring station. In a similar way, we preprocess data of all monitoring stations in all the cities being considered. The order of the input features and output features should be same for each monitoring station.

Finally, a large data set is made by combining all these preprocessed data. The total number of patterns in one year can be expressed as

$$N_{\mathrm{v}} = N_{\mathrm{days}} \sum_{\mathrm{city}=1}^{\mathrm{N_c}} (N_{\mathrm{s}})_{\mathrm{city}} \tag{5.1}$$

where N_c, N_s are the numbers of cities and monitoring stations respectively; N_{days} represents the number of days in a year. These data files prepared for the years 2010, 2011, 2012, 2013, and 2014 can be used for making training, validation and testing data. A practical multi-city neural net forecasting system is shown in Figure 5-1.

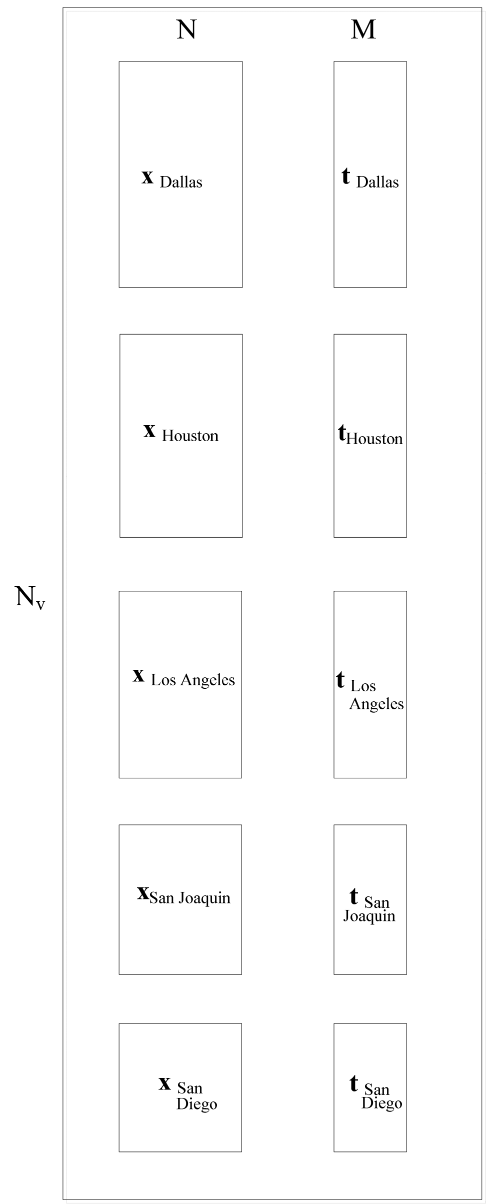


Figure 5-1 A practical multi-city neural network pollutant forecasting system

5.1.2 Median Approach

In the median approach, hourly data is preprocessed in such a way that each city generates one training pattern each day. Each variable is a median taken over the city's monitoring stations. Even though this approach helps reduce the number of input features from that of the example system, we might not get better results compared to the tall file data approach with more patterns and fewer inputs. One reason why this approach could fail is that the meteorological variables such as wind speed, solar radiation, wind direction and temperature show complex behavior and these vary from station to station. Also, these meteorological variables are influenced by topographical features and hence are site specific. Preprocessing can be done in the following way:

(i) Find the median of all the corresponding temperatures at all monitoring stations in a city. This median value becomes the representative temperature for that particular city. Likewise, find a median value for each input variable, namely wind speed, wind direction, nitric oxide, nitrogen dioxide and ozone. Now, each city contains representative median values of all input variables.

(ii) Choose a reference location (say, city center), represented by latitude and longitude for each city which is the average of the latitude and longitude of all the monitoring stations in the particular city.

(iii) After preprocessing, the input vector will have daily minimum, mean and maximum values of the above inputs (medians) with time delays (up to 3 days). For example, median values of meteorological variables such as temperature take values that can be expressed as $MT_k = MT(–k)$, where $MT(–k)$ represents median temperature of that particular city at the past k days. For example, $MT_0 = MT(0)$ represents median temperature at current day, $MT_1 = MT(–1)$ represents median temperature on the previous day, $MT_2 = MT(–2)$ represents median temperature 2 days ago, and so on.

Likewise, median values of wind speed and wind direction can be expressed as $MWS_k = MWS(-k)$, $MWD_k = MWD(-k)$, where k has values from 0 to 2 respectively. Also, the median wind speed and wind direction have to be expressed in complex form as (MWS (-k) · Cos MWD (-k)), and (MWS (-k) · Sin MWD (-k)). Likewise, median values of pollutant variables NO, NO_2 and ozone take values that can be expressed as $MNO_k = MNO(-k)$, $MNO_{2k} = MNO_2(-k)$, $MO_k = MO(-k)$ where k has values from 0 to 2 respectively.

(iv) The output vector consists of future 3 days ahead median values of daily maximum ozone that can be expressed as $MO_j = MO(j)$ where j has values from 1 to 3. The output will have M = 3 features. For example, $MO_1 = MO(1)$ represents ozone concentration one day ahead of time, $MO_2 = MO(2)$ represents ozone concentration two days ahead of time, and $MO_3 = MO(3)$ represents ozone concentration three days ahead of time.

(v) Combine all preprocessed files (one for each city) one below the other. Each city is represented by a reference location (L_C, G_C) say, city's center.

The total number of patterns in a median approach in each year can be expressed as:

$$N_v = (N_c \cdot N_{days}) \tag{5.2}$$

To avoid overfitting problem and improve generalization performance we can use feature selection. This will reduce number of input features by considering only the significant input features.

Between the two approaches, the tall file data approach has the most patterns and fewer input features compared to the example system. More information is available to the network, so tall file approach may perform better than the median approach.

Chapter 6

Feature Selection

The performance of a network can be bad and consume more processing time if the training data sets are large with high dimensionality. Feature selection aims to solve the dimensionality problem by removing redundant and irrelevant inputs. Feature selection improves network performance by preventing memorization, and by increasing the generalization capability.[43, 45, 48, 93, 94.] Feature selection retains most of the information underlying the data by selecting the optimal subset of available features or inputs thereby reducing noise. In feature selection, features retain their original characteristics as opposed to the transformed features in feature extraction. [49,50,57,60.]

Given training data ($\mathbf{x_p}$, $\mathbf{t_p}$) with N_v patterns and N input features, feature selection finds the best subset of size N_1 that gives minimum training error. A feature selection method requires a subset evaluation function (SEF) *J*, a scalar function of $\mathbf{x}^{\mathbf{N1}}$ that evaluates the effectiveness of a candidate feature subset. The subset generation algorithm (SGA) generates subsets for the SEF. Common types of SEF are mean squared error (MSE)[45-54], feature goodness (FG),[63] Bayes Probability of error,[63] filters,[46,63] SEFs based upon scatter matrices,[63] and wrappers[45,46,63,93,94]. Common SGA types include exhaustive search or brute force or optimal subset method [63], branch and bound method[63], sequential backward selection,[45,63,66] sequential forward selection[45,63,66] and plus - L minus - R method. [63]

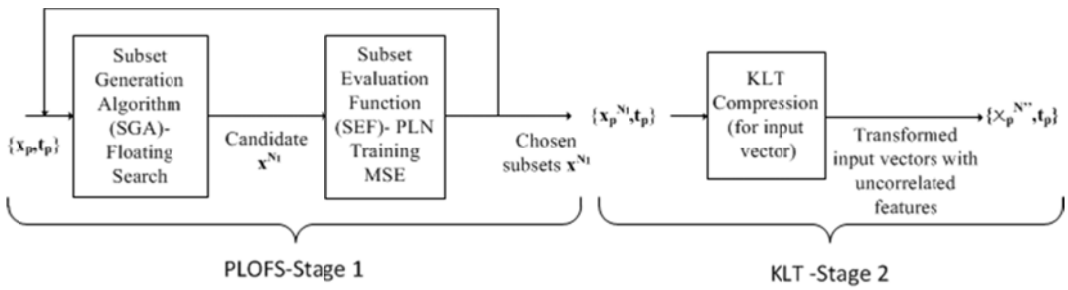
The exhaustive search method gives optimal subsets but it is time consuming and incurs additional computational burden[43, 47, 48, 54]. The branch and bound algorithm proposed by Narendra and Fukunaga results in optimal feature subsets provided the subset evaluation function or the criterion function satisfies monotonicity, a property where error decreases as features increase. Further, as the number of features exceed 30, the branch and bound algorithm becomes unusable.[43, 54] To obtain a near optimal or sub optimal feature set, a sequential backward selection (SBS) or “top down” search method or pruning method and a sequential forward selection (SFS) or “bottom up” search method or growing method could be used. But these suboptimal search methods suffer from nesting. The plus - *l* - minus - *r* method, a suboptimal search method, prevents nesting of feature sets, but this method lacks method for determining the *l* and *r* values. To obtain near optimal feature sets, and reduce the computational complexity involved in high dimensional feature selection problems, Pudil et al used floating search method that dynamically change the number of features included or excluded at each step.[54] The floating search methods namely, the sequential forward floating search (SFFS) and sequential backward floating search (SBBS) methods have better computational performance and yield results comparable to that of branch and bound.[54]

Also, there are transformation methods of feature selection. In a transformation approach[53] to feature selection we look for a transformation

$$\mathbf{z} = f(\mathbf{x}) \qquad (6.1)$$

where $\mathbf{z}$ is a vector of reduced dimension N_1 and $\mathbf{x}$ is the original input feature vector of dimension N. Transformation method of feature selection can result in smaller feature vectors when compared to subset selection methods. However, transformation methods are slower (consume more processing time) as all features of $\mathbf{x}$ are combined to produce vector $\mathbf{z}$. Subset selection methods produce more efficient feature extraction as all the features of $\mathbf{x}$ are not needed for evaluation.[50, 55, 63]

In this study, a two stage feature selection: piecewise linear orthonormal floating search method (PLOFS) followed by Karhunen-Loeve Transform (KLT), is used to help discard redundant input features and retain linearly independent input features. Figure 6-1. shows the two stage feature selection method.



Note: N''< N_1<N

Figure 6-1 Two-stage Feature Selection

6.1 Piecewise linear orthonormal floating search method

In order to obtain optimal feature subsets for a given SEF, one has to use branch and bound. But, branch and bound suffers with combinatorial explosion and involves lot of computational complexity especially when the number of input features is large. To avoid combinatorial explosion and still obtain near optimal subsets coupled with better efficiency and moderate computational complexity, a piecewise linear orthonormal floating search method (PLOFS) [43, 65] is used in this study.

In PLOFS, a forward floating search algorithm, that dynamically adds and removes features predominantly in the forward direction is used as the SGA; and training mean squared error (MSE) that has monotonocity is used as an SEF. A detailed description of PLOFS can be found in [43].

6.2 Karhunen – Loeve Transform (KLT)

In the KLT [55, 56, 57, 60, 67, 79, 91, 92, 96, 97, 98], the transformation kernel is actually derived from the data over which the transformation has to be performed, whereas in transforms such as the Discrete Fourier Transform (DFT), the transformation kernels are fixed and are independent of the data. KL transforms have many applications in data alignment, data compression and object recognition. Previous studies on ozone forecasting using neural networks based on the KL transform can be found in[16, 52, 57, 60, 67, 79, 96, 97, 98].

KL transform or principal component analysis (PCA) can be used to eliminate multi-collinearity (problem associated with highly correlated input features) and generate linearly independent input features. KL transform is basically a rotation transformation that establishes a new coordinate system in such a way that the transformed axes are orthogonal and transformed features are uncorrelated to each other. A detailed description of KL transform can be found in [55, 56, 64].

Chapter 7

Results and Discussion

7.1 Results

Tall file approach described in section 5.1.1 was carried out using five year data (2010-2014) from the 50 monitoring sites selected based on the availability of data from the regions Dallas-Fort Worth (13 sites), Houston-Galveston-Brazoria (12 sites), Los Angeles (10 sites), San Joaquin (8 sites) and San Diego (7 sites). The details of the sites used are described in Appendix B and site maps are shown in Appendix C.

In this approach, the four year data corresponding to the years 2010-2013 were randomly divided in the ratio 3:1 into training and validation data. Random division of data into training and validation sets achieved better results. The data corresponding to the year 2014 was used as testing data. The tall training file formed in this manner had 54050 patterns, tall validation file had 18000 and the tall testing file had 18000 patterns respectively. For comparison, the testing pattern file from the year 2014 of each individual site was used. Each pattern (observation or data point) had 71 input features formed as described in section 5.1.1. Results are shown in the Tables 7-1, 7-4. To reduce the input vector dimension and redundancy, PLOFS and KLT were implemented in stages and the corresponding results are shown in Tables 7-2,7-5 and Table 7-3, 7-6.

Median approach described in section 5.1.2 was carried out and the results are shown in Tables 7-7, 7-8, 7-9 respectively. In median approach, too, the data is randomly divided in the ratio of 3:1 into training and validation data from the years 2010-2013. The 2014 data was used as testing data. Each pattern had 69 input features formed as described in section 5.1.2. Results are shown in the Table 7-7. To reduce the input vector dimension and redundancy, PLOFS and KLT were implemented in stages and the corresponding results are shown in Table 7-8 and Table 7-9.

Table 7-1 Tall file results with all input features

TALL FILE (N=71, M = 3) – Based on Ozone 8hr Average										
Site No.	Station name (region)	Tall file (made with 50 stations from 5 cities)			Individual station file			Tall Method Better? (Yes/No)		
	Dallas Fort Worth (Moderate)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
1	Fort Worth Northwest	8.99	11.16	11.71	9.43	11.75	12.55	Yes	Yes	Yes
2	Arlington Municipal Airport	8.02	10.03	10.72	8.42	10.59	11.51	Yes	Yes	Yes
3	Italy	7.35	8.95	9.59	7.64	9.64	10.42	Yes	Yes	Yes
4	Midlothian	7.32	9.00	9.64	7.67	9.85	10.81	Yes	Yes	Yes
5	Greenville	8.44	10.34	10.83	8.80	10.49	10.92	Yes	Yes	Yes
6	Kaufman	7.44	9.13	9.64	7.63	9.90	10.69	Yes	Yes	Yes
7	Corsicana Airport	7.47	9.19	9.79	7.61	9.69	10.52	Yes	Yes	Yes
8	Eagle Mountain Lake	8.59	10.51	11.05	8.67	10.69	11.56	Yes	Yes	Yes
9	Keller	8.63	10.46	11.03	9.05	11.11	12.07	Yes	Yes	Yes
10	Grapevine Fairway	9.31	11.08	11.48	9.67	11.66	12.56	Yes	Yes	Yes
11	Dallas Executive Airport	7.81	9.75	10.55	7.95	9.95	10.93	Yes	Yes	Yes
12	Dallas Hilton	8.54	10.44	11.05	8.75	10.66	11.56	Yes	Yes	Yes
13	Denton South Airport	9.38	10.95	11.28	9.73	11.43	12.05	Yes	Yes	Yes

Table 7.1—*Continued*

TALL FILE (N=71, M = 3) – Based on Ozone 8hr Average											
Site No.	Station name (region)	Tall file (made with 50 stations from 5 cities)			Individual station file			Tall Method Better? (Yes/No)			
	Houston (Marginal)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
14	Houston Aldine	9.96	11.28	11.90	10.13	11.66	12.49	Yes	Yes	Yes
15	Clinton	9.29	10.98	11.70	9.64	11.51	12.16	Yes	Yes	Yes
16	Conroe (Relocated)	9.78	10.48	10.80	9.73	10.63	11.08	No	Yes	Yes
17	Channel View	9.21	10.97	11.60	9.79	11.53	11.90	Yes	Yes	Yes
18	Galveston 99th Street	9.33	11.68	12.21	9.64	12.11	12.65	Yes	Yes	Yes
19	Houston Bayland Park	9.87	11.66	12.45	10.18	12.22	12.83	Yes	Yes	Yes
20	Houston Deer Park 2	9.01	10.83	11.60	9.36	11.45	12.13	Yes	Yes	Yes
21	Lynchbury Ferry	8.81	10.45	11.12	8.79	10.38	11.04	No	No	No
22	Lake Jackson	8.64	10.46	11.20	8.80	10.95	11.35	Yes	Yes	Yes
23	Northwest Harris	8.98	10.66	11.21	9.40	11.05	11.67	Yes	Yes	Yes
24	Park Place	9.69	11.58	12.16	9.96	12.09	12.70	Yes	Yes	Yes
25	Seabrook Friendship Park	8.70	11.02	11.85	9.07	11.58	12.15	Yes	Yes	Yes

Table 7.1—*Continued*

TALL FILE (N=71, M = 3) – Based on Ozone 8hr Average										
Site No.	Station name (region)	Tall file (made with 50 stations from 5 cities)			Individual station file			Tall Method Better? (Yes/No)		
	Los Angeles (Extreme)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
26	Azusa	8.82	10.16	10.58	8.88	10.69	10.96	Yes	Yes	Yes
27	Compton-700 North Bullis Road	7.16	8.34	8.83	7.49	8.92	9.49	Yes	Yes	Yes
28	Glendora Laurel	9.60	11.62	12.42	9.78	12.26	13.02	Yes	Yes	Yes
29	Lancaster -43301 Division street	8.79	10.16	10.41	8.28	9.72	10.05	No	No	No
30	Los Angeles North Main Street	7.61	8.98	9.62	7.67	9.05	9.62	Yes	Yes	Yes
31	Pasadena S Wilson Avenue	9.87	12.33	13.44	10.61	13.33	14.07	Yes	Yes	Yes
32	Pomona	9.19	11.23	12.39	10.31	13.60	15.55	Yes	Yes	Yes
33	Santa Clarita	9.57	12.21	13.28	9.23	11.62	12.27	No	No	No
34	West Los Angeles-VA Hospital	7.23	8.38	9.06	8.38	12.00	13.91	Yes	Yes	Yes
35	Los Angeles Westchester Parkway	6.69	8.32	9.11	6.81	8.49	9.80	Yes	Yes	Yes

Table 7.1—*Continued*

TALL FILE (N=71, M = 3) – Based on Ozone 8hr Average										
Site No.	Station name (region)	Tall file (made with 50 stations from 5 cities)			Individual station file			Tall Method Better? (Yes/No)		
	San Joaquin (Extreme)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
36	Clovis-N Villa Avenue	9.32	11.47	12.17	9.12	11.62	13.07	No	Yes	Yes
37	Merced S Coffee Avenue	8.38	10.18	10.89	8.17	10.06	10.85	No	No	No
38	Shafter-Walker Street	8.29	9.67	10.05	8.28	9.65	10.10	No	No	Yes
39	Fresno-Sierra Skypark #2	8.83	10.85	11.57	8.65	11.05	12.38	No	Yes	Yes
40	Stockton-Hazelton Street	6.94	8.74	9.29	6.87	8.36	8.88	No	No	No
41	Tracy-Airport	7.48	9.43	9.83	8.88	10.20	10.18	Yes	Yes	Yes
42	Turlock-S Minaret Street	8.00	9.74	10.23	7.81	9.47	9.92	No	No	No
43	Visalia-N Church Street	8.53	10.12	10.53	8.44	10.45	10.87	No	Yes	Yes

Table 7.1—*Continued*

TALL FILE (N=71, M = 3) – Based on Ozone 8hr Average										
Site No.	Station name (region)	Tall file (made with 50 stations from 5 cities)			Individual station file			Tall Method Better? (Yes/No)		
	San Diego (Marginal)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
44	Alpine-Victoria Drive	6.60	8.53	9.32	6.49	8.40	9.47	No	No	Yes
45	Chula Vista	5.31	6.62	7.13	5.55	6.80	7.27	Yes	Yes	Yes
46	El Cajun-Redwood Avenue	6.01	7.58	8.32	7.23	9.06	9.72	Yes	Yes	Yes
47	Escondido-E Valley Parkway	6.52	7.90	8.81	6.68	8.03	8.67	Yes	Yes	No
48	Otay Mesa-Paseo International	5.66	6.98	7.43	6.32	7.43	7.63	Yes	Yes	Yes
49	San Diego-1110 Beardsley Street	6.18	7.26	7.74	6.38	7.57	8.14	Yes	Yes	Yes
50	San Diego - Kearny Villa Road	5.74	7.24	8.04	5.81	7.49	8.10	Yes	Yes	Yes

Table 7-2 Tall file results based on Stage 1 feature selection-PLOFS

Results from **PLOFS** :		TALL FILE (N = 62, M = 3) – Based on Ozone 8hr Average								
Site No.	Station name (region)	Tall file (made with 50 stations from 5 cities)			Individual station file			Tall Method Better? (Yes/No)		
	Dallas Fort Worth (Moderate)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
1	Fort Worth Northwest	8.97	10.92	11.58	9.33	11.67	12.61	Yes	Yes	Yes
2	Arlington Municipal Airport	8.11	10.05	10.82	8.39	10.52	11.54	Yes	Yes	Yes
3	Italy	7.39	9.08	9.81	7.60	9.81	10.57	Yes	Yes	Yes
4	Midlothian	7.36	9.08	9.77	7.69	9.77	10.79	Yes	Yes	Yes
5	Greenville	8.49	10.47	11.09	8.83	10.66	11.23	Yes	Yes	Yes
6	Kaufman	7.46	9.27	10.16	7.75	9.83	10.69	Yes	Yes	Yes
7	Corsicana Airport	7.51	9.43	10.26	7.48	9.33	10.25	No	No	No
8	Eagle Mountain Lake	8.44	10.19	10.95	8.68	10.76	11.64	Yes	Yes	Yes
9	Keller	8.53	10.18	10.93	8.93	11.03	11.91	Yes	Yes	Yes
10	Grapevine Fairway	9.22	10.74	11.37	9.65	11.63	12.53	Yes	Yes	Yes
11	Dallas Executive Airport	7.96	9.90	10.80	7.95	9.96	10.97	No	Yes	Yes
12	Dallas Hilton	8.51	10.28	11.04	8.77	10.66	11.51	Yes	Yes	Yes
13	Denton South Airport	9.26	10.51	11.12	9.78	11.48	12.11	Yes	Yes	Yes

Table 7.2—*Continued*

Results from **PLOFS** : TALL FILE (N=62, M = 3) – Based on Ozone 8hr Average										
Site No.	Station name (region)	Tall file (made with 50 stations from 5 cities)			Individual station file			Tall Method Better? (Yes/No)		
	Houston (Marginal)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
14	Houston Aldine	10.08	11.36	12.01	10.12	11.61	12.45	Yes	Yes	Yes
15	Clinton	9.46	11.23	11.93	9.66	11.58	12.28	Yes	Yes	Yes
16	Conroe (Relocated)	9.91	10.57	10.87	9.65	10.35	10.95	No	No	Yes
17	Channel View	9.44	11.07	11.75	9.55	11.13	11.62	Yes	Yes	No
18	Galveston 99th Street	9.31	11.58	12.24	9.63	12.13	12.64	Yes	Yes	Yes
19	Houston Bayland Park	9.89	11.67	12.28	10.18	12.22	12.85	Yes	Yes	Yes
20	Houston Deer Park 2	9.12	10.92	11.65	9.34	11.44	12.13	Yes	Yes	Yes
21	Lynchbury Ferry	8.87	10.51	11.24	9.41	11.30	11.48	Yes	Yes	Yes
22	Lake Jackson	8.71	10.59	11.19	8.95	10.98	11.88	Yes	Yes	Yes
23	Northwest Harris	9.18	10.73	11.35	9.27	10.94	11.68	Yes	Yes	Yes
24	Park Place	9.77	11.63	12.15	9.89	12.02	12.64	Yes	Yes	Yes
25	Seabrook Friendship Park	8.75	11.03	11.82	9.02	11.55	12.19	Yes	Yes	Yes

Table 7.2—*Continued*

Results from **PLOFS** : TALL FILE (N=62, M = 3) – Based on Ozone 8hr Average										
Site No.	Station name (region)	Tall file (made with 50 stations from 5 cities)			Individual station file			Tall Method Better? (Yes/No)		
	Los Angeles (Extreme)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
26	Azusa	8.92	10.347	10.56	8.60	10.35	10.73	No	Yes	Yes
27	Compton-700 North Bullis Road	7.17	8.23	8.65	7.32	8.79	9.32	Yes	Yes	Yes
28	Glendora Laurel	9.68	11.57	12.26	9.62	11.97	12.70	No	Yes	Yes
29	Lancaster - 43301 Division street	8.43	9.64	9.83	8.34	9.67	10.08	No	Yes	Yes
30	Los Angeles North Main Street	7.66	8.73	9.30	7.68	9.04	9.65	Yes	Yes	Yes
31	Pasadena S Wilson Avenue	9.86	12.30	13.32	10.50	13.17	13.93	Yes	Yes	Yes
32	Pomona	9.36	11.39	12.19	12.01	15.82	17.90	Yes	Yes	Yes
33	Santa Clarita	9.56	11.87	12.55	9.26	11.51	12.18	No	No	No
34	West Los Angeles-VA Hospital	7.31	8.47	9.11	8.27	11.66	13.80	Yes	Yes	Yes
35	Los Angeles Westchester Parkway	6.80	8.76	9.69	6.86	8.64	9.78	Yes	No	Yes

Table 7.2—*Continued*

Results from **PLOFS** :		TALL FILE (N=62, M = 3) – Based on Ozone 8hr Average								
Site No.	Station name (region)	Tall file (made with 50 stations from 5 cities)			Individual station file			Tall Method Better? (Yes/No)		
	San Joaquin (Extreme)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
36	Clovis-N Villa Avenue	9.36	11.35	12.01	9.04	11.46	12.70	No	Yes	Yes
37	Merced S Coffee Avenue	8.41	10.06	10.54	8.15	9.99	10.77	No	No	Yes
38	Shafter-Walker Street	8.36	9.66	9.85	8.27	9.64	10.11	No	No	Yes
39	Fresno-Sierra Skypark #2	8.83	10.87	11.37	8.65	11.08	12.32	No	Yes	Yes
40	Stockton-Hazelton Street	6.77	8.29	8.81	6.89	8.56	9.06	Yes	Yes	Yes
41	Tracy-Airport	7.48	9.10	9.44	7.50	9.16	9.77	Yes	Yes	Yes
42	Turlock-S Minaret Street	7.83	9.49	9.94	7.80	9.46	9.92	No	No	No
43	Visalia-N Church Street	8.46	9.89	10.19	8.31	10.26	10.94	No	Yes	Yes

Table 7.2—*Continued*

Results from **PLOFS** :		TALL FILE (N=62, M = 3) – Based on Ozone 8hr Average								
Site No.	Station name(region)	Tall file (made with 50 stations from 5 cities)			Individual station file			Tall Method Better? (Yes/No)		
	San Diego (Marginal)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
44	Alpine-Victoria Drive	6.95	8.89	9.32	6.47	8.35	9.35	No	No	Yes
45	Chula Vista	5.29	6.53	6.93	5.54	6.89	7.40	Yes	Yes	Yes
46	El Cajun-Redwood Avenue	5.87	7.40	8.10	7.06	9.13	10.12	Yes	Yes	Yes
47	Escondido-E Valley Parkway	6.55	7.90	8.60	6.68	8.12	8.77	Yes	Yes	Yes
48	Otay Mesa-Paseo International	5.67	6.83	7.27	6.27	7.47	7.58	Yes	Yes	Yes
49	San Diego-1110 Beardsley Street	6.30	7.28	7.79	6.43	7.63	8.13	Yes	Yes	Yes
50	San Diego - Kearny Villa Road	5.76	7.20	7.79	6.08	7.96	8.34	Yes	Yes	Yes

Table 7-3 Tall file results based on Stage 2 feature selection (transformation)-KLT

KLT Results: TALL FILE (**N=58**, M = 3) – Based on Ozone 8hr Average										
Site No.	Station name (region)	Tall file (made with 50 stations from 5 cities)			Individual station file			Tall Method Better? (Yes/No)		
	Dallas Fort Worth (Moderate)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
1	Fort Worth Northwest	9.04	11.09	11.68	9.34	11.73	12.51	Yes	Yes	Yes
2	Arlington Municipal Airport	8.06	10.04	10.75	8.43	10.25	11.22	Yes	Yes	Yes
3	Italy	7.47	9.06	9.66	7.70	9.42	10.38	Yes	Yes	Yes
4	Midlothian	7.32	9.04	9.64	7.74	9.80	10.81	Yes	Yes	Yes
5	Greenville	8.49	10.33	10.86	8.57	10.36	11.04	Yes	Yes	Yes
6	Kaufman	7.59	9.20	9.70	7.72	9.71	10.47	Yes	Yes	Yes
7	Corsicana Airport	7.52	9.27	9.83	7.63	9.69	10.53	Yes	Yes	Yes
8	Eagle Mountain Lake	8.49	10.54	11.26	8.68	10.76	11.64	Yes	Yes	Yes
9	Keller	8.58	10.22	10.96	8.99	11.09	12.04	Yes	Yes	Yes
10	Grapevine Fairway	9.35	10.91	11.46	9.65	11.63	12.53	Yes	Yes	Yes
11	Dallas Executive Airport	7.87	9.90	10.67	7.95	9.96	10.97	Yes	Yes	Yes
12	Dallas Hilton	8.69	10.45	11.12	8.74	10.66	11.54	Yes	Yes	Yes
13	Denton South Airport	9.29	10.63	11.18	9.58	11.36	12.11	Yes	Yes	Yes

Table 7.3—*Continued*

KLT Results-	TALL FILE (N=58, M = 3) – Based on Ozone 8hr Average									
Site No.	Station name(region)	Tall file (made with 50 stations from 5 cities)			Individual station file			Tall Method Better? (Yes/No)		
	Houston (Marginal)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
14	Houston Aldine	10.00	11.36	12.03	10.12	11.61	12.45	Yes	Yes	Yes
15	Clinton	9.33	10.95	11.51	9.52	11.57	12.14	Yes	Yes	Yes
16	Conroe (Relocated)	9.88	10.64	10.92	9.61	10.40	10.93	No	No	Yes
17	Channel View	9.29	10.97	11.55	9.55	11.12	11.61	Yes	Yes	Yes
18	Galveston 99th Street	9.29	11.74	12.36	9.63	12.13	12.64	Yes	Yes	Yes
19	Houston Bayland Park	10.02	11.81	12.35	10.18	12.22	12.85	Yes	Yes	Yes
20	Houston Deer Park 2	9.13	10.84	11.56	9.34	11.44	12.13	Yes	Yes	Yes
21	Lynchbury Ferry	8.90	10.50	11.35	8.82	10.38	11.04	No	No	No
22	Lake Jackson	8.58	10.53	11.15	8.91	11.10	11.90	Yes	Yes	Yes
23	Northwest Harris	9.03	10.70	11.34	9.27	10.93	11.68	Yes	Yes	Yes
24	Park Place	9.71	11.54	12.09	9.91	12.08	12.72	Yes	Yes	Yes
25	Seabrook Friendship Park	8.75	11.05	11.96	9.02	11.55	12.20	Yes	Yes	Yes

Table 7.3—*Continued*

KLT Results TALL FILE (N=58, M = 3) – Based on Ozone 8hr Average										
Site No.	Station name(region)	Tall file (made with 50 stations from 5 cities)			Individual station file			Tall Method Better? (Yes/No)		
	Los Angeles (Extreme)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
26	Azusa	8.58	10.29	10.47	8.61	10.36	10.72	Yes	Yes	Yes
27	Compton-700 North Bullis Road	6.92	8.20	8.69	7.43	9.02	9.43	Yes	Yes	Yes
28	Glendora Laurel	9.50	11.74	12.30	9.63	12.04	12.75	Yes	Yes	Yes
29	Lancaster -43301 Division street	9.30	11.17	10.82	7.88	8.90	9.17	No	No	No
30	Los Angeles North Main Street	7.54	8.93	9.39	7.72	9.03	9.57	Yes	Yes	Yes
31	Pasadena S Wilson Avenue	9.73	12.42	13.42	10.44	13.14	13.90	Yes	Yes	Yes
32	Pomona	8.97	11.38	12.44	10.99	13.38	15.43	Yes	Yes	Yes
33	Santa Clarita	9.39	11.89	12.89	9.34	11.69	12.38	No	No	No
34	West Los Angeles-VA Hospital	7.19	8.32	9.05	8.38	11.79	13.91	Yes	Yes	Yes
35	Los Angeles Westchester Parkway	6.68	8.24	9.19	6.92	8.56	9.62	Yes	Yes	Yes

Table 7.3—*Continued*

KLT Results: TALL FILE (N=58, M = 3) – Based on Ozone 8hr Average										
Site No.	Station name(region)	Tall file (made with 50 stations from 5 cities)			Individual station file			Tall Method Better? (Yes/No)		
	San Joaquin (Extreme)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
36	Clovis-N Villa Avenue	9.28	11.38	12.05	9.10	11.87	13.15	No	Yes	Yes
37	Merced S Coffee Avenue	8.25	9.84	10.40	8.04	9.67	10.58	No	No	Yes
38	Shafter-Walker Street	8.51	9.74	10.07	8.27	9.64	10.14	No	No	Yes
39	Fresno-Sierra Skypark #2	10.90	11.85	12.18	8.76	11.56	12.82	No	No	Yes
40	Stockton-Hazelton Street	8.78	9.16	9.39	6.77	8.36	8.99	No	No	No
41	Tracy-Airport	9.29	10.35	10.76	7.50	9.09	9.85	No	No	No
42	Turlock-S Minaret Street	7.87	9.69	10.31	7.84	9.51	10.00	No	No	No
43	Visalia-N Church Street	8.49	10.09	10.36	8.38	10.26	10.87	No	Yes	Yes

Table 7.3—*Continued*

KLT Results - TALL FILE (N=58, M = 3) – Based on Ozone 8hr Average										
Site No.	Station name(region)	Tall file (made with 50 stations from 5 cities)			Individual station file			Tall Method Better? (Yes/No)		
	San Diego (Marginal)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
44	Alpine-Victoria Drive	6.84	8.64	9.25	6.45	8.34	9.35	No	No	Yes
45	Chula Vista	5.31	6.63	7.07	5.39	6.62	7.10	Yes	No	Yes
46	El Cajun-Redwood Avenue	5.92	7.67	8.29	7.15	9.39	10.23	Yes	Yes	Yes
47	Escondido-E Valley Parkway	6.55	8.07	8.87	6.65	7.93	8.53	Yes	No	No
48	Otay Mesa-Paseo International	5.73	7.01	7.54	6.08	7.33	7.52	Yes	Yes	No
49	San Diego-1110 Beardsley Street	6.11	7.13	7.60	6.57	7.74	8.19	Yes	Yes	Yes
50	San Diego - Kearny Villa Road	5.72	7.40	8.05	5.93	7.62	8.13	Yes	Yes	Yes

Table 7-4 Best and poorly predicted sites in each city based on tall file results (with all inputs, N= 71)

No	(City)	Best Predicted site in each city (Testing RMSE, ppb)			Poorly predicted site in each city (Testing RMSE, ppb)		
		One day ahead daily maximum ozone (ppb)	Two day ahead daily maximum ozone (ppb)	Three day ahead daily maximum ozone (ppb)	One day ahead daily maximum ozone (ppb)	Two day ahead daily maximum ozone (ppb)	Three day ahead daily maximum ozone (ppb)
1	Dallas-Fort Worth	Midlothian (7.32)	Italy (8.95)	Italy (9.59)	Denton South Airport (9.38)	Fort Worth North West (11.16)	Fort Worth North West (11.71)
2	Houston	Lake Jackson (8.64)	Lynchbury Ferry (10.45)	Conroe (Relocated) (10.80)	Aldine (9.96)	Galveston 99th Street (11.68)	Houston Bayland Park (12.45)
3	Los Angeles	Westchester Parkway (6.69)	Westchester Parkway (8.32)	Compton (8.83)	Pasadena (9.87)	Pasadena (12.33)	Pasadena (13.44)
4	San Joaquin	Stockton Hazelton (6.94)	Stockton Hazelton (8.74)	Stockton Hazelton (9.29)	Clovis-N Villa Avenue (9.32)	Clovis-N Villa Avenue (11.47)	Clovis-N Villa Avenue (12.17)
5	San Diego	Chula Vista (5.31)	Chula Vista (6.62)	Chula Vista (7.13)	Alpine-Victoria Drive (6.6)	Alpine-Victoria Drive (8.53)	Alpine-Victoria Drive (9.32)

Key Findings from Table 7-4

- Chula Vista (San Diego) is the best predicted site among all the 50 monitoring sites for one day ahead, two day ahead and three ahead daily maximum ozone concentrations.
- Aldine (Houston) is the most poorly predicted site among all the 50 monitoring sites for one day ahead, Pasadena (Los Angeles) is the most poorly predicted site among all the 50 monitoring sites for two day ahead and three ahead daily maximum ozone concentrations.

Table 7-5 Best and poorly predicted sites in each city based on tall file results after stage 1 feature selection (after PLOFS, N =62)

No	**(City)**	**Best Predicted site in each city (Testing RMSE, ppb)**			**Poorly predicted site in each city (Testing RMSE, ppb)**		
		One day ahead daily maximum ozone (ppb)	Two day ahead daily maximum ozone (ppb)	Three day ahead daily maximum ozone (ppb)	One day ahead daily maximum ozone (ppb)	Two day ahead daily maximum ozone (ppb)	Three day ahead daily maximum ozone (ppb)
1	Dallas-Fort Worth	Midlothian (7.36)	Midlothian (9.08)	Midlothian (9.77)	Denton South Airport (9.26)	Fort Worth North West (10.92)	Fort Worth North West (11.58)
2	Houston	Lake Jackson (8.71)	Lynchbury Ferry (10.51)	Conroe (Relocated) (10.87)	Aldine (10.08)	Houston Bayland Park (11.67)	Houston Bayland Park (12.28)
3	Los Angeles	Westchester Parkway (6.80)	Compton (8.23)	Compton (8.65)	Pasadena (9.86)	Pasadena (12.30)	Pasadena (13.32)
4	San Joaquin	Stockton Hazelton (6.77)	Stockton Hazelton (8.29)	Stockton Hazelton (8.81)	Clovis-N Villa Avenue (9.36)	Clovis-N Villa Avenue (11.35)	Clovis-N Villa Avenue (12.01)
5	San Diego	Chula Vista (5.29)	Chula Vista (6.53)	Chula Vista (6.93)	Alpine-Victoria Drive (6.95)	Alpine-Victoria Drive (8.89)	Alpine-Victoria Drive (9.32)

Key Findings from Table 7-5

- Chula Vista (San Diego) is the best predicted site among all the 50 monitoring sites for one day ahead, two day ahead and three ahead daily maximum ozone concentrations.
- Aldine (Houston) is the most poorly predicted site among all the 50 monitoring sites for one day ahead, Pasadena (Los Angeles) is the most poorly predicted site among all the 50 monitoring sites for two day ahead and three ahead daily maximum ozone concentrations.

Table 7-6 Best and poorly predicted sites in each city based on tall file results after stage 2 feature selection (after KLT, N= 58)

No	(City)	Best Predicted site in each city (Testing RMSE, ppb)			Poorly predicted site in each city (Testing RMSE, ppb)		
		One day ahead daily maximum ozone (ppb)	Two day ahead daily maximum ozone (ppb)	Three day ahead daily maximum ozone (ppb)	One day ahead daily maximum ozone (ppb)	Two day ahead daily maximum ozone (ppb)	Three day ahead daily maximum ozone (ppb)
1	Dallas-Fort Worth	Midlothian (7.32)	Midlothian (9.04)	Midlothian (9.64)	Grapevine Fairway (9.35)	Fort Worth North West (11.09)	Fort Worth North West (11.68)
2	Houston	Lake Jackson (8.58)	Lynchbury Ferry (10.50)	Conroe (Relocated) (10.92)	Houston Bayland Park (10.02)	Houston Bayland Park (11.81)	Galveston 99th Street (12.36)
3	Los Angeles	Westchester Parkway (6.68)	Compton (8.20)	Compton (8.69)	Pasadena (9.73)	Pasadena (12.42)	Pasadena (13.42)
4	San Joaquin	Turlock-S Minaret Street (7.87)	Stockton Hazelton (9.16)	Stockton Hazelton (9.39)	Fresno-Sierra Skypark # 2 (10.90)	Fresno-Sierra Skypark # 2 (11.85)	Fresno-Sierra Skypark # 2 (12.18)
5	San Diego	Chula Vista (5.31)	Chula Vista (6.63)	Chula Vista (7.07)	Alpine-Victoria Drive (6.84)	Alpine-Victoria Drive (8.64)	Alpine-Victoria Drive (9.25)

Key Findings from Table 7-6

- Chula Vista (San Diego) is the best predicted site among all the 50 monitoring sites for one day ahead, two day ahead and three ahead daily maximum ozone concentrations.
- Fresno-Sierra Skypark # 2 (San Joaquin) is the most poorly predicted site among all the 50 monitoring sites for one day ahead, Pasadena (Los Angeles) is the most poorly predicted site among all the 50 monitoring sites for two day ahead and three ahead daily maximum ozone concentrations.

Table 7-7 Median file results with all features

Median File : (N=69) - Ozone 8 hour avg										
No	City	Tall City (made with 5 cities) Median file			City Median file			Comparison results		
		1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
1	Dallas	7.37	9.30	9.92	7.53	9.58	10.31	Yes	Yes	Yes
2	Houston	8.36	10.30	11.05	8.55	10.37	11.01	Yes	Yes	No
3	Los Angeles	6.84	8.73	9.35	6.95	8.66	9.19	Yes	No	No
4	San Joaquin	6.35	8.13	8.65	6.22	8.18	9.02	No	Yes	Yes
5	San Diego	5.84	7.52	8.05	5.64	7.12	7.60	No	No	No

Table 7-8 Median file results based on stage 1 feature selection: PLOFS

Median File : PLOFS (N=35) - Ozone 8 hour avg										
No	City	Tall City (made with 5 cities) Median file			City Median file			Comparison results		
		1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
1	Dallas	7.33	9.49	10.20	7.44	9.23	9.94	Yes	No	No
2	Houston	8.05	9.88	10.64	8.41	10.44	11.23	Yes	Yes	Yes
3	Los Angeles	6.72	8.59	9.17	7.03	8.81	9.32	Yes	Yes	Yes
4	San Joaquin	6.15	7.80	8.50	6.19	8.04	8.72	Yes	Yes	Yes
5	San Diego	5.72	7.17	7.59	5.71	7.19	7.68	No	Yes	Yes

Table 7-9 Median file results based on stage 2 feature selection (transformation): KLT

Median File : After PLOFS and KLT (N=33) - Ozone 8 hour avg										
No	City	Tall City (made with 5 cities) Median file			City Median file			Comparison results		
		1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)	1 day ahead (RMSE ppb)	2 day ahead (RMSE ppb)	3 day ahead (RMSE ppb)
1	Dallas	7.22	9.15	9.92	7.53	9.41	10.05	Yes	Yes	Yes
2	Houston	8.29	10.18	10.93	8.33	10.10	10.84	Yes	No	No
3	Los Angeles	6.88	8.64	9.20	7.05	8.87	9.50	Yes	Yes	Yes
4	San Joaquin	6.31	8.05	8.57	6.21	8.11	8.87	No	Yes	Yes
5	San Diego	5.91	7.48	7.90	5.69	7.26	7.80	No	No	No

The number of exceedance days in Los Angeles, San Joaquin Valley, and San Diego region (California state) during the period of 2010-2014 is listed in Table 7-10. Table 7-11, and Table 7-12 show the statistical properties of all the input variables considered in this study from all the five regions.

Table 7-10 Number of ozone exceedance days (National 8-hour ozone) in California

	Number of exceedance days (National 8-hr ozone)				
	2010	2011	2012	2013	2014
Los Angeles					
Azusa	3	12	10	6	11
Compton-700 North Bullis Road	4	0	0	1	2
Glendora Laurel	20	30	45	24	38
Lancaster -43301 Division street	45	53	39	34	36
Los Angeles North Main Street	1	0	1	0	2
Pasadena S Wilson Avenue	3	5	9	0	7
Pomona	4	16	15	15	33
Santa Clarita	23	31	57	40	45
West Los Angeles-VA Hospital	1	0	0	0	4
Los Angeles Westchester Parkway	0	0	0	1	3
Total number of exceedance days	**104**	**147**	**176**	**121**	**181**
San Joaquin					
Clovis-N Villa Avenue	39	49	57	38	56
Merced S Coffee Avenue	14	19	9	31	22
Shafter-Walker Street	22	18	30	5	11
Fresno-Sierra Skypark #2	35	45	19	25	32
Stockton-Hazelton Street	2	0	2	0	1
Tracy-Airport	3	8	16	2	8
Turlock-S Minaret Street	10	17	35	14	12
Visalia-N Church Street	34	17	37	2	10
Total number of exceedance days	**159**	**173**	**205**	**117**	**152**
San Diego					
Alpine-Victoria Drive	12	10	7	6	10
Chula Vista	2	0	1	0	0
El Cajon-Redwood Avenue	3	1	0	1	0
Escondido-E Valley Parkway	3	2	0	0	5
Otay Mesa-Paseo International	0	1	0	0	0
San Diego-1110 Beardsley Street	0	0	0	0	0
San Diego - Kearny Villa Road	0	1	1	0	1
Total number of exceedance days	**20**	**15**	**9**	**7**	**16**

Table 7-11 Statistical properties of annual hourly pollutant and meteorological parameters in five regions

	Nitric oxide (ppb)	Nitrogen dioxide (ppb)	Ozone (ppb)	Resultant wind speed (mile/hour)	Solar Radiation (Langleys/min)	Temperature (^{o}F)
Dallas Fort Worth 2010						
Mean	1.9278	7.4543	25.157	3.6547	0.278	60.9
Minimum	0	0	0	0	0	0
Maximum	304.2	72.8	108	30.2	1.53	106.9
Standard Deviation	7.632	7.608	16.008	5.09	0.394	24.125
Dallas Fort Worth 2011						
Mean	1.8344	7.2766	28.114	8.07	0.288	68.28
Minimum	0	0	0	0.2	0	10.6
Maximum	272.8	70	101	33.4	1.54	110.4
Standard Deviation	7.5144	7.708	17.021	4.3817	0.412	19.743
Dallas Fort Worth 2012						
Mean	1.9424	6.7413	27.552	7.4696	0.28	68.457
Minimum	0	0	0	0.1	0	18.1
Maximum	215	70.1	122	31.6	1.51	108.4
Standard Deviation	7.608	7.18	16.813	4.052	0.3977	16.32
Dallas Fort Worth 2013						
Mean	1.618	6.4	28.004	7.5	0.269	65.32
Minimum	0	0	0	0	0	15.9
Maximum	342.1	63.4	112	30.1	1.763	105.5
Standard Deviation	6.887	6.872	17.394	4.25	0.394	18.076
Dallas Fort Worth 2014						
Mean	1.337	5.57	28.123	7.9864	0.265	64.778
Minimum	0	0	0	0.1	0	12
Maximum	215.4	62.8	105	29.4	1.549	102.8
Standard Deviation	6.123	6.256	16.165	4.33	0.384	17.838

Table 7-11 —*Continued*

	Nitric oxide (ppb)	Nitrogen dioxide (ppb)	Ozone (ppb)	Resultant wind speed (mile/hour)	Solar Radiation (Langleys/minute)	Temperature (^{o}F)
Houston 2010						
Mean	3.2364	7.876	24.488	5.228	0.272	69.37
Minimum	0	0	0	0	0	17.8
Maximum	479.9	111.9	127	20.8	1.545	100.6
Standard Deviation	12.107	7.9678	18.137	3.277	0.386	15.288
Houston 2011						
Mean	3	8.64	24.94	5.8178	0.294	71.012
Minimum	0	0	0	0	0	18
Maximum	366.2	53.9	129	17.8	1.926	107.8
Standard Deviation	11.42	7.315	17.63	3.275	0.407	15.199
Houston 2012						
Mean	3.03	7.5	23.364	4.99	0.266	71.8
Minimum	0	0	0	0	0	29.2
Maximum	289.7	134.7	112	22.9	1.57	102.7
Standard Deviation	10.573	7.58	17.256	3.09	0.385	12.36
Houston 2013						
Mean	2.883	7.116	22.345	5.39	0.265	69.18
Minimum	0	0	0	0	0	28.2
Maximum	350.5	68	102	22.4	1.557	103.2
Standard Deviation	11.176	7.44	16.175	3.183	0.388	14.322
Houston 2014						
Mean	2.66	6.866	22.16	5.28	0.254	68.495
Minimum	0	0	0	0	0	19.5
Maximum	351.9	97.8	93	19.8	1.64	98.7
Standard Deviation	9.526	7.163	15.52	2.99	0.376	14.26

Table 7-11 —*Continued*

	Nitric oxide (ppb)	Nitrogen dioxide (ppb)	Ozone (ppb)	Resultant wind speed (mile/hour)	Solar Radiation (Langleys/minute)	Temperature (°F)
Los Angeles 2010						
Mean	11.864	17.779	25.534	4.967	0.284	61.528
Minimum	0	0	0	2.01	0	9
Maximum	440	97	104	48.094	3.3475	113
Standard Deviation	27.154	12.677	16.738	3.1746	0.4348	11.904
Los Angeles 2011						
Mean	11.852	17.624	24.776	4.5536	0.307	61.744
Minimum	0	0	0	2.0132	0	23
Maximum	417	110	111	70.016	1.5878	105
Standard Deviation	27.914	12.594	17.585	3.1545	0.429	11.72
Los Angeles 2012						
Mean	9.995	16.3	24.799	4.2267	.224	63.479
Minimum	0	0	0	0	0	0
Maximum	377	82	134	55.923	1.188	106.3
Standard Deviation	23.591	11.87	18.746	2.787	0.308	12.012
Los Angeles 2013						
Mean	9.87	16.207	27.38	4.37	0.31	62.59
Minimum	0	0	0	0	0	1
Maximum	388	90	115	53.91	1.586	110
Standard Deviation	24.051	12.023	17.09	2.84	0.4317	12.109
Los Angeles 2014						
Mean	8.067	15.41	28.66	4.2878	0.322	65.778
Minimum	0	0	2	0	0	28
Maximum	341	89	117	65.99	1.6694	104
Standard Deviation	20.16	11.89	17.78	2.7997	0.442	11.082

Table 7-11 —*Continued*

	Nitric oxide (ppb)	Nitrogen dioxide (ppb)	Ozone (ppb)	Resultant wind speed (mile/hour)	Solar Radiation (Langleys/minute)	Temperature (^{o}F)
San Joaquin 2010						
Mean	4.936	10.18	30.084	8.83	0.294	61.411
Minimum	0	0	0	0	0	25
Maximum	258	82	133	38.028	1.59	108
Standard Deviation	12.07	7.645	22.327	6.76	0.419	14.56
San Joaquin 2011						
Mean	6.1	10.574	31.386	8.5	0.301	60.848
Minimum	0	0	0	0	0	4
Maximum	683	62	133	49.213	1.57	106
Standard Deviation	15.129	7.922	23.082	6.74	0.42	15.603
San Joaquin 2012						
Mean	5.51	10.41	34.3	9.14	0.028	64.059
Minimum	0	0	1	0	0	23
Maximum	282	78	124	46.976	1.05	108
Standard Deviation	13.57	7.902	23.475	6.65	0.13	15.317
San Joaquin 2013						
Mean	6.667	10.853	31.714	8.874	0.323	63.88
Minimum	0	0	1	0	0	23
Maximum	219	118	123	40.265	1.659	108
Standard Deviation	15.814	8.669	22.972	6.62	0.434	15.845
San Joaquin 2014						
Mean	4.91	9.951	34.08	8.7	0.33	65.6
Minimum	0	0	0	0	0	5
Maximum	234	67	119	38.699	2.06	107
Standard Deviation	12.382	8.17	23.68	6.01	0.45	14.548

Table 7-11 —Continued

	Nitric oxide (ppb)	Nitrogen dioxide (ppb)	Ozone (ppb)	Resultant wind speed (mile/hour)	Solar Radiation (Langleys/minute)	Temperature (^{o}F)
San Diego 2010						
Mean	8.25	13.548	40.75	5.786	0.305	61.344
Minimum	0	0	1	0	0	29
Maximum	390	91	105	49.2	2.09	109
Standard Deviation	19.72	10.39	14.01	6.03	0.433	9.32
San Diego 2011						
Mean	8.46	12.988	40.927	4.38	0.322	61.56
Minimum	0	0	1	0	0	30
Maximum	405	100	114	42.5	1.927	102
Standard Deviation	20.14	10.5	14.647	5.2	0.448	9.85
San Diego 2012						
Mean	7.37	12.465	40.814	4.06	0.31	63.09
Minimum	0	0	2	0	0	31
Maximum	447	77	101	24.6	1.66	106
Standard Deviation	18.467	10.053	14.62	4.55	0.437	10.424
San Diego 2013						
Mean	7.26	12.224	42.545	4.433	0.311	62.88
Minimum	0	0	5	0	0	30
Maximum	501	91	95	22.369	1.51	102
Standard Deviation	18.987	10.677	13.554	4.62	0.43	10.27
San Diego 2014						
Mean	4.66	10.722	43.944	4.374	0.317	66.043
Minimum	0	0	5	0	0	23
Maximum	329	87	92	34.67	1.52	108
Standard Deviation	13.538	9.918	13.872	4.7	0.43	10.263

Table 7-12 Statistical properties of annual hourly pollutant and meteorological parameters in five regions used in training and validation averaged over the years (2010-2013).

	Nitric oxide (ppb)	Nitrogen dioxide (ppb)	Ozone (ppb)	Resultant wind speed (mile/hour)	Solar Radiation (Langleys/minute)	Temperature (^{o}F)
Dallas Fort Worth						
Mean	1.83	6.97	27.21	6.67	0.28	65.74
Minimum	0	0	0	0.08	0	11.15
Maximum	283.52	69.08	110.75	31.33	1.59	107.80
Standard Deviation	7.41	7.34	16.81	4.44	0.40	19.57
Houston						
Mean	3.04	7.78	23.79	5.36	0.27	70.34
Minimum	0	0	0	0	0	23.3
Maximum	371.57	92.13	117.50	20.98	1.65	103.57
Standard Deviation	11.32	7.57	17.30	3.21	0.39	14.29
Los Angeles						
Mean	10.90	16.98	25.62	4.53	0.28	62.34
Minimum	0	0	0	1.01	0	7.25
Maximum	405.5	94.75	116.0	56.99	1.95	108.58
Standard Deviation	25.68	12.29	17.54	2.99	0.51	11.94
San Joaquin						
Mean	5.81	10.51	31.87	8.84	0.24	62.55
Minimum	0	0	0.5	0	0	18.75
Maximum	360.50	85.00	128.25	43.62	1.47	107.50
Standard Deviation	14.15	8.03	22.96	6.69	0.35	15.33
San Diego						
Mean	7.84	12.81	41.26	4.67	0.31	62.22
Minimum	0	0	2.25	0	0	30.0
Maximum	435.75	89.75	103.75	34.67	1.8	104.75
Standard Deviation	19.33	10.42	14.21	5.12	0.44	9.97

Conclusions based on Tall file approach results (Table 7-1, 7-2, and 7-3), Table 7-11, Table 7-12, and Appendix C (monitoring site maps):

- Monitoring sites from Los Angeles and San Joaquin (that are classified as extreme based on 8-hour ozone non attainment classifications by USEPA) were poorly predicted when compared to the predicted sites in San Diego (marginal), Houston (marginal) and Dallas Fort Worth (moderate). See Tables 7-1 to Table 7-3. This indicates that extreme non attainment areas might not be properly predicted using the tall file approach.
- Based on the tall file results and annual ozone summary data (from California Environmental Protection Agency Air Resources Board), shown in Table 7-10, the sites from Los Angeles have more exceedance days based on 8-hour ozone national standard (75 ppb) in the year 2014 that was used as testing data, compared to the years used to develop the model (2010-2013). This could explain why the neural network model did not perform as well in predicting ozone concentrations for Los Angeles.
- The average of standard deviation of annual ozone concentration from 2010-2013 from SanJoaquin (22.96 ppb) is high. This indicates large variability in ozone concentration across the monitoring sites in the San Joaquin Valley region was and a possible reason for poor ozone prediction results.
- The number of exceedance days in 2014 in Los Angeles county sites are: Azusa (11 days), Glendora Laurel (38 days), Lancaster Lancaster - 43301 Division Street (17 days), and Santa Clarita (45 days). These sites showed poorer performance compared to other sites , Pasadena S- Wison Avenue (7 days), West Los Angeles-VA Hospital (4 days), Compton (2 days), and North Main Street (2 days) in Los Angeles. Only Pomona was an exception (33 days).

- The number of exceedance days in 2014 in San Joaquin valley sites are: Clovis-N Villa Avenue (56 days), Fresno-Sierra Skypark # 2 (32 days), Shafter Walker Street (11 days) , Merced - S Coffee Avenue (22 days), Turlock-S Minaret Street (12 days), Visalia- N Church (10 days). These sites showed poorer performance compared to other sites, Tracy Airport (8 days) and Stockton-Hazelton (1 day) in San Joaquin valley.
- The number of exceedance days in 2014 in San Diego sites are Alpine- Victoria Drive (10 days), and Escondido- E Valley Parkway (5 days). These sites showed poorer performance compared to other sites, Kearny Villa Road (1 days), Chula Vista (0 days), El Cajon-Redwood Avenue (0 days), 1110 Beardsley Street (0 days), and Otay Mesa-Paseo International (0 days) in San Diego.
- The following sites that were relatively far from the remaining sites in the respective regions and where the tall file approach did not work better are as follows (See Appendix C – Monitoring site maps):
    - Corsicana airport (Dallas Fort Worth region)
    - Conroe (Relocated) (Houston-Galveston-Brazoria region)
    - Lancaster – 43301 Division Street, Santa Clarita (Los Angeles)
    - Escondido-E Valley Parkway (San Diego)
- The comprehensive ground level ozone forecasting model using the neural network MLP-HWOMOLF should be able to be used for any region with meteorological parameters falling in the ranges given in Table 7.12, or with NO values up to 436 ppb, NO_2 values up to 95 ppb, ozone values up to 128 ppb, wind speeds up to 57.0 mph, solar radiation up to 1.95 Langleys/minute, and temperatures ranging from 7 to 109°F.

7.2 Comparison Work

Comparison work was done to show the better performance of the MLP-HWOMOLF over other neural networks.

Comparison work 1 – Riuz's approximate methodology

Description of methodology adopted by Riuz:

- Number of inputs: 10 (daily mean values of input variables of temperature, NO, NO_x, NO_2, ozone, wind speed, wind direction, solar radiation, carbon monoxide of the previous day, previous day daily maximum ozone).
- Data collected from the years 1999, 2000, 2002, 2003, and 2004.
- No temporal inputs included and missing values removed.
- Data from all the years randomly divided into training, validation, and testing in the ratio (65:5:30) five times and best results noted.

	Riuz's Methodology (MLP-LM) (Neural Network Toolbox) Testing RMSE(ppb)	HWO-MOLF (Testing RMSE) (ppb)
1	Testing RMSE: 16.37	Testing RMSE: 15.03
2	Testing RMSE: 15.029	Testing RMSE: 14.727
3	Testing RMSE: 15.874	Testing RMSE: 14.89
4	Testing RMSE: 14.93	Testing RMSE: 14.28
5	Testing RMSE: 14.38	Testing RMSE: 13.85

Description of methodology 2: (difference in preprocessing the data)).

- Inclusion of temporal variables in continuous format
- $\mathrm{Cos}(\frac{2\pi}{365} \times DOY)$, $\mathrm{Sin}(\frac{2\pi}{365} \times DOY)$ for non leap year and $\mathrm{Cos}(\frac{2\pi}{366} \times DOY)$, $\mathrm{Sin}(\frac{2\pi}{366} \times DOY)$ for leap year. $\mathrm{Cos}(\frac{2\pi}{7} \times DOW)$, $\mathrm{Sin}(\frac{2\pi}{7} \times DOW)$ that represent the day of the week. ($X_{WS} \cdot \mathrm{Cos}(X_{WD})$), and ($X_{WS} \cdot \mathrm{Sin}(X_{WD})$) to account for continuity in wind speed and direction.
- Data collected from the years 1999, 2000, 2002, 2003, and 2004.

- Linearly interpolating missing values and using lagged inputs.
- Data from all the years randomly divided into training, validation, and testing in the ratio (65:5:30) five times and best results noted

	Riuz's Methodology (MLP-LM) (Neural Network Toolbox) Testing RMSE(ppb)	HWO-MOLF (Testing RMSE) (ppb)
1	Testing RMSE : 17.6918	Testing RMSE : 17.08
2	Testing RMSE : 26.85	Testing RMSE : 15.14
3	Testing RMSE : 20.51	Testing RMSE : 16.46
4	Testing RMSE : 20.2731	Testing RMSE : 15.39
5	Testing RMSE : 17.1756	Testing RMSE : 14.49

Description of methodology 3: (difference in preprocessing the data and data division)

- Inclusion of temporal variables in continuous format
- Co$s(\frac{2\pi}{365} \times DOY)$, Si$n(\frac{2\pi}{365} \times DOY)$ for non leap year and Co$s(\frac{2\pi}{366} \times DOY)$, Si$n(\frac{2\pi}{366} \times DOY)$ for leap year. Cos($\frac{2\pi}{7} \times DOW$), Sin($\frac{2\pi}{7} \times DOW$) that represent the day of the week. ($X_{WS} \cdot$ Cos(X_{WD})), and ($X_{WS} \cdot$ Sin(X_{WD})) to account for continuity in wind speed and direction.
- Data collected from the years 1999, 2000, 2002, 2003, and 2004.
- Linearly interpolating missing values and using lagged inputs.
- Data from all the years randomly divided into training, validation, and testing in the ratio (3:1:1) five times and best results noted.

	Riuz's Methodology (MLP-LM) (Neural Network Toolbox) Testing RMSE(ppb)	HWO-MOLF (Testing RMSE) (ppb)
1	Testing RMSE : 22.20	Testing RMSE : 17.01
2	Testing RMSE : 17.57	Testing RMSE : 16.03
3	Testing RMSE : 17.32	Testing RMSE : 14.278
4	Testing RMSE : 15.06	Testing RMSE : 15.62
5	Testing RMSE : 23.53	Testing RMSE : 16.58

Description of methodology 4: (difference in preprocessing the data and data division and using feature selection)

- Inclusion of temporal variables in continuous format

$Cos(\frac{2\pi}{365} \times DOY)$, $Sin(\frac{2\pi}{365} \times DOY)$ for non leap year and $Cos(\frac{2\pi}{366} \times DOY)$, $Sin(\frac{2\pi}{366} \times DOY)$ for leap year. $Cos(\frac{2\pi}{7} \times DOW)$, $Sin(\frac{2\pi}{7} \times DOW)$ that represent the day of the week.

- $(X_{WS} \cdot Cos(X_{WD}))$, and $(X_{WS} \cdot Sin(X_{WD}))$ to account for continuity in wind speed and direction.
- Data collected from the years 1999, 2000, 2002, 2003, and 2004.
- Linearly interpolating missing values, using lagged inputs.
- Training and validation data is made up in the ratio (3:1) by randomly dividing data from the years 1999, 2000, 2002, 2003. The testing data is 2004 data.

	Riuz's Methodology (MLP-LM) (Neural Network Toolbox) Testing RMSE(ppb)	HWO-MOLF (Testing RMSE) (ppb)
1	Testing RMSE : 13.44	Testing RMSE : 11.618
2	Testing RMSE : 19.84	Testing RMSE : 11.952
3	Testing RMSE : 13.78	Testing RMSE : 11.63
4	Testing RMSE : 16.4	Testing RMSE : 11.623
5	Testing RMSE : 19.307	Testing RMSE : 11.632

Comparison work 2 – Prybutok's approximate methodology

Description of methodology adopted by Prybutok:

- Variables considered: hourly ozone, carbon dioxide (CO_2), nitric oxide (NO), nitrogen dioxide (NO_2), oxides of nitrogen (NO_x), temperature, wind speed and wind direction during the period (June 1-Oct 10) 1994, near Aldine, Houston (chosen). No temporal inputs included.

- Hourly data collected from the summer months (June 1, 1994-September 30, 1994) divided into training and validation in the ratio 4:1 randomly.
- Testing data (October 1- October 10, 1994).

- Algorithm used: multilayer perceptron based on Levenberg Marquardt (MLP-LM)
- Inputs in the pattern file :

  (1) X_1 = dummy variable (holidays vs working days)

  (2) X_2 = hourly ozone level at 9 am

  (3) X_3 = actual maximum daily temperature

  (4) X_4 = average concentration of CO_2 between 6:00 am and 9:00 am on the day of interest

  (5) X_5 = average concentration of NO between 6:00 am and 9:00 am on the day of interest

  (6) X_6 = average concentration of NO_2 between 6:00 am and 9:00 am on the day of interest

  (7) X_7 = average concentration of oxides of nitrogen between 6:00 am and 9:00 am on the day of interest

  (8) X_8 = average concentration of wind speed between 6:00 am and 9:00 am on the day of interest

  (9) X_9 = average concentration of wind direction between 6:00 am and 9:00 am on the day of interest
- Output: daily maximum ozone.

Results: One day ahead daily maximum ozone prediction

Neural network	One day ahead daily maximum ozone (RMSE)
MLP- LM (Neural Network Tool box)	19.57 ppb
MLP (HWO-MOLF)	18.76 ppb

After applying feature selection (PLOFS) results improved as shown below

Neural network	One day ahead daily maximum ozone (RMSE)
MLP- LM (Neural Network Toolbox)	16.58 ppb
MLP (HWO-MOLF)	16.29 ppb

Conclusion: Our methodology and MLP-HWOMOLF improve results and perform better than Prybutok's methodology and MLP-LM. Use of KLT did not help both the networks.

Comparison work 3 – Comrie's approximate methodology

- Inputs considered (1 hour ozone maximum from previous day, daily maximum temperature, average daily dew point temperature, average daily wind speed, mean UV radiation. No time inputs used.
- May- September data from 1991-1995, DeKalb Jr. College, Atlanta, Georgia.
- Data randomly divided in the ratio of 50:15:35 without following chronological order

Neural network	One day ahead daily maximum ozone (RMSE)
MLP- LM (Neural Network Toolbox)	18.24 ppb
MLP (HWO-MOLF)	17.88 ppb

Improved methodology after inclusion of time inputs and PLOFS

Neural network	One day ahead daily maximum ozone (RMSE)
MLP- LM (Neural Network Toolbox)	16.67 ppb
MLP (HWO-MOLF)	16.545 ppb

Conclusion: Our methodology and MLP-HWOMOLF improve results and perform better than Comrie's methodology and MLP-LM. Use of KLT did not help both the networks.

Chapter 8

Final Conclusions & Future Work

8.1 Final Conclusions

This work represents the first neural network developed to forecast ozone in multiple regions, as well as at multiple sites in the same region. Previous studies have developed separate neural network models to forecast ozone at each location. The following conclusions can be drawn from the work presented here:

- Tall file approach helps better prediction as it helped most of the monitoring sites.
- The tall file approach didn't perform well in Los Angeles (based on the results from PLOFS) and San Joaquin (based on the results without feature selection, PLOFS and KLT). Both Los Angeles and San Joaquin are designated as "extreme" ozone non-attainment areas by EPA indicating that the current model might not predict well for extreme ozone pollutant levels. The results could be improved if more stations from these two cities are included in making the tall data files.
- The comprehensive ground level ozone forecasting model using the neural network MLP-HWOMOLF with the aid of two stage feature selection (PLOFS and KLT) could predict
  - ✓ one day ahead daily maximum ozone in the range of 5.29 ppb to 10.9 ppb.
  - ✓ two day ahead daily maximum ozone in the range of 6.53 ppb to 12.42 ppb.
  - ✓ three day ahead daily maximum ozone in the range of 6.93 ppb to 13.44 ppb.
- Median approach cannot be site specific and might not be reliable as they do not truly represent any particular site.

- MLP-HWOMOLF proves to be better network compared to other networks trained with different algorithms based on the comparison work.

8.2 Recommendations for Future Work

The following are recommendataions for future work:

- To determine the statistical significance of the tall file approach results.
- To check the prediction performance of the tall file ozone neural network forecasting system based on testing data from a site not used in the training of the tall file approach (i.e., testing data from a monitoring site not picked from these 50 sites in this study).
- To evaluate whether the inclusion of more San Joaquin sites in model development help improve prediction performance in the San Joaquin region.
- To compare the developed NN model to current models (e.g., TCEQ regression models) used for ozone forecasting.
- To develop similar neural network forecasting models for other pollutants.

Appendix A

Literature Review

Author(s) (Year)	Study Location	Air Pollutant(s) Modeled	Predictor Variables	Years of data	Model	Model Performance Comparison
Sekar et al. (2015)[103]	Pollutant data from a heavy traffic intersection, and meteorological data from Safdarjung station in Delhi, India.	Hourly ozone(O_3), oxides of nitrogen (NO_X)	O_3, NO_X, traffic data, atmospheric pressure (P), temperature (OT), wind speed (WS) wind direction (WD), cloud cover (CC), sunshine, rainfall, stability class, mixing height, visibility, solar insolation, temporal variables: day of the week and time of the day.	2008 – 2010	Multilayer perceptron using Levenberg-Marquardt (MLP-LM) Algorithm, Decision tree algorithms: reduced error pruning tree (REPTree), and M5 P tree.	MP 5 tree performed better than MLP-LM and REPTree.
de Souza et al. (2015)[110]	Campo Grande, Brazil	Hourly O_3	O_3, maximum OT, RH, WS, and precipitation.	2004 – 2010	Multilayer perceptron using back propagation (MLP-BP)	
Biancofiore et al. (2014)[101]	Pescara in Central Italy	Hourly O_3 (up to 48 hours)	O_3, nitrogen dioxide (NO_2), OT, relative humidity (RH), WS, WD and ultraviolet radiation.	2005	Recurrent MLP (ELMAN network), multiple linear regression (MLR)	ELMAN recurrent network performed better than MLR.
Tamas et al. (2014)[104]	Urban and suburban stations (Canetto, Sposata) in Ajaccio, and (Giraud, Montesoro) in Bastia from the French island of Corsica, France	24 hour ahead O_3	O_3, NO_2, wind force, WD, global SR, OT, precipitation, and hour of the day. Cos (2πh/24), sin (2πh/24), and weekday number.	2008 – 2012	MLP-LM, persistence models	MLP-LM performed better than persistence models.
Luna et al. (2014)[67]	Mobile automatic monitoring	Hourly O_3	Hourly O_3, nitric oxide (NO), NO_x and NO_2, solar radiation (SR), scalar WS,	2011 and 2012.	MLP-LM, Support Vector Machines (SVMs) SVMs. PCA	SVM’s and MLP-LM performance was remarkably

	station at Pontifical Catholic University, and Rio de Janeiro State University in the city of Rio de Janeiro, Brazil.		carbon monoxide (CO), moisture content in the air.		was used for dimension reduction.	close.
Zahedi (2014)[69]	Mobile station at Shuaiba industrial area in Kuwait.	O_3	O_3, WS, WD, RH, OT, SR, methane, CO, CO_2, NO, NO_2, SO_2, non-CH_4 hydrocarbons, dust.	March and April 1995.	(Sugeno-Takagi-Gang fuzzy inference and hybrid algorithm), and MLP-BP.	(Sugeno-Takagi-Gang fuzzy inference and hybrid algorithm) performed better than MLP-BP.
Alkasassbeh (2013)[68]	Chenbagaramanputhur in Kanyakumari district, India.	O_3	Seven readings per day for ozone, two readings per day for NO_2.	May–July 2009.	Radial Basis Function (RBF), SVMs, MLP-BP.	RBF's performed better than SVM'S and SVM's performed better than MLP-BP.
Arhami et al. (2013)[71]	Fatemi Station, Iran.	Hourly CO, NO_X, NO_2, NO, O_3, particulate matter of 10 μm (PM_{10})	Hourly CO, NO_X, NO, NO_2, O_3, PM_{10}, air OT, wet bulb OT, CC, RH, WS, WD, P, vapor pressure, visibility code, and temporal variables cos (2πh/24), cos (2πm/12), where h = hour of the day, m = month of the year.	2009	MLP-BP coupled with Monte Carlo simulation.	
Pires et al. (2012) [102]	Oporto, North Portugal	One day ahead hourly average O_3	Hourly CO, NO, NO_2, O_3, OT, RH, SR, WS.	May–August 2004	MLP-BP aided with Genetic Algorithms (GA)	
Kandya et al. (2012)[86]	Monitoring site located at Indian Institute	8-hourly averaged O_3.	8-hourly averaged values O_3, NO, NO_2, SO_2, CO, P, respirable suspended	Sept. 2008 –	MLP-BP	

	of Madras Madras, India.		particulate matter, hydrocarbons, WS, WD, solar intensity.	March 2010.		
Paoli et al. (2011)[72]	Suburban station at Sposata located near Ajaccio on the island of Corsica, France	One hour ahead O_3.	Hourly O_3, NO_2, WS, WD, OT, RH, hour of the day, and day of the month, and month of the year.	October 2007 – May 2010.	MLP-LM	
Taormina et al. (2011)[73]	Pollutant data from Harlington station, London Hillington-Harlington (Heathrow airport zone) and meteorological daily data from a monitoring station located in Heathrow airport	Daily maximum hourly O_3	Hourly CO, NO, NO_2, NO_X, O_3, SR.	2004 – 2009.	MLP-LM, Persistence method	Results improved when the optimal model returned by MLP-LM was employed on test data after dynamically updating the weights using an adaptive neural network based on back propagation. Also, forecasting prediction was better than persistence model.
Ibarra-Berastegi et al. (2009)[75]	Six locations in Bilbao, Spain	Hourly SO_2, CO, NO_2, NO, and up to 8 hour ahead eight O_3.	Hourly traffic data, WS, WD, pollutants- SO_2, CO, NO_2, NO, O_3	2000 and 2001.	MLP-BP, MLP-LM, and generalized regression neural network (GRNN), linear models, persistence models	Performance of 4 out of 24 cases showed that persistence models outperformed other models, 13 out of 24 cases linear models performed better

						than any other model, in 6 out of 24 cases non-linear models performed better, and in one case, RBF outperformed other models.
Salcedo-Sanz et al. (2009)[30]	27 monitoring stations in Madrid, Spain	Hourly O_3, NO_X	Hourly O_3, NO_X.	2002 – 2007.	Gaussian RBF and evolutionary based RBF.	Evolutionary based RBF showed better performance compared to other RBFs. Results from evolutionary based RBFs used as initial points in developing Land Use regression models with the aid of GIS.
Coman et al. (2008)[58]	Prunay, Aubervilliers stations, Paris, France.	Hourly O_3 for a 24-hour horizon	Hourly O_3, NO_2, RH, T, SR, sunshine duration, WS, sin(2πh/24), cos(2πh/24).	August 2000 –July 2001.	A “static” MLP (A single MLP) and a “dynamic” MLP (a cascade of 24 MLPs, each MLP feeds the following MLP) based on two training algorithms (MLP-SCG) and limited memory Broyden, Fletcher, Goldfarb, and Sahanno (BFGS) quasi-Newton.	Static model performed little bit better than dynamic model and persistence model. Results showed similar levels of performance when the two trained with MLP-SCG and MLP- BFGS algorithms.
Salazur-	Mexicali	Daily	Daily mean, and mean of	1999	A persistence	Prediction

Ruiz et al (2008)[22]	(Mexico)-Calexico (California, US) border area	maximum hourly O_3.	first six hours of the day of O_3, OT, NO_2, NO, CO, resultant WS, and RH.	– 2004 (excluding 2001)	model, multilinear regression model, semi parametric ridge regression model, a MLP-BP model, an ELMAN recurrent neural network model and an SVM model.	performance of the artificial intelligence (AI) based models was better than the linear models, and among the AI based models, MLP-BP showed better performance than the ELMAN network and the SVM; the ELMAN network performed better than the SVM.
Liu (2007)[79]	Ta-Liao at Kaohsiung in Taiwan	Daily maximum hourly O_3.	Maximum OT, dew OT, PM_{10}, WS, WD, sunshine, O_3, and NO_X	1997 – 2001.	Box-Jenkins univariate autoregressive integrated moving average (ARIMA), regression with time-series error (RTSE) models; PCA was used in the development of RTSE model.	RTSE model with PCA is superior to ARIMA, RTSE models.
Dutot et al. (2007)[8]	Three monitoring stations namely, Prefecture, La Source and Saint Jean de Braye in the city of Orleans,	Daily maximum hourly O_3.	Cloudiness, rainfall, WS, WD, OT gradient, and O_3.	April – Sept. 1999 – 2003.	Linear model, deterministic model, persistence model, MLP-LM.	Real time neural network model, NEUROZONE, performed better than linear model, deterministic model, and persistence model.

	France					
Sousa et al. (2007)[16]	A monitoring site in Oporto, Northern Portugal.	Hourly O_3.	NO, NO_2, O_3, OT, RH, wind velocity.	July 2003.	MLR, MLP-BP model based on original data, principal component regression and MLP-BP based on principal components.	MLP-BP based on principal components, MLP-BP on original data showed better accuracy prediction compared to the two linear regression models.
Lu et al. (2006)[108]	Four air quality stations (Cutin, Chungming, Chiayi, and Chianjin) and meteorological stations (Taipei, Taichung, Chiayi, and Kaohsuing) in Taiwan.	Hourly O_3.	Hourly average O_3, CO, NO_X, SO_2, PM_{10}, WS, WD, OT, average P, RH, CC, precipitation, global radiation	1998 – 2002.	MLP based on two stage clustering (unsupervised self-organizing map neural network (SOM) followed by K- means clustering), multilinear regression (MLR). PCA used to obtain component scores of eigen values used as inputs in SOM.	MLP based on two stage clustering performed better than MLP, MLR and two level clustering followed by MLR.
Wang et al. (2006)[21]	Two stations: Tseun Wan, and Tung Chung in Hong Kong	Daily maximum hourly O_3.	NO_2, NO_X, NO, CO, OT, SR, WS and temporal variable (day of the year).	2000 – 2003.	MLP based on synergistically coupled particle swarm optimization (PSO) and Levenberg-Marquardt (LM) algorithm, (MLP-PSO-LM).	MLP-PSO-LM performed better than MLP-LM, and MLP-PSO.
Paster-Barcenas	A rural monitoring	Hourly O_3.	NO, NO_2, O_3, WS, WD, OT, solar irradiance, P, RH	April 2002.	MLP-BP. Sensitivity analysis was used to	

et al. (2005)[10]	station located in “Centre de Capacitacio Agraria de Carcaixent” in Valencia, Spain.				find relatively important inputs.	
Abdul-Wahab et al. (2005)[96]	Kuwait University mobile laboratory at Khaldiya, Kuwait.	Hourly O_3.	O_3, NO_X, NO, CO, SO_2, non CH_4 hydrocarbons OT, SR, WS and WD	June 1997	MLR using PCA.	
Wirtz et al. (2005)[81]	Edmonton East monitoring station and Stony Plain station in Edmonton, Alberta, Canada.	Hourly O_3 (up to 2 hours ahead)	CO, NO, NO_2, SO_2, O_3 total hydrocarbons, mixing height, opacity, RH, WS, WD, temporal variables (hour of the day, month of the year, day of the week).	May to Sept. 1999 – 2003.	MLP-BP	
Heo et al. (2004)[97]	Ssangmun, Bangi, Guro, and Gwangghwamun stations in Seoul in Korea.	Daily maximum hourly O_3.	CO, NO_2, SO_2, O_3, surface WS, surface WD, upper WS, upper WD, surface OT, upper OT, RH, surface SR.	1989 – 1999.	Two forecast models MLP-BP, and Fuzzy expert systems.	MLP-BP model forecasts daily maximum hourly ozone model one day ahead. Fuzzy expert system forecasts the high ozone levels.
Kumar et al. (2004)[89]	Brunei Darussalam airport, Brunei	Daily maximum hourly O_3	Hourly O_3	July 1998 – March 1999.	ARIMA	
Zolgadri et	Bordeaux	Daily	Hourly OT, radiation, solar	1998	Non-linear adaptive	An integrated

al. (2004)[95]	Grand Parc station, Bordeaux, France	maximum O_3.	intensity, barometric pressure, WS, RH, WD, trend of seasonal variation of ozone, and $[NO_2]/[NO]$.	– 2001	state space estimator (NASSE), gain scheduling defined for modeling threshold exceedance for extreme O_3 concentration and an (MLP-LM) was used in an integrated monitoring system	operational ozone warning system was developed.
Chaloulakou et al. (2003)[109]	N. Smirni, Liossia, Maroussi, and Likovrissi stations in Athens, Greece.	Daily maximum hourly O_3.	WS, SR, RH, surface OT, OT at 850 hPa (850 millibars), WD index and ozone	April – Oct 1992 – 1999.	MLP-LM, MLR	MLP-LM performed better than MLR.
Rohli et al. (2003)[98]	Eleven sites in Baton Rouge, Louisiana.	Daily maximum hourly O_3. (8 hr average)	NO_X, O_3, surface OT, dew point OT, WS, sea level P, visibility, dew point depression, vertical mixing features, synoptic scale weather features, and transport from upwind regions.	1995 - 2000.	MLR using PCA, and Decision tree.	Each site has a separate model.
Wang et al. (2003)[25]	Three monitoring stations, namely, Tsuen Wan, Kwai Chung, and Kwun Tong in Hong Kong.	Daily maximum O_3	O_3, NO_2, NO, NO_X,CO, SO_2, respirable suspended particles, WS, WD, SR, indoor OT and outdoor OT	1999 - 2000.	RBF, and Adaptive RBF.	Adaptive RBF performed better than RBF.
Vautard et	Paris, France	O_3 (up to	Horizontal wind, surface P,	Sum	A hybrid statistical-	Model developed

al. (2001)[99]	using data collected from European center for Medium range weather forecasts.	three days ahead)	humidity, OT, cloudiness, SO_2, VOC's, NO_X, CO, boundary conditions, anthropic and biogenic emissions.	mer 1999.	deterministic chemistry transport model.	applicable to continental cities like Paris only.
Kaprara et al. (2001)[100]	Nine monitoring stations in Athens, Greece.		O_3, NO_2, SO_2, CO, smoke, OT, RH, WS, WD,	1990 – 1999.	Classification and Regression Trees (CART), MLR	CART model performed better than MLR.
Cobourn et al (2000)[85]	Seven monitoring stations in Louisville.	Daily maximum hourly O_3 (8 hour average)	Daily 8-hour average of O_3, clear-sky atmospheric transmittance daily minimum OT, WS, CC, humidity.	May–Sept. 1993 – 1999.	MLP-BP	
Prybutok et al. (2000)[23]	A monitoring station in Houston	Daily maximum hourly O_3.	NO, NO_2, O_3, OT, WS, WD, carbon dioxide.	June-Oct. 1993.	MLP-BP, ARIMA, Stepwise regression model.	MLP-BP showed superior performance compared to ARIMA, and stepwise regression.
Hadjiiski et al (2000)[6]	Galleria and Clinton stations in Houston.	Hourly O_3 (up to 5 hours).	Fifty three hydrocarbons (C_2-C_{10} compounds), O_3, NO_X, NO, NO_2, OT, ultraviolet radiation.	June – Nov. 1993.	MLP-BP aided with Sensitivity Analysis	
Sohn et al. (2000)[11]	Seoul, South Korea.	Short term (1-6 hour) and long term (16-21 hour) O_3.	O_3, NO_2, CO, SO_2, OT, WS, sunlight, humidity.	(August and September) in 1997.	MLP-CG aided with spatio-analysis that includes the effects of advection and dispersion	
Benvenuto et al.	Ente Zona Industriale di	I hour, 3 hours and	Hourly measurements of OT, WS, WD, global	1995.	MLP-BP	

(2000)[82]	Porto Margera and Venice municipality monitoring network areas, Venice, Italy.	daily maximum concentrations of O_3, CO, NO_2.	radiation, humidity, precipitation, P, vehicle flow rate, SO_2, O_3, NO,NO_2, O_3, non CH_4 hydrocarbons, PM_{10}.			
Gardner and Dorling (2001, 2000, 1999)[83, 14,12]	Bristol, Edinburgh, Eskdalemuir, Leeds, and Southampton in UK.	Hourly O_3, NO_X, NO_2.	O_3, NO_X, NO_2, amount of low cloud, base of lowest cloud, visibility, dry bulb OT, vapor pressure, WS, and WD. $\cos(2\pi h/24)$, $\sin(2\pi h/24)$, $\cos(2\pi d/365)$, $\sin(2\pi d/365)$, where h is the hour of the day, and d is the day of the year.	1993 – 1996 and for – Southampton 1994 – 1997.	Multilayer perceptron-scaled conjugate gradient (MLP-SCG), regression trees, and linear models.	MLP-SCG performed better than regression trees and linear models even though regression trees were readily interpretable.
Spellman (1999)[15]	Five sites with different topographical and demographical features (Bloomsbury, Leeds and Birmingham being urban sites; Harwell (Oxfordshire) being rural and Strath Vaich being a remote site).	Daily maximum hourly O_3.	Hourly O_3, SO_2, PM_{10}, WS, WD, OT.	May – Sept. 1993 – 1996	MLP-BP, MLR.	
Comrie (1997)[7]	Eight monitoring sites from different	Daily maximum hourly O_3.	Daily maximum OT, average daily WS, daily total sunshine, and O_3.	May– Sept. 1991	MLP-BP, MLR	MLP-BP performed slightly better than MLR.

	cities (Atlanta, Boston, Charlotte, Chicago, Phoenix, Pittsburgh, Seattle, and Tucson) in USA			– 1995.		
Yi et al. (1996)[24]	A monitoring site in Dallas-Fort Worth (DFW) region, Texas	Daily maximum hourly O_3.	Hourly values of NO, NO_2, O_3, CO_2, OT, WS, and WD.	June –Oct. 1993 – 1994.	MLP-BP, MLR, and Box-Jenkins model.	MLP-BP showed better performance than MLR, and Box-Jenkins model.
Ryan (1995)[106]	Baltimore metropolitan area (Baltimore-Washington region).	Daily maximum hourly O_3.	Skycover, WS, OT, pressure, O_3, and dew point temperature.	1983 – 1993	CART, Stepwise MLR, Subjective or Expert Analysis.	Subjective or Expert Analysis performs better than stepwise MLR and CART in strong ozone episodes. Stepwise MLR is better than CART.
Clark et al (1982)[90]	27 monitoring stations from Northeastern quadrant of the US	Daily maximum based on one hour average ozone.	35 prognostic variables including hourly OT, absolute humidity, WS, O_3, NO_X, precipitation, sea level pressure, altitude.	June – Sept. 1975 – 1977.	(Stepwise) MLR	Stepwise MLR model was developed separately for each of the 27 sites.
Karl (1979)[107]	25 sites divided into three different groups of sites from Greater St.Louis, Missouri.	Daily maximum based on one hour average ozone (up to 48 hours)	Boundary layer, WS, OT, precipitation, RH, P, O_3, dew point, OT, vertical velocity, and day of the week, sine, and cosine of Julian date.	April –Oct. 1975 – 1976.	Model Output Statistics (MOS) based on derived from National Meteorological Center's Limited area Fine Mesh (LFM) model.	

Wolff et al (1978)[105]	Approximately 75 sites from Northeastern quadrant of the US.	Daily maximum based on one hour average ozone.	Hourly T, absolute humidity, WS, O_3, NO_X, and hydrocarbons.	April – Sept. 1976.	(Stepwise) MLR	A stepwise regression model was calibrated based on New Jersey data and tested on sites at Northeastern Ohio, Marquette, MI, Norfolk, VA, Cook County, IL and Connecticut.

Appendix B

Monitoring station/site details

Dallas Fort Worth region data			
Site No	**Ozone Monitoring site details (Name, AQS_ID/EPA site number, (latitude, longitude))**	**Pollutant data (O_3, NO, NO_2 measured in parts per billion (ppb))details**	**Meteorological data (Resultant wind speed (miles/hour), resultant wind direction(degrees), outdoor temperature (^{o} F), solar radiation (Langleys/minute)) details**
1	Fort Worth Northwest AQS_ID: 484391002 (32.8058183, - 97.3565675)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Fort Worth Northwest	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013, 2014 collected from Fort Worth Northwest.
2	Arlington Municipal Airport AQS_ID: 484393011 (32.6563574, - 97.0885849)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Arlington Municipal Airport	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013, 2014 collected from Arlington Municipal Airport.
3	Italy AQS_ID: 481391044 (32.1754166, - 96.8701892)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Italy	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013, 2014 collected from Italy.
4	Midlothian OFW AQS_ID: 481390016 (32.4820829, - 97.0268987)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013, 2014 collected from Midlothian OFW	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013, 2014 collected from Midlothian OFW.
5	Greenville AQS_ID: 482311006 (33.1530882, - 96.1155717)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013, 2014 collected from Greenville	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013, 2014 collected from Greenville.
6	Kaufman AQS_ID: 482570005 (32.5649684, - 96.3176873)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Kaufman.	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013, 2014 collected from Kaufman.
7	Corsicana Airport AQS_ID: 483491051 (32.0319335,- 96.3991408)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013, 2014 collected from Corsicana Airport	Temperature, Wind for the years 2010, 2011, 2012, 2013, 2014 collected from Corsicana Airport. Solar radiation data for the years 2010, 2011, 2012, 2013, 2014 was made using the mean of the three sites: Italy,

			Midlothian and Kaufman stations.
8	Eagle Mountain Lake AQS_ID: 484390075 (32.9878908, -97.4771754)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013, 2014 collected from Eagle Mountain Lake. For the missing months of 2010 NO, NO_2 data, stations Midlothian, Arlington Municipal Airport, Dallas Executive Airport, Fort Worth North West, Dallas Hilton, Grapevine Fairway, and Denton Airport South	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013, 2014 collected from Eagle Mountain Lake.
9	Keller AQS_ID: 484392003 (32.9225007, -97.2820936)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013, 2014 collected from Keller. For the missing months of 2010 NO, NO_2 data, stations Midlothian, Arlington Municipal Airport, Dallas Executive Airport, Fort Worth North West, Dallas Hilton, Grapevine Fairway, and Denton Airport South	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013, 2014 collected from Keller.
10	Grapewine Fairway AQS_ID: 484393009 (32.9842596,-97.0637211)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Grapewine Fairway	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013,2014 collected from Grapewine Fairway
11	Dallas Executive Airport AQS_ID: 484393011 (32.6563574, -97.0885849)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Dallas Executive Airport.	Temperature, Wind for the years 2010, 2011, 2012, 2013, 2014 collected from Dallas Executive Airport. Solar radiation data for the years 2010, 2011, 2012, 2013, 2014 was made using the mean of the Italy, Midlothian, Keller, Kaufman, Arlington Municipal Airport, Dallas Hilton, Eagle Mountain Lake, Fort

			Worth North West, Grapevine Fairway, Dallas Executive Airport, and Denton Airport South.
12	Dallas Hilton AQS_ID: 481130069 (32.8200608, -96.8601165)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Dallas Hilton	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013,2014 collected from Dallas Hilton.
13	Denton South Airport AQS_ID: 481210034 (33.2190690, -97.1962836)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013, 2014 collected from Denton South Airport	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013, 2014 collected from Denton South Airport.

Houston-Galveston-Brazoria region data			
Site No	**Ozone Monitoring site details (Name, AQS_ID/EPA site number, (latitude, longitude))**	**Pollutant data (O_3, NO, NO_2 measured in parts per billion (ppb))details**	**Meteorological data (Resultant wind speed (miles/hour), resultant wind direction(degrees), outdoor temperature (^oF), solar radiation (Langleys/minutes)) details**
1	Houston Aldine AQS_ID: 482010024 (29.9010364, - 95.3261373)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Houston Aldine.	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013,2014 collected from Houston Aldine.
2	Clinton AQS_ID: 482011035 (29.7337263, - 95.2575931)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013, 2014 collected from Clinton.	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013,2014 collected from Clinton.
3	Conroe (Relocated) AQS_ID: 483390078 (30.3503017, - 95.4251278	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013, 2014 collected from Conroe (Relocated)	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013,2014 collected from Conroe (Relocated).
4	Channel View AQS_ID: 482010026 (29.8027073, - 95.1254948)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013, 2014 collected from Channel View.	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013,2014 collected from Channel View.
5	Galveston 99th Street AQS_ID: 481671034 (29.2544736,-94.8612886)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013, 2014 collected from Galveston 99th Street.	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013,2014 collected from Galveston 99th Street.
6	Houston Bayland	O_3, NO, NO_2 data for	Temperature, Wind, and Solar

	Park AQS_ID: 482010055 (29.6957294, - 95.4992190)	the years 2010, 2011, 2012, 2013, 2014 collected from Houston Bayland Park	Radiation data for the years 2010, 2011, 2012, 2013,2014 collected from Houston Bayland Park.
7	Houston Deer Park2 AQS_ID: 482011039 (29.670025, - 95.1285077)	O_3, NO, NO_2 data for the years 2010, 2011,2012,2013, 2014 collected from Houston Deer Park 2.	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013,2014 collected from Houston Deer Park 2.
8	Lynchbury Ferry AQS_ID: 482011015 (29.7616528, - 95.0813861)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Lynchbury Ferry.	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013,2014 collected from Lynchbury Ferry.
9	Lake Jackson AQS_ID: 480391016 (29.0437592, - 95.4729462)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Lake Jackson.	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013,2014 collected from Lake Jackson.
10	Northwest Harris AQS_ID: 482010029 (30.0395240, - 95.6739508)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Northwest Harris	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013,2014 collected from Northwest Harris.
11	Park Place AQS_ID: 482010416 (29.6863890, - 95.2947220)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Park Place	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013, 2014 collected from Park Place.
12	Seabrook Friendship Park AQS_ID: 482011050 (29.5830473, - 95.0155437)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013, 2014 collected from Seabrook Friendship Park.	Temperature, Wind, and Solar Radiation data for the years 2010, 2011, 2012, 2013,2014 collected from Seabrook Friendship Park.

Los Angeles county data			
Site No	**Ozone Monitoring site details (Name, AQS_ID/EPA site number, (latitude, longitude))**	**Pollutant data (O_3, NO, NO_2 measured in parts per billion (ppb))details**	**Meteorological data (Resultant wind speed (miles/hour), resultant wind direction(degrees), outdoor temperature(^{o} F), solar radiation (Watt/m^2)) details**
1	Azusa AQS_ID: 060370002 (34.1364,- 117.9239)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Azusa.	Temperature data for 2010, 2012, 2013, 2014 was collected from Azusa, 2011 from Santa Fe Dam ((34.12111,-117.94611),1.6504 miles from Azusa)).

			Solar radiation for the year 2010,2012 was collected from Azusa, 2011, 2013,2014 from Santa Fe Dam (34.12111,-117.94611) Wind data for the year 2010, 2011,2012,2013, and 2014 was collected from Santa Fe Dam (34.12111,-117.94611)
2	Compton-700 North Bullis Road AQS_ID:060371002 (33.901389, -118.205)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013, 2014 collected from Compton-700 North Bullis Road.	Temperature data for all the years 2010, 2011, 2012, 2013, 2014 collected from Compton-700 North Bullis Road. Solar Radiation data for all the years 2010,2011,2012,2013, and 2014 collected from (Long Beach # 2 (33.79699,-118.09399) 9.61 miles away from Compton-700 North Bullis Road). Wind data for the years 2010, 2011, 2012, 2013 collected from Long Beach # 2 and 2014 data from Compton-700 North Bullis Road.
3.	Glendora Laurel AQS_ID: 060370016 (34.14437,-117.85038)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Glendora Laurel.	Temperature data for all the years 2010, 2011, 2012, 2013, 2014 collected from Glendora Laurel. Solar Radiation and Wind data for all the years 2010, 2011, 2012, 2013 and 2014 were collected from Santa Fe Dam (34.12111,-117.94611) 5.6879 miles away from Glendora Laurel.
4	Lancaster -43301 Division street AQS_ID: 060379033 (34.669586,-118.13076)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Lancaster -43301 Division street.	Temperature data for all the years 2010, 2011, 2012, 2013, 2014 collected from Lancaster -43301 Division street. Wind data for all the years 2010, 2011, 2013 and 2014 were collected from (Palmdale #4 (34.6150,-118.033) 6.7076 miles away from Lancaster -43301 Division street) and 2012 data from Lancaster -43301 Division street.

			Solar Radiation for all the years 2010, 2011, 2012, 2013 and 2014 were collected from Palmdale #4 (34.6150,-118.033) 6.7076 miles away from Lancaster -43301 Division street.
5	Los Angeles North Main Street AQS_ID: 060370016 (34.06638,-118.22666)	O_3 data for the years 2010, 2011, 2012, 2013,2014 collected from Los Angeles North Main Street. NO, NO_2 data for the years 2010, 2011, 2013,2014 collected from Los Angeles North Main Street and 2012 data from Pasadena-S Wilson Avenue (34.132778,-118.127222) 16.85 miles away from Los Angeles North Main Street.	Temperature data for the years 2010, 2011, 2014 collected from Los Angeles North Main Street and for the years 2012, 2013 from Los Angeles USC_Campus Downtown (34.0167, -118.283), 4.7045 miles away from Los Angeles North Main Street Solar Radiation data for the years 2010, 2011, 2012, 2013 was collected from (Glendale # 2 (34.2, -118.232) 9.2249 miles away from Los Angeles North Main Street) and for the year 2014 data was collected from Los Angeles North Main Street. Wind data for the for the years 2010, 2011, 2012, 2013 was collected from Glendale # 2 and for the year 2014 data was collected from Los Angeles USC Campus Downtown.
6	Pasadena S Wilson Avenue AQS_ID: 060372005 (34.132778,-118.127222)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Pasadena S Wilson Avenue and for the missing months during the years 2010, 2012, 2013 data was used from Los Angeles North Main Street.	Temperature and Solar Radiation data for all the years 2010, 2011, 2012, 2013, 2014 collected from (Glendale # 2 (34.2, -118.232) 21.865 miles away from Pasadena S Wilson Avenue. Wind data for all the years 2010, 2011, 2012, 2013 collected from Glendale # 2, and for the year 2014, data was collected from Pasadena S Wilson Avenue.
7	Pomona AQS_ID: 060371701 (34.066696,-117.751358)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Pomona.	Temperature data, Wind data and Solar Radiation data for the years 2010, 2011, 2012, 2013 and 2014 data was collected from Pomona # 2 (34.058, -117.812)
8	Santa Clarita	O_3, NO, NO_2 data for	Temperature data, Wind data and

	AQS_ID: 060376012 (34.38340,-118.528471)	the years 2010, 2011, 2012, 2013, 2014 collected from Santa Clarita.	Solar Radiation data for the years 2010, 2011, 2012, 2013 and 2014 data was collected from Santa Clarita.
9	West Los Angeles-VA Hospital AQS_ID:060370113 (34.050556, -118.456665)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013, 2014 collected from West Los Angeles-VA Hospital.	Temperature data, Wind data and Solar Radiation data for the years 2010, 2011, 2012, 2013 and 2014 data was collected from Santa Monica (34.04399,-118.476) 1.1945 miles away from West Los Angeles-VA Hospital.
10	Los Angeles Westchester Parkway AQS_ID:060375005 (33.955055,-118.430442)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013, 2014 collected from Los Angeles Westchester Parkway.	Temperature data, Wind data and Solar Radiation data for the years 2010, 2011, 2012, 2013 and 2014 data was collected from Santa Monica (34.04399,-118.476) 6.6681 miles away from Los Angeles Westchester Parkway.

San Joaquin air basin data			
Site No	**Ozone Monitoring site details (Name, AQS_ID/EPA site number, (latitude, longitude))**	**Pollutant data (O_3, NO, NO_2 measured in parts per billion (ppb))details**	**Meteorological data (Resultant wind speed (miles/hour), resultant wind direction(degrees), outdoor temperature(^{o} F), solar radiation (Watt/m^2)) details**
1	Clovis-N Villa Avenue AQS_ID: 060195001 (36.81944,-119.71638)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Clovis-N Villa Avenue.	Temperature and Wind data for all the years 2010, 2011, 2012, 2013, 2014 collected from Clovis-N Villa Avenue. Solar radiation data for all the years 2010, 2012, 2013 collected from Clovis-N Villa Avenue. 2011 and missing monthly data for the year 2014 collected from Fresno State # 2 (36.820999,-119.742), 1.4188 miles away from Clovis-N Villa Avenue.
2	Merced S Coffee Avenue AQS_ID: 060470003 (37.28166,-120.43361)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Merced S Coffee Avenue.	Temperature and Wind data for all the years 2010, 2011, 2012, 2013, 2014 collected from

			Merced S Coffee Avenue. Solar radiation data for all the years 2010, 2011, 2012, 2013, 2014 collected from Merced (37.314,-120.386), 3.4364 miles away from Merced S Coffee Avenue.
3	Shafter-Walker Street AQS_ID: 060296001 (35.503307,-119.272807)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Shafter-Walker Street.	Temperature and Wind data for all the years 2010, 2011, 2012, 2013 collected from Shafter-Walker Street and data for the missing months in 2014 was collected from Shafter-USDA (35.533,-119.2810), 2.1 miles away Shafter-Walker Street. Solar radiation data for the years 2010, 2012, and 2013 collected from Shafter-Walker Street. 2011 data and data for the missing months in 2014 were collected from Shafter-USDA.
4	Fresno-Sierra Skypark #2 AQS_ID: 060190242 (36.84170,-119.8828)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Fresno-Sierra Skypark #2 and ozone data for the missing months in the year 2014 was collected from Fresno Garland (36.78532,-119.774174).	Temperature and Wind data for all the years 2010, 2011, 2012, 2013, 2014 collected from Fresno-Sierra Skypark#2. Solar radiation data for all the years 2010, 2011, 2012, 2013, 2014 collected from Fresno State # 2 (36.820999,-119.742) 7.9068 mile away from Fresno-Sierra Skypark #2.
5	Stockton-Hazelton Street AQS_ID: 060771002 (37.951667, -121.26888)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Stockton-Hazelton Street.	Temperature and Wind data for all the years 2010, 2011, 2012, 2013, 2014 collected from Stockton-Hazelton Street. Solar radiation data for all the years 2010, 2011,

			2012, 2013, 2014 collected from Manteca (37.8350, -121.223), 8.4296 miles away from Stockton-Hazelton Street.
6	Tracy-Airport AQS_ID: 060773005 (37.682499,-121.4406)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Tracy-Airport.	Temperature and Wind data for all the years 2010, 2011, 2012, 2013, 2014 collected from Tracy- Airport. Solar radiation data for all the years 2010, 2011, 2012, 2013, 2014 collected from Tracy (37.72599, -121.474), 3.5122 miles away from Tracy Airport.
7	Turlock-S Minaret Street AQS_ID: 060990006 (37.488236,-120.835886)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Turlock-S Minaret Street.	Temperature data for the years 2010, 2011 was collected from Rose Peak (37.50194,-120.73555). 2012, 2013, and 2014 data was collected from Turlock-S Minaret Street. Wind data for all the years 2010, 2011, 2012, 2013, 2014 collected from Turlock-S Minaret Street. Solar radiation data for all the years 2010, 2011 , 2012, 2013, 2014 collected from Rose Peak (37.50194, -120.735556), 5.5739 miles away from Turlock-S Minaret Street.
8	Visalia-N Church Street AQS_ID: 061072002 (36.3325,-119.290833)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Visalia-N Church Street.	Temperature data for all the years 2010, 2011, 2012, 2013, 2014 collected from Visalia-Airport (36.31388,-119.39222), 5.7814 miles away from Visalia-N Church Street. Wind data for all the years 2010, 2011, 2012,

			2014 collected from Visalia-N Church Street and 2013 data was collected from Visalia-Airport. Solar radiation data for all the years 2010, 2012, 2013, 2014 collected from Visalia-Airport and 2011 data collected from Lindcove (36.35699,-119.059), 12.996 miles away from Visalia-N Church Street.

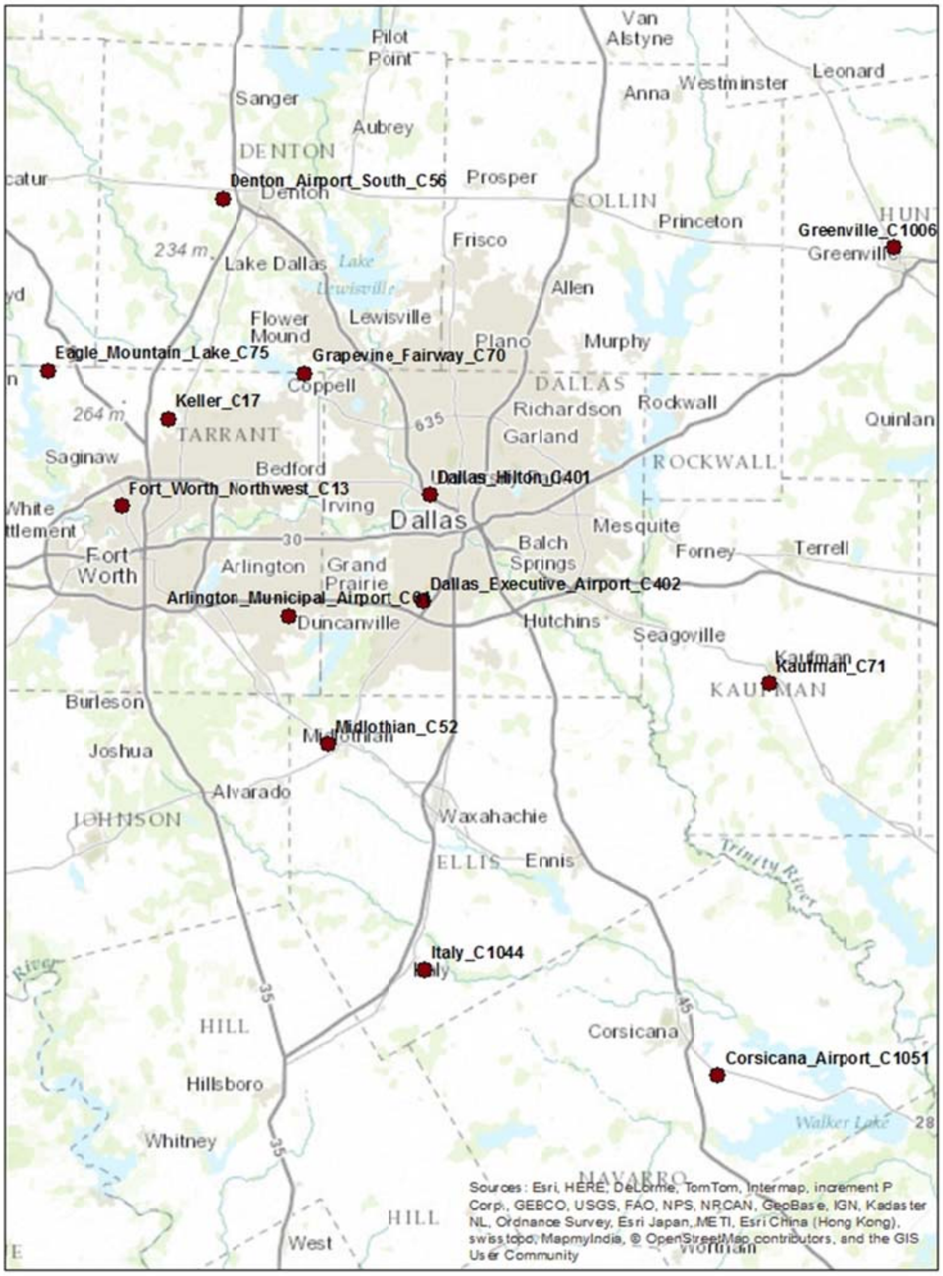
San Diego air basin data			
Site No	**Ozone Monitoring site details (Name, AQS_ID/EPA site number, (latitude, longitude))**	**Pollutant data (O_3, NO, NO_2 measured in parts per billion (ppb))details**	**Meteorological data (Resultant wind speed (miles/hour), resultant wind direction(degrees), outdoor temperature(^{o} F), solar radiation (Watt/m^2)) details**
1	Alpine-Victoria Drive AQS_ID: 060731006 (32.84219, -116.7683)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Alpine-Victoria Drive	Temperature data for all the years 2010, 2011, 2012, 2013, 2014 collected from Alpine-Victoria Drive. Solar radiation data for all the years 2010, 2011, 2012, 2013, 2014 collected from (Alpine (RAWS)-(32.83361, -116.73916)), 1.7434 miles away from Alpine-Victoria Drive. Wind data for the year 2010, 2011 was collected from Alpine (RAWS) for the missing months and for the years 2012, 2013, 2014 data was collected from Alpine-Victoria Drive.
2	Chula Vista AQS_ID: 060730001 (32.631258, -117.05907)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from Chula Vista.	Temperature and Wind data for all the years 2010, 2011, 2012, 2013, 2014 collected from Chula Vista. Solar radiation data for all the years 2010, 2011, 2012, 2013, 2014 collected from San Miguel # 1(32.68,-116.97), 6.1977

			miles away Chula Vista.
3	El Cajon-Redwood Avenue AQS_ID: 060730003 (32.79083, -116.94249)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013 collected from El Cajun-Redwood Avenue. 2014 data for the missing months collected from San Diego Kearny Villa Road (32.84546,-117.12389).	Temperature and Wind data for all the years 2010, 2011, 2012, 2013 collected from El Cajon-Redwood Avenue. 2014 data for the missing months collected from Gillespie Field (32.826099,-116.97199), 2.975 miles away from El Cajon-Redwood Avenue Solar radiation data for all the years 2010, 2011, 2012, 2013, 2014 collected from San Miguel # 1(32.68,-116.97), 7.5229 miles away from El Cajon-Redwood Avenue.
4	Escondido-E Valley Parkway AQS_ID: 060731002 (33.127707,-117.07532)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013, 2014 collected from Escondido-E Valley Parkway.	Temperature and Wind data for all the years 2010, 2011, 2012, 2013, 2014 collected from Escondido-E Valley Parkway. Solar radiation data for all the years 2010, 2011, 2012, 2013, 2014 collected from Escondido SPV (33.081,-116.978), 6.5843 miles away from Escondido-E Valley Parkway.
5	Otay Mesa-Paseo International AQS_ID: 060732007 (32.552216, -116.93793)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013 collected from Otay Mesa-Paseo International. 2014 data collected from Otay Mesa-Donovan (32.57936,116.929486)	Temperature and Wind data for the years 2010, 2011, 2012, 2013 collected from Otay Mesa-Paseo International. 2014 data collected from Otay Mesa-Donovan. Solar radiation data for all the years 2010, 2011, 2012, 2013, 2014 collected from Otay Lake (32.63,-116.938), 5.3675 miles away from Otay Mesa-Paseo International.
6	San Diego-1110 Beardsley Sreet AQS_ID: 060731010 (32.70139,-117.1528)	O_3, NO, NO_2 data for the years 2010, 2011, 2012, 2013,2014 collected from San Diego-1110 Beardsley Street.	Temperature and Wind data for all the years 2010, 2011, 2012, 2013, 2014 collected from San Diego-1110 Beardsley Street. Solar radiation data for all the years 2010, 2011, 2012, 2013, 2014 collected from San Diego # 6 (32.72999,-117.139), 2.1306 miles away San Diego-1110 Beardsley Street.
7	San Diego - Kearny Villa Road AQS_ID: 060731016	O_3, NO, NO_2 data for the missing months for the years 2010, 2011 collected from San Diego Overland	Temperature and Wind data for the missing months for the years 2010, 2011 collected from San Diego Overland Avenue. 2012, 2013, 2014 data collected from San Diego-Kearny Villa Road.

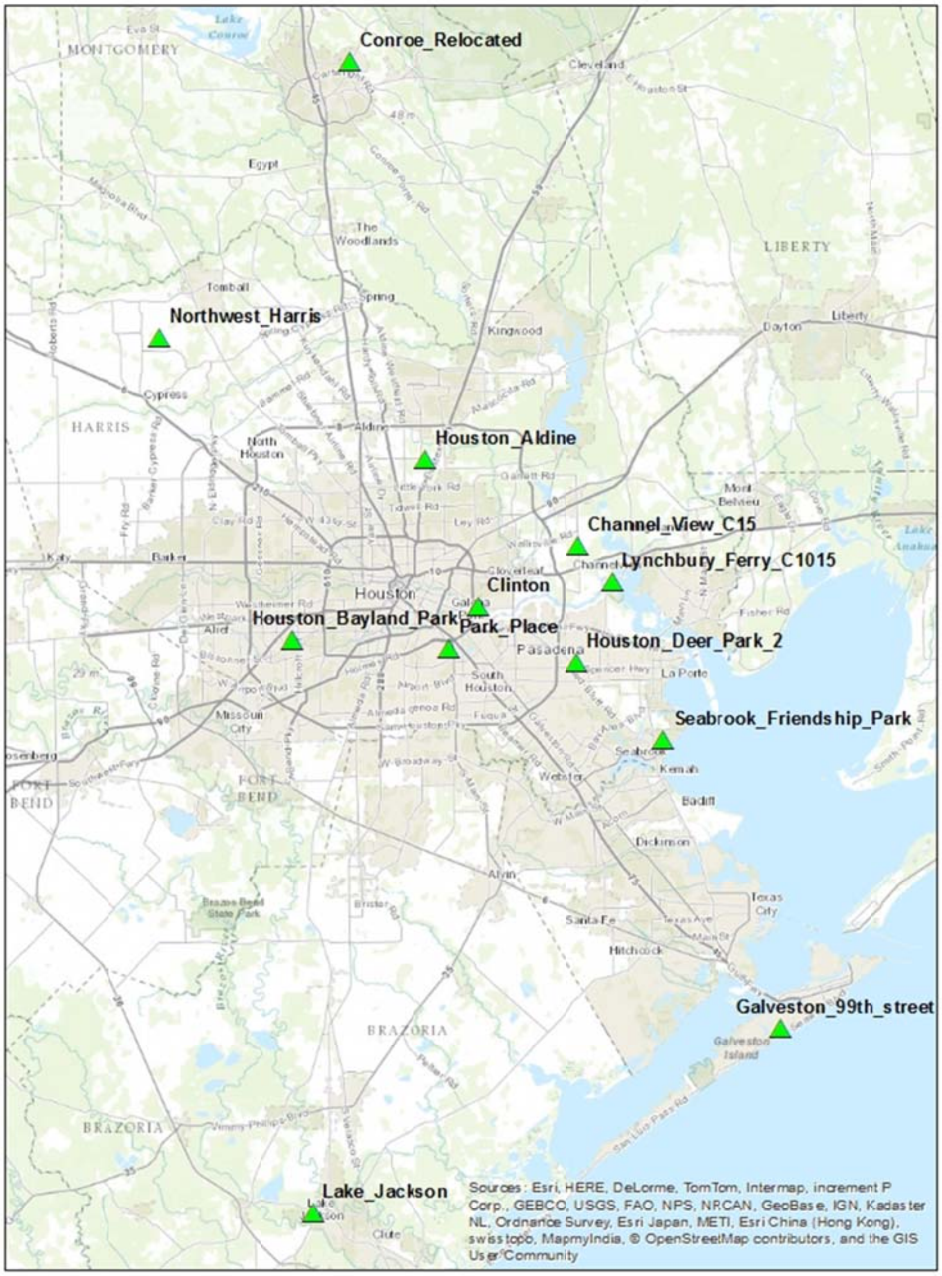
	(32.845467, -117.12389)	Avenue.(32.836461,-117.12869). 2012, 2013, 2014 data collected from San Diego - Kearny Villa Road	Solar radiation data for the missing months for the year 2010 collected from San Diego Overland Avenue, 2011 collected from Miramar (32.886,-117.142), 2.9874 mile away from San Diego - Kearny Villa Road. 2012, 2013 and 2014 data collected from San Diego-Kearny Villa Road.

Appendix C

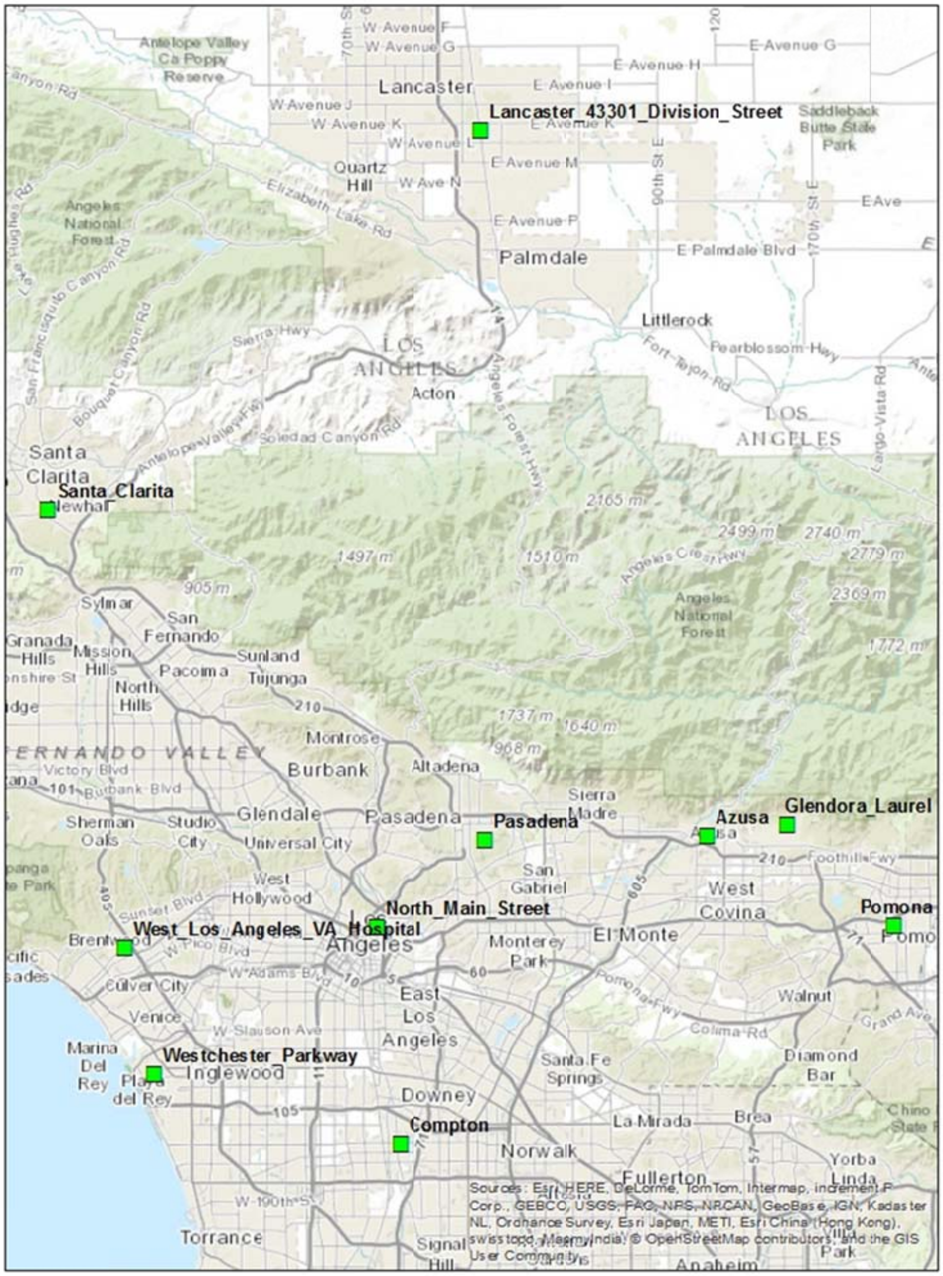
Monitoring station/site maps



Dallas Fort Worth region

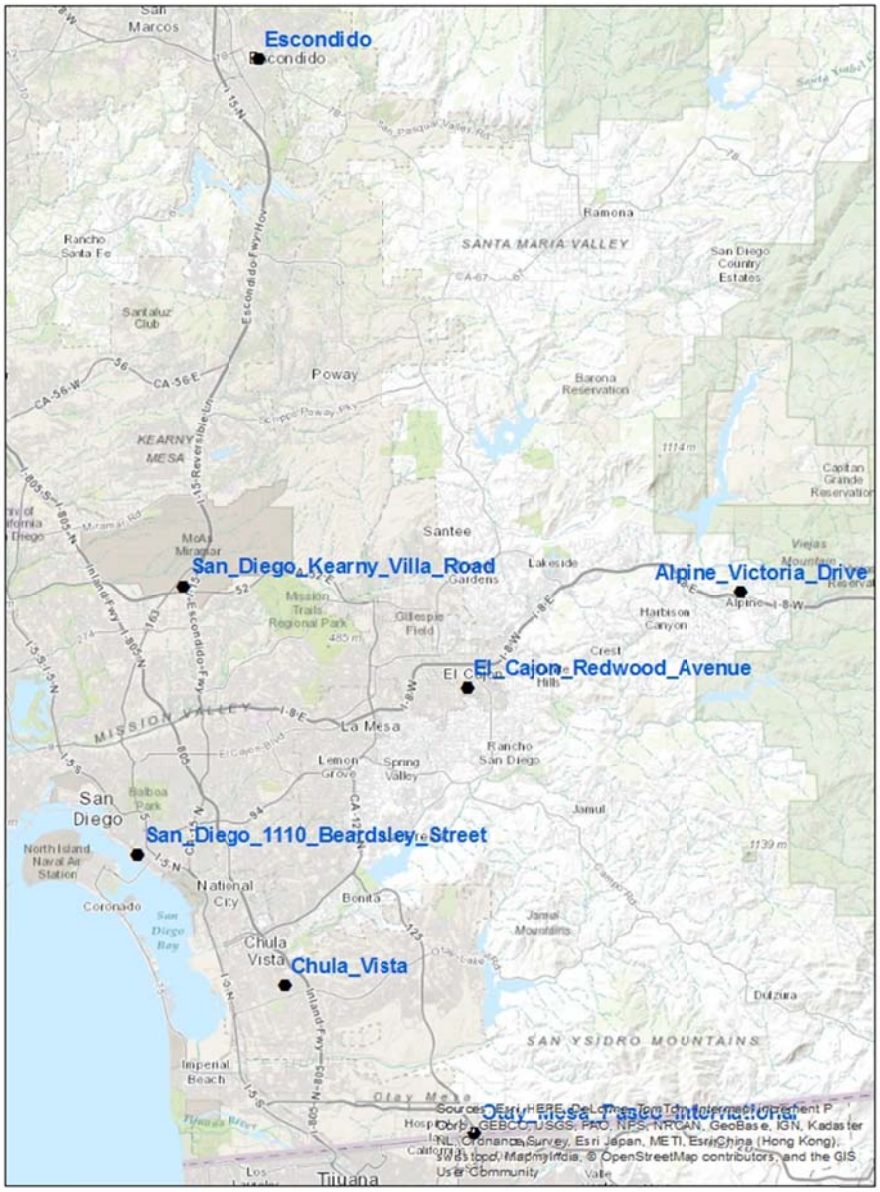


Houston Galveston Brazoria region



Los Angeles air basin

San Joaquin air basin



San Diego air basin


NEVADA
UTAH
CALIFORNIA
ARIZONA
GREAT BASIN
Salt Lake City
Carson City
Sacramento
San Francisco
San Jose
Las Vegas
Tijuana
Mexicali
Phoenix
Medford
Stockton_Hazelton
Tracy_Airport
Turlock
Merced
Skypark
Clovis
Visalia_N_Church
Shafter_Walker_Street
Lancaster_43301_Division_Street
Santa_Clarita
Pomona
Azusa
Pasadena
North_Main_Street
Compton
Escondido
El_Cajon_Redwood_Avenue
Chula_Vista
San_Diego_Kearny_Villa_Road
Otay_Mesa_Paseo_International
Sources: Esri, HERE, DeLorme, TomTom, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

California

References

1. Jacob, D.J. *Introduction to Atmospheric Chemistry*; Princeton: New Jersey, 1999.
2. Turner, D.B.; Schulze, R.H. *Practical Guide to Atmospheric Dispersion Modeling*; Trinity Consultants: Texas, 2007.
3. Haykin, S. *Neural Networks*; Prentice Hall: New Jersey, 1999.
4. Hagan, M.T.; Demuth, H.B.; Beale, M. *Neural Network Design*; PWS Publishing Company: Boston, MA, 1996.
5. Schlink, U.; Helbarth, O.; Richter, M.; Dorling, S.; Nunnari, G.; Cawley, G.; Pelikan, E. Statistical models to assess the health effects and to forecast ground-level ozone; *Environmental Modelling & Software* **2006**, *21*, 547-558.
6. Hadjiiski, L.; Hopke, P. Application of Artificial Neural Networks to Modeling and Prediction of Ambient Ozone Concentrations; *J. Air Waste Manage. Assoc.* **2000**, *50*, 894-901.
7. Comrie, A. C. Comparing Neural Networks and Regression Models for Ozone Forecasting; *J. Air Waste Manage. Assoc.* **1997**, *47*, 653-663.
8. Dutot, A. L.; Rynkiewicz, J.; Steiner, F. E.; Rude, J. A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions; *Environmental Modelling & Software* **2007**, *22*, 1261-1269.
9. Wang, D.; Lu, W. Z. Ground – level ozone prediction using multiple perceptron trained with an innovative hybrid approach; *Ecological Modelling* **2006**, *198*, 332-340.
10. Barcenas, O. P.; Olivas,E. S.; Guerrero, J. D. M.; Valls, G.C.; Rodriguez, J. L. C.; Tascon, S.D.V. Unbiased sensitivity analysis and pruning techniques in neural networks for surface ozone modeling; *Ecological Modelling* **2005**, *182*, 149-158.

11. Sohn, S. H.; Oh, S. C.; Jo, B. W.; Yeo, Y. K. Prediction of Ozone Formation Based on Neural Network; *Journal of Environmental Engineering* **2000**, *126*(8) , 688-696.

12. Gardner, M. W.; Dorling, S. R. Neural network modeling and prediction of hourly NO_X and NO_2 concentrations in urban air in London; *Atmos.Environ.* **1999**, *33*, 709-719.

13. Gardner, M. W.; Dorling, S. R. Meteorologically adjusted trends in UK daily maximum surface ozone concentrations; *Atmos.Environ.* **2000**, *34*, 171-176.

14. Gardner, M. W.; Dorling, S. R. Statistically surface ozone models: an improved methodology to account for non-linear behavior; *Atmos.Environ.* **2000**, *34*, 21-34.

15. Spellman, G. An application of artificial neural networks to the prediction of surface ozone concentrations in the United Kingdom; *Applied Geography* **1999**, *19*, 123-136.

16. Sousa, S. I. V.; Martins, F. G.; Ferraz, M. C. M. A.; Pereira, M.C. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations; *Environmental Modelling & Software* **2006**, *22*, 97-103.

17. Zhang, G.; Pattuwo, B. E.; Hu, M. Y. Forecasting with artificial neural networks: The state of the art; *International Journal of Forecasting* **1998**, *14*, 35-62.

18. Basheer, I.A.; Hajmeer, M. Artificial neural networks: fundamentals, computing, design, and application; *Journal of Microbiological Methods* **2000**, *43*, 3-31.

19. Malalur, S.S.; Manry, M.T. Feed-Forward Network Training Using Optimal Input Gains; *Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia* **2009**, 1953-1960.

20. Rohit Rawat, Kunal Vora, Michael T.Manry, and Gautam R.Eapi. Multivariable neural network forecasting using two stage feature selection, *13th International Conference on Machine Learning and Application (ICMLA), Detroit, MI*, Dec **2014**.

21. Wang, D.; Lu, W.Z. Forecasting of ozone level in time series using MLP model with a novel hybrid training algorithm; *Atmos.Environ.* **2006**, *40*, 913-924.
22. Ruiz, E. S.; Ordieres, J. B.; Vergara, E. P.; Rizo, S. F. C. Development and comparative analysis of tropospheric ozone prediction models using linear and artificial intelligence-based models in Mexicali, Baja California(Mexico) and Calexico, California (US); *Environmental Modelling & Software* **2008**, *23*, 1056-1069.
23. Prybutok, V. R.; Yi, J.; Mitchell, D. Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations; *European Journal of Operational Research* **2000**, *122*, 31-40.
24. Yi, J.; Prybutok, V. R. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area; *Environmental Pollution* **1996**, *92*(3), 349-357.
25. Wang, W.; Lu, W.; Wang, X.; Leung, A. Y. T. Prediction of maximum daily ozone level using combined neural network and statistical characteristics; *Environment International* **2003**, *29*, 555-562.
26. U.S. Environmental Protection Agency. *8 – Hour Ozone Nonattainment Areas*; accessed at www.epa.gov/airquality/greenbook/gntc.html, February **2012**.
27. Texas Commission on Environmental Quality. *Introduction to Air Quality Modeling: Photochemical Modeling*; accessed at www.tceq.texas.gov/airquality/airmod/overview, February **2012**.
28. Rawat, R. An efficient piecewise linear network; Masters Thesis, University of Texas at Arlington December **2009**.
29. Seinfeld, J.H.; Pandis, S.N. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*; John Wiley & Sons: New York, **1998**.

30. Salcedo-Sanz, S.; Portilla – Figueras, J. A.; Ortiz – Garcia, E. G.; Perez – Bellido, A. M.; Garcia - Herrera, R.; Elorrieta, J.I. Spatial Regression analysis of NO_x and O_3 concentrations in Madrid urban area using Radial Basis Function networks; *Chemometrics and Intelligent Laboratory Systems* **2009**, *99*, 79-90.
31. Sarle, W.S. Stopped training and other remedies for overfitting. *Proceedings of the 27th Symposium on the Interface. Available via ftp://ftp.sas.com/pub/neural/inter95.ps.Z* **1995**.
32. Gopalakrishnan, A.; Jiang, X.; Chen, M.S.; Manry, M.T. Constructive proof of Efficient Pattern Storage in the Multilayer Perceptron; *Twentyseventh Asilomar Conference on Signals, Systems & Computers* **1993**, *1*, 386-390.
33. Chen, M. S.; Manry, M.T. Conventional modeling of the multilayer perceptron using polynomial basis functions; *Neural Networks, IEEE Transactions* **1993**, *4*(1) 164-166.
34. Mintz, R.; Young, B.R.; Svrcek W.Y. Fuzzy logic modeling of surface ozone concentrations; *Computers & Chemical Engineering* **2005,** 29, 2049-2059.
35. Peton, N.; Dray, G.; Pearson, D.; Mesbah, M.; Vuillot, B. Modelling and analysis of ozone episodes; *Environmental Modelling and Software* **2000**, 15, 647-652.
36. Celikoglu, H.B.; Cigizoglu H.K. Modelling public transport trips by radial basis function neural networks; *Mathematical and Computer Modeling* **2006**, 45, 480 - 489.
37. Costa,A.; Markellos, R.N. Evaluating public transport efficiency with neural network models; *Transpn Res.-C* **1997,** 5, 301-312.
38. Yilmaz, I.; Kaynar, O. Multiple regression, ANN (RBF,MLP) and ANFIS models for prediction of swell potential of clayey soils; *Expert Systems with Applications* **2011,** 38, 5958-5966.

39. Hurtado, J.E.; Londono,J.M.; Meza, M.A. On the applicability of neural networks for soil dynamic amplification analysis; *Soil Dynamics and Earthquake Engineering* **2001**, 21, 579-591.

40. Chen,J.; Adams, B.J. Integration of artificial neural networks with conceptual models in rainfall-runoff modeling; *Journal of Hydrology* **2005**, 318, 232-249.

41. Mjalli, F. S.; Al-Asheh, S.; Alfadala, H.E. Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance; *Journal of Environmental Management* **2007**,83, 329-338.

42. Narasimha, P.L.; Manry, M.T.; Maldonado, F. Upper bound on pattern storage in feedforward networks; *Neurocomputing* **2008**, 71, 3612-3616.

43. Li, Jiang.; Manry, M.T.; Narasimha, P.L. Feature Selection Using a Piecewise Linear Network; *Neural Networks, IEEE Transactions* **2006,** 17, 1101-1115.

44. Vaidyalingam, J. Cubic model of a multilayer perceptron; Masters Thesis, University of Texas at Arlington December **2003**.

45. Guyon, I.; Elisseeff, A. An introduction to Variable and Feature Selection; *Journal of Machine Learning Research* **2003**, 3, 1157-1182.

46. Crone, S. F.; Kourentzes, N. Feature selection for time series prediction – A combined filter and wrapper approach for neural networks; *Neurocomputing* **2010**, 73, 1923 - 1936.

47. Dash, M.; Liu, H. Feature Selection for Classification; *Intelligent Data Analysis* **1997**, 1, 131-156.

48. Verikas, A.; Bacauskiene, M. Feature selection with neural networks; *Pattern Recognition Letters* **2002**, 23, 1323-1335.

49. Castellano,G.; Fanelli, A.M. Variable selection using neural-network models; *Neurocomputing* **2000**, 31, 1-13.

50. Zheng, H.; Zhang, Y. Feature selection for high-dimensional data in astronomy; *Advances in Space Research* **2008**, 41, 1960-1964.

51. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern Classification*; John Wiley & Sons, Inc: New York, 2001.

52. Vora, K. Neural network based forecaster using feature selection; *Masters Thesis*, University of Texas at Arlington December **2012**.

53. Fukunaga,K. *Introduction to Statistical Pattern Recognition*; Academic Press, California, 1990.

54. Pudil,P.; Novovicova, J.; Kittler, J. Floating search methods in feature selection; *Pattern Recognition Letters* **1994**, 15, 1119 -1125.

55. K L Transforms accessed at http://nptel.iitm.ac.in. **2008**.

56. Karhunen –Louve Theorem accessed at http://en.wikipedia.org/wiki/Karhunen%E2%80%93Lo%C3%A8ve_theorem **2012.**

57. Chattopadhyay, S.; Chattopadhyay, G. Modeling and Prediction of Monthly Total Ozone Concentrations by Use of an Artificial Neural Network Based on Principal Component Analysis; Pure and Applied Geophysics **2012**, 169, 1891-1908.

58. Coman,A.; Ionescu, A.; Candau, Y. Hourly ozone prediction for a 24-h horizon using neural networks. *Environmental Modeling & Software* **2008**, 23, 1407-1421.

59. Chattopadhyay, S.; Bandyopadhyay, G. Artificial neural network with backpropagation learning to predict mean monthly total ozone in Arosa, Switzerland; *International Journal of Remote Sensing* **2007**, 4471-4482.

60. Pires, J.C.M.; Sousa, S.I.V.; Pereira, M.C.; Alvim-Ferraz, M.C.M.; Martins, F.G. Management of air quality monitoring using principal component and cluster analysis– Part II: CO, NO_2 and O_3; *Atmospheric Environment* **2008**, 42, 1261-1274.

61. Malalur, S.S.; Manry, M.T. Multiple optimal learning factors for feed-forward networks; Proceedings of SPIE: Independent Component Analysis, Wavelets, Neural Networks, Biosystems, and Nano Engineering VIII, Orlando, Florida, Vol.7703, pp.77030F-1–77030F-12, April 7-9, 2010

62. Manry, M.T. Neural Networks, Course material, Fall 2010, University of Texas at Arlington.

63. Manry, M.T. Statistical Pattern Recognition, Course material, Summer 2011, University of Texas at Arlington.

64. Manry, M.T. Statistical Signal Processing, Course material, Spring 2011, University of Texas at Arlington.

65. Piecewise linear orthonormal floating search method accessed at https://www.youtube.com/watch?v=z87BMktA3vU 2012.

66. Theodoridis, S.; Koutroumbas, K. *Pattern Recognition*; Academic Press, Elsevier (USA), 3rd edition, 2006.

67. Luna, A.S.; Paredes, M.L.L.; de Oliveira, G.C.G.; Correa, S.M. Prediction of ozone concentration in tropospheric levels using artificial networks and support vector machine at Rio de Janeiro, Brazil; *Atmospheric Environment* **2014**, 98, 98-104.

68. Alkasassbeh, M. Prediction of surface ozone using artificial neural networks and support vector machines; International Journal of Advanced Science and Technology 2013, 55, 1-12.

69. Zahedi, G.; Saba, S.; Elkamel, A.; Bahadori, A. Ozone pollution prediction around industrial areas using fuzzy neural network approach; *Clean – Soil, Air , Water* **2014**, 42(7), 871-879.

70. Russo, A.; Raischel, F.; Lind, P, G. Air quality prediction using optimal neural networks; 2013.

71. Arhami, M.; Kamali, M.; Rajabi,M.M. Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations; *Environmental Science Pollution Research* , **2013**, *20*, 4777-4789.

72. Paoli, C.; Notton, G.; Nivet, M.; Padovani, M.; Savelli, J. A neural network model forecasting for prediction of hourly ozone concentration in Corsica; IEEE 2011.

73. Taormina, R.; Mesin. L.; Orione, F.; Pasero. E. Forecasting tropospheric ozone concentrations with adaptive neural networks; *Proceedings of International Joint Conference on Neural Networks, IEEE*, **2011**, 1857-1863.

74. Pasero, E.; Mesin, L. Artificial neural networks for pollution forecast, *Air pollution, Villanyi (Ed.), ISBN: 978-953-307-143-5, InTech,* 2010.

75. Ibarra-Berastegi, G.; Saenz, J.; Ezcurra, A.; Elias, A.; Barona, A. Using neural networks for short-term prediction of air pollution levels; *ACTEA*, *IEEE*, **2009**, 498-502.

76. Kaminski, W.; Skrzypski, J.; Jach-Szakiel, E. Application of artificial neural networks (ANNs) to predict air quality classes in big cities; *19th International Conference on System Engineering*, *IEEE*, **2008**, 135-140.

77. Sun, G.; Hoff, S. J.; Zelle, B.C., Nelson, M.A. Forecasting daily source air using multivariate statistical analysis and radial basis function networks; *J. Air & Waste Manage. Assoc.* **2008**, *58*, 1571-1578.

78. Thomas, S.; Jacko, R.B. Model for forecasting expressway fine particulate matter and carbon monoxide concentration: application of regression and neural network models; *J. Air & Waste Manage. Assoc.***2007**, *57*, 480-488.

79. Liu, P.G. Establishment of a Box-Jenkins Multivariate time-series model to simulate ground-level peak daily one-hour ozone concentrations at Ta-Liao in Taiwan; *J. Air & Waste Manage. Assoc.***2007**, *57*, 1078-1090.

80. Niska, H.; Hiltunen, T.; Karppinen, A.; Russkanen, J.; Kolehmainen, M. Evolving the neural network model for forecasting air pollution time series; J*. Engineering Applications of Artificial Intelligence*, **2004**, *17*, 159-167.

81. Wirtz, D. S.; El-din, M.G.; El-Din, A.; Idriss, A. Systematic development of an Artificial neural network model for real time prediction of ground level ozone in Edmonton, Alberta, Canada; *J. Air & Waste Manage. Assoc.***2005**, *55*, 1847-1857.

82. Benvenuto, F.; Marani, A. Neural networks for environmental problems: Data quality control and air pollution nowcasting; *Global Nest*, **2000**, *2(3)*, 281-292

83. Gardner, W.; Dorling, S. Artificial neural network-derived trends in daily maximum surface ozone concentrations; *J. Air & Waste Manage. Assoc.* **2001**, *51*, 1202-1210.

84. Gardner, M. W.; Dorling, S. R. Artificial neural networks (The multilayer perceptron)-A review of applications in the atmospheric sciences; *Atmos.Environ.***1998**, *32*, 2627-2636.

85. Cobourn, G.W.; Dolcine, L.; French, M.; Hubbard, M.C. A comparison of nonlinear regression and neural network models for ground level ozone forecasting; *J. Air & Waste Manage. Assoc.* **2000**, *50*, 1999-2009.

86. Kandya.A.; Nagendra,S.S.M.; Tiwari. V.K. Forecasting the tropospheric ozone using artificial network modeling approach: A case study of megacity Madras, India; *J Civil Environ Engg.* **2012**, *S1*, 1-5

87. Cobourn, G.W.; Lin, Y. Trends in meteorological adjusted ozone concentrations in six Kentucky metro areas, 1998-2002; *J. Air & Waste Manage. Assoc.* **2004**, *54*, 1383-1393.

88. Al-Alawi, S.M.; Abdul-Wahab S.A.; Bakheit, C.S. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone; *Environmental Modeling & Software.***2008**, *23*, 396-403.

89. Kumar,K.; Yadav, A.K.; Singh,M.P.; Hassan,H.; Jain, V.K. Forecasting daily maximum surface ozone concentrations in Brunei Darussalam- An ARIMA modeling approach; *J. Air & Waste Manage. Assoc.* 2004, 54, 809-814.

90. Clark, T. L.; Karl, T. R. Application of prognostic meteorological variables to forecasts of daily maximum one-hour ozone concentrations in the Northeastern United Sates; J. Applied Meteorology. **1982**, *21*, 1662-1671.

91. Pires, J.C.M.; Pereira, M.C.; Alvim-Ferraz, M.C.M.; Martins, F.G. Identification of redundant air quality measurements through the use of principal component analysis. *Atmos.Environ.***2009**, *43*, 3837-3842.

92. Pires, J.C.M.; Martins, F.G.; Sousa, S.I.V.; Alvim-Ferraz, M.C.M.; Pereira, M.C. Selection and validation of parameters in multiple linear and principal component regressions; *Environmental Modeling & Software.***2008**, *23*, 50-55.

93. Piramuthu, S. Evaluating feature selection methods for learning in data mining applications; *European Journal of Operational Research.* **2004**, *156*, 483-494.

94. Gheyas, I.A.; Smith, L.S. Feature subset selection in large dimensionality domains; Pattern Recognition. **2010**, *43*, 5-13.

95. Zolghadri, A.; Monsion, M.; Marchionini, C.; Petrique, O. Development of an operational model-based warning system for tropospheric ozone concentrations in Bordeaux, France; *Environmental Modeling & Software*. **2004**, *19*, 369-382.

96. Abdul-Wahab, S.A., Bakheit, C.S., Al-Alawi, S.M. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software*. **2005**, *20 (10)*, 1263-1271.

97. Heo, J.S.; Kim, D.S. A new method of ozone forecasting using fuzzy expert system and neural network systems. *Science of the Total Environment*, **2004**, *325 (1-3),* 221-237.

98. Rohli, R.V.; Hsu, S.A., Blanchard, B.W., Fontenot, R.L. Short-Range Prediction of Tropospheric Ozone Concentrations and Exceedances for Baton Rouge, Louisiana. *American Meteorological Society*, **2003**, *18*, 371-383.

99. Vautard, R.; Beekmann, M., Roux, J., Gombert, D. Validation of a hybrid forecasting system for the ozone concentrations over the Paris area. Atmospheric Environment, **2001**, *35*, 2449-2461.

100. Kaprara. A.; Karatzas, K., Moussiopoulos,N. Maximum ozone level prediction in Athens with the aid of the CART system. *7th International conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes,* **2001**, Belgirate, Italy.

101. Biancofiore, F.; Verdecchia, M., Carlo, P, D., Tomasetti, B., Aruffo, E., Busilacchio, M., Bianco, S., Tommaso, S.D., Colangeli, C. Analysis of surface ozone using a recurrent neural network. *Science of Total Environment,* **2015,** *514,* 379-387.

102. Pires, J. C. M.; Goncalves. B.; Azevedo, F. G.; Carneiro, A. P.; Rego, N.; Assembleia, A.J.B.; Lima, J.F.B.; Silva, P.A.; Alves. C.; Martins, F. G. Optimization of artificial neural network models through genetic algorithms for surface ozone concentration forecasting. *Environmental Science Pollution Research,* **2012**, *19*, 3228-3234.

103. Sekar, C.; Ojha, C. S. P.; Gurjar, B. R.; Goyal. M. K. Modeling and prediction of hourly ambient ozone and oxides of nitrogen and decision tree algorithms for an urban intersection in India. *Journal of hazardous toxic and radioactive waste, American Society of Civil Engineers*, **2015**.

104. Tamas, W.; Notton, G.; Paoli, C.; Voyant, C.; Nivet, M.; Balu, A. Urban ozone concentration forecasting with artificial neural network in Corsica. *Mathematical modeling in Civil Engineering*, **2014**, 10(1), 1-9.

105. Wolff, G. T.; Lioy, P. J. An empirical model for forecasting maximum daily ozone levels in the Northeastern U.S. *Journal of Air Pollution Control Association*, **1978**, *28(10)*, 1034-1038.

106. Ryan, W.F. Forecasting severe ozone episodes in the Baltimore metropolitan area. *Atmospheric Environment*, **1995**, *29(17)*, 2387-2398.

107. Karl, T.R. Potential application of model output statistics (MOS) to forecasts of surface ozone concentrations. *Journal of Applied Meteorology*, **1979**, *18*, 254-265.

108. Lu, H-C.; Hsieh, J-C.; and Chang, T-S. Prediction of daily maximum ozone concentrations from meteorological conditions using a two-stage neural network. *Atmospheric Research*, **2006**, 81,124-139.

109. Chaloulakou, A.; Saisana, M.; Spyrellis, N. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. *The Science of Total Environment*, **2003**, *313*, 1-13.

110. deSouza, A.; Aristone, F.; Sabbah, I. Modeling the surface ozone concentration in Campo Grande (MS) - Brazil using Neural Networks. *Natural Science*, **2015**, 171-178.

111. Rawat, R.; Patel, J.K.; Manry, M.T. Minimizing validation error with respect to network size and number of training epochs. Neural Networks. *The 2013 International Joint Conference (IJCNN)*, Dallas, Texas, Aug. **2013,** 1-7. doi:10.1109/IJCNN.2013.6706919

112. Donald F. Gatz. *Feasibility of Forecasting Surface Ozone Concentrations in the Chicago Area*. Prepared for the Illinois Department of Natural Resources Department of Commerce and Community Affairs. April 2003.

Biographical Information

Gautam R. Eapi was born in Visakhapatnam, Andhra Pradesh, India, in 1979. He received his B.Tech degree in Civil Engineering from Jawaharlal Nehru Technological University, Kakinada, India in 2003. He received his M.Tech degree in Environmental Engineering from Motilal Nehru National Institute of Technology, Allahabad, India in 2007. He received his Ph.D. degree in Environmental Engineering from the University of Texas at Arlington, Texas, U.S.A in 2015. In the past, he worked in the Environmental Defense Fund and Go Green projects. His current research interest is in the area of application of artificial neural networks and statistical pattern recognition in the field of Environmental Engineering.