FUNCTIONAL DATA ANALYSIS FOR ENVIRONMENTAL AND BIOMEDICAL PROBLEMS

by

CHIVALAI TEMIYASATHIT

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 2008

To my Dad, Mom, Sister, Grandma, Uncles, and Aunties for never lost any faith in me.

ACKNOWLEDGEMENTS

This dissertation would not be possible without many great people that have provided me with numerous supports and generosity throughout the years of my graduate work, a million thanks to them. Regardless of the extent and thoroughness of this acknowledgement, it is incomparable to how grateful I am to all of my supporters for every memory we shared and every moment we experienced. I will cherish this feeling forever.

My first gratitude goes to my supervising professor, Dr. Seoung Bum Kim, for his invaluable advices during the course of my doctoral study. I have been extremely fortunate to have an advisor who gave me the freedom to explore the researches independently, while patiently and consistently guided me towards my goal. Seoung had not only provided academic opportunities and economic supports, he also instilled me with the qualities to become a good researcher. Without his continual supports and encouragements, this dissertation would not have been completed.

I am highly grateful to my supervising committee, Dr. H. W. Corley, Dr. Victoria C. P. Chen, and Dr. Melanie Sattler, for their guidance over the years and generously given their time and expertise to better my dissertation. I am especially grateful to Dr. Victoria C. P. Chen and Dr. Melanie Sattler for their thoughtful criticisms and feedbacks on my manuscripts during busy semesters.

To Dr. Kevin Schug who allowed me to collaboratively work on the mass spectrometry research with him and his students, I am thankful to him for these wonderful opportunities and for his insightful advices on many chemistry aspects. I would also like to extend my gratitude to Dr. Soontorn Oraintara for the guidance and valuable suggestions on the NMR research at many different stages. His constructive criticisms

have enlightened me with various aspects on the signal processing. Furthermore, I would like to take this opportunity to thank Dr. Hanli Liu for kindness discussions and also provided the dataset for the prostate cancer research. I am indebted to Dr. Kwok-Leung Tsui for his invaluable advices during our meetings at INFORMS and Arlington. His comments really helped me to understand and enhance my research ideas. I appreciate the hospitality and generosity from Dr. K. R. Rao, thank you very much for your kindness. After all, I am highly appreciated the hardness and efforts, from all my teachers in Chulalongkorn University, Thailand and The University of Texas at Arlington (UTA), which has equipped me with knowledge and wisdom to be able to pursue my doctorate degree at UTA. My thanks also go to the Industrial and Manufacturing Systems Engineering staffs: Christie Murphy, Kimetha Williams, Julie Estill, Rose, and Joyce for their assistants with the necessary administrative tasks during my graduate years.

Many friends have helped me stay sane through these difficult years. Their support and care helped me overcome obstacles and encourage me through my graduate study. I am greatly value our friendships. My special thanks go to all the student collaborators: Yodchanan Wongsawat, Neelesh V. Sule, Hien P. Nguyen, and Raji Misjudeen. Workings closely with all of you were my marvelous experience. Many thanks to my COSMOS colleagues: Dr. Prattana Punnakitikashem, Siriwat Visoldilokpun, Chatabush Roongrat, Thuntee Sukchotrat, Panaya Rattakorn, Weerawat Jitpitaklert, Poovich Phaladiganon, Passakorn Phananiramai, Surachai Charoensri, Bancha Ariyajunya, Dr. Dachuan (Thomas) Shih, Dr. Huiyuan Fan, Ching-feng Lin, Dr. Hee-su Wang, Dr. Prashant Tarun, and Dr. Duraikannan Sundaramoorthi for their constructive comments and discussions on my research. Very special thanks go to my close friends: Natthaphat Leenutapong, Marina (Horikawa) Kobayashi, and Chutima Thumrattranaprut, who always listen, comfort, and support me during my hardships. My

thanks must also go to Dr. Chai Chompooinwai, Att Kruafak, Yothin Rakvongthai, and Panita Suebvisai for making my life here pleasurable.

Most importantly, all of these would not have happened without love and supports from my beloved family. I would like to express my deepest gratitude to my parents who have been a constant and endless source of love. I am thankful for their sacrifices, concerns, patience, and supports, thank you very much Mom and Dad for never ever lost any faith in me. My special heartfelt gratitude goes to my sister who is always the one and only best friend to me for all my life. Thank you for your continually aids and supports, for your forgiveness, and for your encouragements; without you, I would not have survived this far. I am especially grateful to my aunt, Wanna Chen, and my cousin, Paradee Wattanasin, for taken a very good care of me during my stay in TX. Lastly, I am extremely indebted to my Grandmom, Uncles, Aunts, Cousins, and relatives for their love, blessings, and supports they have given to me.

November 11, 2008

ABSTRACT

FUNCTIONAL DATA ANALYSIS FOR ENVIRONMENTAL AND BIOMEDICAL PROBLEMS

CHIVALAI TEMIYASATHIT, Ph.D.

The University of Texas at Arlington, 2008

Supervising Professor: Dr. Seoung Bum Kim

Vast amounts of data are being generated due to the development of sensing technology. Among those, one of the common types of data usually found in various discipline is the functional data. Because the functional data are generally collected in a wide area of interest over a relatively long period, such analyses should take into account both temporal and spatial characteristics. Furthermore, combinations of observations from multiple locations, each with a large number of serially correlated values, lead to a situation that poses a great challenge to analytical and computational capabilities.

In contrast, data obtained from medical and biomedical researches usually collected from a very small number of testing subjects. Since all medical data collection procedures require direct interaction with testing subjects, these procedures need to be carried with high attention and caution to ensure that there is no side effects or consequences from the experiments. Generally, these experiments required approvals from the review board in order to proceed. Therefore, over a long period of time, only a very small set of data can be obtained from a medical study.

To efficiently extract implicit patterns from these datasets, data mining methods are beneficial tools for analyzing such large and complicated as well as small and scarce data. Despite the great potential of applying data mining methods to such complicated data, the appropriate methods remain premature and insufficient. The major aim of this dissertation is to present some data mining methods, along with the real data, as a tool for analyzing the complex behavior of functional data.

In the first part, this dissertation presents a data mining application to: (1) Identify an efficient way to characterize the spatial variations of $PM_{2.5}$ concentrations based solely upon their temporal patterns, and (2) Analyze the temporal and seasonal patterns of PM2.5 concentrations in spatially homogenous regions. This study used 24-hour average $PM_{2.5}$ concentrations measured every third day during the period between 2001 and 2005 at 522 monitoring sites in the continental United States. A k-means clustering algorithm using the correlation distance was employed to investigate the similarity in patterns between temporal profiles observed at the monitoring sites. A k-means clustering analysis produced six clusters of sites with distinct temporal patterns which were able to identify and characterize spatially homogeneous regions of the United States. The study also presents a rotated principal component analysis (RPCA) that has been used for characterizing spatial patterns of air pollution and discusses the difference between the clustering algorithm and RPCA.

Data mining application for investigating the behavior of ozone concentration will be presented in the followed chapter. Ozone has been known to be associated with human health. Ozone data are generally collected over a long period of time from interested locations. However, constructing ozone monitoring sites may not possible or cost effective due to some limitations such as hazardous environment or inaccessible area. The objective of this present study is: (1) To interpolate ozone concentrations as a functional response at an unsampled location, and (2) To reduce model complexity by constructing a data

compression and reduction model which achieve the highest accuracy as much as possible. This study used daily maximum 8-hour ozone concentrations between 2003 and 2006 at 14 monitoring sites in Dallas-Fort Worth area. Wavelet decomposition broke down the data into multiscale data analysis. Regression Analysis was used as a data compression method. Kriging was applied as a spatial interpolation. In addition, model refining step helped tune the ozone concentration with different variability. This study reveals that our model can achieve up to 6.99 ppb in mean absolute error (MAE) and 9.76 ppb in mean absolute error for high ozone day (MAE_{75}).

Finally, an efficient strategy for classification of prostate cancer in near infrared spectra is illustrated. Prostate cancer is the most common male cancer and the second leading cause of cancer death in the United States. The main purpose of this study is to develop an efficient tool that classifies the near infrared (NIR) spectroscopic data taken from ex vivo human prostate glands as normal or cancer. Our proposed procedure consists of several steps. First, to ensure the comparability between spectra, normalization was done by dividing each spectral point by the area of the total intensity of the spectrum. Second, clustering analysis was performed with these normalized spectra to separate the spectra that represent the normal pattern from a mixed group that contains both normal and tumor spectra. Third, we conducted two-stage classification, the first being an effort to construct a classification model with the labels obtained from the preceding clustering analysis and the second being a classification to focus on the mixed group classified from the first classification model. To increase the accuracy, the second classification model was constructed based on the selected features that capture important characteristics of the spectral data. Our proposed procedure was evaluated by its classification ability in testing samples using a leave-one-out cross validation technique, yielding an accuracy of 90%.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Vast amounts of data are being generated to extract implicit patterns of ambient air pollution. Because air pollution data are generally collected in a wide area of interest over a relatively long period, such analyses should take into account both temporal and spatial characteristics. Furthermore, combinations of observations from multiple monitoring stations, each with a large number of serially correlated values, lead to a situation that poses a great challenge to analytical and computational capabilities. Data mining methods are efficient for analyzing such large and complicated data. Despite the great potential of applying data mining methods to such complicated air pollution data, the appropriate methods remain premature and insufficient. The major aim of this dissertation is to present some data mining methods, along with the real data, as a tool for analyzing the complex behavior of ambient air pollutants.

1.1 Organization of the Dissertation

This dissertation begins with background of data mining for air pollution modeling in Chapter 1. Chapter 2 presents a model for spatial and temporal characterization for Particulate Matter 2.5 in the Continental United States while Chapter 3 presents the interpolation model for ozone concentrations at an unsampled location. Finally, Chapter 4 discusses the future research directions.

1.2 Background

In 1990, under the Clean Air Act., the U.S. Environmental Protection Agency (EPA) set the National Ambient Air Quality Standards (NAAQS) for six pollutants, also known as criteria pollutants, which are particulate matter, ozone, sulfur dioxide, nitrogen dioxide, carbon monoxide, and lead (US EPA, 1990). Any exceedance of the NAAQS results in non-attainment of the region for that particular pollutant.

Well-known consequences of air pollution include the greenhouse effect (global warming), stratospheric ozone depletion, tropospheric (ground-level) ozone, and acid rain [1]. In this dissertation, we present applications concerning tropospheric ozone and the less publicized air pollution problem of particulate matter. High concentrations of tropospheric ozone affect human health by causing acute respiratory problems, chest pain, coughing, throat irritation, or even asthma [2]. Ozone also interferes with the ability of plants to produce and store food, damages the leaves of trees, reduces crop yields, and impacts species diversity in ecosystems [3, 4]. Particulate matter is an air contaminant that results from various particle emissions and gaseous precursor. For example, $PM_{2.5}$ (particulate matter that is 2.5 micrometers or smaller in size) has the potential to cause adverse health effects in humans, including premature mortality, nose and throat irritation, and lung damage [5]. Furthermore, $PM_{2.5}$ has been associated with visibility impairment, acid deposition, and regional climate change. To reduce pollutant concentrations and establish the relevant pollution control program, a clear understanding of the pattern of pollutants in particular regions and time periods is necessary. Data mining techniques can help investigate the behavior of ambient air pollutants and allow us to extract implicit and potentially useful knowledge from complex air quality data. Figure 1.1 illustrates the five primary stages in the data mining process in air pollution problems: data collection, data preprocessing, explanatory analysis and visualization, model construction, and model evaluation.

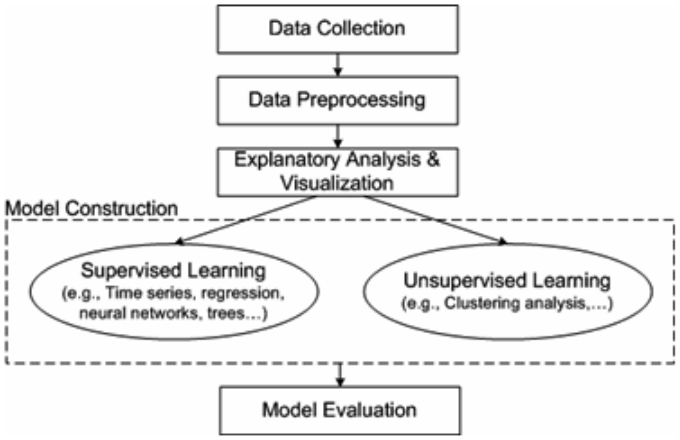


Figure 1.1. Overview of data mining in air pollution problems.

1.3 Data Collection

Because air pollution data are generally collected in a wide region of interest over a relatively long time period, the data are composed of both temporal and spatial information. A typical air pollution database consists of pollutant observations, for monitoring site at time for $i = 1, 2, ..., m$, and $j = 1, 2, ..., n$, where m and n is the number of monitoring sites and time points, respectively. Since most air pollution data hold these two properties, spatial and temporal variability should be incorporated into the analysis in order to accurately analyze the air pollution characteristics. Table 1.1 provides a list of publicly accessible databases and their web addresses that contain a variety of air pollution data.

Table 1.1. Example of publicity accessible database for air pollution data

Network	Locations	Parameters	Time Range	Sources
AIRS-Gaseous	United States (U.S.)	O_3, CO, SO_2, NO_2, PM Mass concentrations	1990-present	AIRS/AQS http://www.epa.gov/ttn/airs/airsaqs/
PAMS, AIRS-Gaseous	U.S. Ozone Nonattainment area	O_3, NO_2, NOX, Nitric Acid	1994-present	Same as AIRS-Gaseous
AIRS-Speciated	U.S.	$PM_{2.5}$ Mass Concentration, Speciated Aerosol	2001-present	Same as AIRS-Gaseous
SEARCH - Continuous	Southeastern U.S.	$PM_{2.5}$ Mass Concentration, Speciated Aerosol, Gaseous, Surface Meteorology	1998-present	http://www.atmospheric-research.com/public/index.html
SEARCH - 24 hour	Southeastern U.S.	$PM_{2.5}$ Mass Concentration, Speciated Aerosol	1998-present	Same as SEARCH-Continuous

1.4 Data Preprocessing

Preprocessing of air pollution data is a crucial task because inadequate preprocessing can result in low-quality data and make it difficult to extract meaningful information from subsequent analyses. The collected air pollution data typically contain a number of potential outliers that are far away from the rest of the observations and missing values possibly due to measurement or instrumental errors. It is necessary to process missing values and outliers in both the time and space domains. Imputing missing values or replacing potential outliers with a sample average is the simplest method because it can be calculated without any pre-specified assumptions or complex mathematical formulas. However, the sample average assumes that each observation is equally important and

does not take into account the fact that the data are collected over time and space. A weighted average can be an efficient method to replace the outliers or impute the missing values. One example of using a weighted average is the inverse-distance-squared weighted method [6]. This method determines weights based on spatial proximity to the query points. Other approaches for processing outliers and missing values include functional or maximum likelihood imputation schemes. Polynomial functions and splines can be used to interpolate regularly-spaced data. Maximum likelihood, which typically requires high computation, uses an iterative approach based on model parameter estimation. Examples of this approach include Expectation-Maximization [7] and kriging [8].

1.5 Exploratory Data Analysis

The main purpose of exploratory analysis and visualization is to provide initial guidelines that enable the subsequent analyses to be more efficient. Principal component analysis (PCA) is a multivariate data analysis technique that helps reduce the dimensions of a data set via an orthogonal linear transformation [9]. The transformed variables, called principal components (PCs), are uncorrelated, and generally, the first few PCs are sufficient to account for most of variability of the entire data. Thus, plotting the observations with these reduced dimensions facilitates the visualization of high-dimensional data. PCA has been used in a variety of air pollution applications [10, 11]. Lengyel et al. [12] observed the diurnal pattern (day and night) of tropospheric ozone concentrations using reduced dimensions in PCA. Lehman et al. [13] applied a rotated PCA approach to study the spatial and temporal variability of tropospheric ozone concentrations in the eastern United States.

Correspondence analysis is another useful explanatory technique that analyzes the relationship between two or more categorical variables. Correspondence analysis examines the contingency table containing the frequency data to investigate how it deviates

from expectation assuming the columns and rows are independent [14]. Similar to PCA, correspondence analysis provides low-dimensional data that facilitate the visualization of the association in multiple levels in contingency tables. Multiple correspondence analysis that extends to the case of more than three categorical variables was used to examine the relationship between nitrogen dioxide exposure levels and related qualitative variables [15].

1.6 Model Construction

Data mining tools for constructing models can be divided into two categories, supervised and unsupervised approaches. Supervised approaches require both the explanatory variable and the response variable, while unsupervised approaches rely solely upon the explanatory variables. Time series analysis is one of the classical supervised approaches for analyzing data collected over time. Numerous studies have used time-series analysis to investigate and predict the behavior of air pollution [13, 16]. Recently, Chelani and Devotta [17] proposed a hybrid autoregressive integrated moving average (ARIMA) model that combined the Box and Jenkins ARIMA model with nonlinear dynamical modeling to forecast nitrogen dioxide concentrations.

Regression analyses aim to build the models based on the relationship between the explanatory and response variables. Regression analysis has been applied to identify the representative monitoring locations [18], and to predict a variety of air pollutant concentrations [19, 12]. Artificial neural networks have also been widely used for predicting ozone concentrations in different locations around the world [11, 20].

Unsupervised approaches aim to extract the information purely from the explanatory variables. Although visualization techniques elicit the natural groupings of the observations, the interpretation of graphical results is not necessarily straightforward. Clustering analysis is an unsupervised approach that systematically partitions the obser-

vations by minimizing within-group variations and maximizing between-group variations, then assigns a cluster label to each observation. Numerous clustering methods have been introduced for grouping air pollution data [21, 22]; however, no consensus exists about the best method to satisfy all conditions. A previous study applied the k-means clustering algorithm with Euclidean distance to sulfur dioxide data from 30 sites in the eastern United States. They obtained six clusters in which the sites within a cluster have a similar pattern of meteorological factors and sulfur dioxide levels [23].

1.7 Model Evaluation

The significance of constructed models should be evaluated for predicting the future behavior of air pollution. The basic approach for model evaluation is to separate data into two data sets, a training set and a testing set. The training set is used to construct the models and these models are then evaluated by their prediction ability on the testing set. Prediction errors typically measure the difference between the actual and fitted values. Table 1.2 lists the frequently used model performance evaluation measures for air pollution modeling.

Table 1.2. Performance measurements for models

Performance Measurement	Equation
Mean Biased (mg/m^3)	$\text{MB} = \frac{1}{m\times n}\sum_{j=1}^{n}\sum_{i=1}^{m}[x_a(s_i,t_j) - x_f(s_i,t_j)]$
Mean Absolute Error (mg/m^3)	$\text{MAE} = \frac{1}{m\times n}\sum_{j=1}^{n}\sum_{i=1}^{m}\lvert x_a(s_i,t_j) - x_f(s_i,t_j)\rvert$
Root Mean Square Error (mg/m^3)	$\text{RMSE} = \sqrt{\frac{1}{m\times n}\sum_{j=1}^{n}\sum_{i=1}^{m}[(x_a(s_i,t_j) - x_f(s_i,t_j))^2]}$
Mean Normalized Bias (%)	$\text{MNB} = \frac{1}{m\times n}\sum_{j=1}^{n}\sum_{i=1}^{m}\left[\frac{x_a(s_i,t_j) - x_f(s_i,t_j)}{x_a(s_i,t_j)}\right]$
Mean Normalized Error (%)	$\text{MNE} = \frac{1}{m\times n}\sum_{j=1}^{n}\sum_{i=1}^{m}\left\lvert\frac{x_a(s_i,t_j) - x_f(s_i,t_j)}{x_a(s_i,t_j)}\right\rvert$
Root Mean Square Normalized Error (%)	$\text{RMSNE} = \sqrt{\frac{1}{m\times n}\sum_{j=1}^{n}\sum_{i=1}^{m}\left[\left(\frac{x_a(s_i,t_j) - x_f(s_i,t_j)}{x_a(s_i,t_j)}\right)^2\right]}$

$x_a(s_i,t_j)$: Actual value,

$x_f(s_i,t_j)$: Fitted value from the model,

m: Total number of monitoring sites,

n: Total number of time points .

CHAPTER 2

CHARACTERIZATION OF SPATIALLY HOMOGENEOUS REGIONS BASED ON TEMPORAL PATTERNS OF PARTICULATE MATTER 2.5 IN THE CONTINENTAL UNITED STATES

Statistical analyses of time-series or spatial data have been widely used to investigate the behavior of ambient air pollutants. Because air pollution data are generally collected in a wide area of interest over a relatively long period, such analyses should take into account both spatial and temporal characteristics. In particular, a number of studies have been devoted to characterization of temporal and (or) spatial correlation(s) in air pollution data collected from a number of monitoring sites in an area of interest. Temporal correlation or spatial correlation can be defined as a correlation between the same variables at different times and locations, respectively, and it measures the strength of the relationship of observations. Sometimes, the term *autocorrelation* is used instead of *correlation* to emphasize its characteristic of self-correlation (i.e., correlation of the variable with itself). Therefore, high temporal or spatial correlation implies a strong relationship of observations (e.g., air pollution concentrations) in time or space.

This chapter focuses on characterizing $PM_{2.5}$, one of the six criteria pollutants identified by the U.S. Environmental Protection Agency under the federal Clean Air Act [24, 25]. The other five criteria pollutants include ozone, sulfur dioxide, nitrogen dioxides, carbon monoxide, and lead [24] $PM_{2.5}$ has the potential to cause adverse health effects in humans, including premature mortality, nose and throat irritation, and lung damage [5, 26].Furthermore, $PM_{2.5}$ has been known to be associated with visibility impairment, acid deposition, and regional climate change [27].

A number of statistical models have been proposed to characterize the spatial correlation of $PM_{2.5}$ concentrations. Descriptive statistical analyses that examined daily, seasonal, and spatial trends in mass, composition, and size distributions of 24-hour average $PM_{2.5}$ concentrations at 16 specific sites in several counties over southeast Texas during the period from 2000 to 2001 showed that mass and composition were generally spatially homogeneous, while particle size distributions were not [28]. A nonnegative factor analytic model was used to analyze the contribution of meteorology (e.g., temperature, humidity, pressure, and wind speed) and other ambient factors (e.g., ozone concentration) to $PM_{2.5}$ concentrations at 300 monitoring sites in the eastern United States during 2000 [29]. Temporal and spatial trends of sulfur dioxide (SO_2), sulfate ($SO_4^=$), nitrogen species, and all major components of $PM_{2.5}$, were investigated from 1989 to 1995 at 34 rural clean air status and trends network (CASTNet) sites in the eastern United States [23]. In their study, a clustering analysis was performed to group 30 sites adjusted for seasonal effects so that the sites within a cluster had a similar pattern of meteorological factors and ozone levels. A more comprehensive study of spatial and temporal trends of $SO_4^=$ was performed over 10 years for 70 monitoring sites in the continental United States [30]. They characterized the spatial trends of $SO_4^=$ concentrations in summer and winter and quantified the temporal change of the $SO_4^=$ level. A number of studies have been conducted to determine the spatial and temporal patterns of aerosol concentrations for impacting haze and visual effect [31, 32, 33].

Analyses of spatial and temporal patterns of pollutants can be used to establish representative monitoring sites. A fixed-effect analysis of variance (ANOVA) model was developed to explore spatial and daily variations of pollutant levels and to identify the representativeness of $PM_{2.5}$ monitoring sites in Seattle, Washington [18]. Furthermore, a statistical model was used to quantify the representativeness of existing monitoring

sites [34]. Principal components analysis was applied to measure the spatial representativeness of ground level ozone concentrations [10].

An understanding of spatial correlations of pollutant concentrations would be useful in improving dynamic air quality models. McNair et al. [6] evaluated the performance of the Carnegie/California Institute of Technology (CIT) model and found that spatial inhomogeneity needed to be taken into account in order to develop model performance guidelines. Jun and Stein [35] compared daily $SO_4^=$ levels between observation data and the Community Multiscale Air Quality (CMAQ) model by space-time correlation. The CMAQ model matches the space-time correlation structure of the observed data; however, CMAQ partially captures time-lagged spatial variation of $SO_4^=$ concentrations. Recently, Park et al. [36] investigated effects of spatial variability on the evaluation of the CMAQ model and observed that slight errors in the model were caused by uncertainties due to the different spatial scales between the point-observations and the volume-averaged simulated concentrations. Their recommendation was to use data at spatially representative monitoring sites in model evaluation.

This chapter seeks to characterize regions of homogenous $PM_{2.5}$ concentrations at 1,402 monitoring sites across the continental United States based solely upon their temporal patterns over multiple years. Each monitoring site provides a profile (or curve) that represents the temporal pattern of $PM_{2.5}$ concentrations. Figure 2.1 shows the 1402 profiles that represent temporal patterns of $PM_{2.5}$ concentrations for 1402 monitoring sites. Combinations of multiple temporal profiles, each with 609 variables (days), lead to a large number of data points and a situation that poses a great challenge to analytical capabilities. Our first objective was to identify an efficient way to characterize $PM_{2.5}$ concentrations based solely upon their temporal patterns. Our approach yielded groupings of the monitoring sites into spatially homogenous regions. Thus, our second objective was to analyze the temporal and seasonal patterns of $PM_{2.5}$ concentrations in these spatially

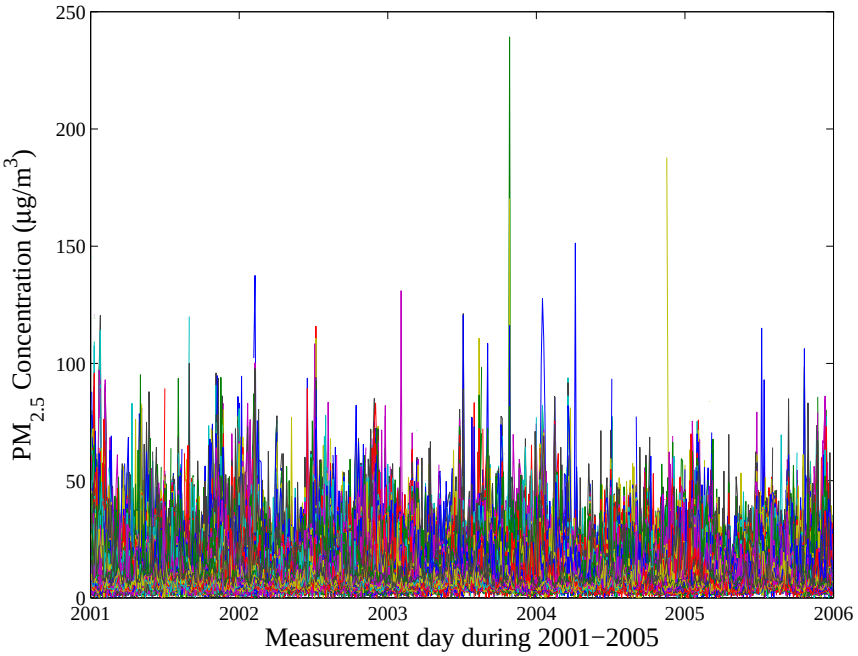


Figure 2.1. 24-hour average $PM_{2.5}$ concentrations measured every third day during five years (January 2001-December 2005) at 1042 monitoring sites in the continental United States. Each curve represents temporal profiles of a monitoring site.

homogenous regions. Finally, our third objective was to examine the feasibility of using spatial and temporal patterns for establishing effective pollutant management programs. The spatial and temporal information play complementary roles in this paper. In other words, temporal patterns at each monitoring site were used to characterize spatial correlations of $PM_{2.5}$ concentrations, and then the identified spatial patterns were used to establish the representative temporal pattern in each spatially homogenous region.

2.1 Data

Monitoring data were obtained from the Aerometric Information System (AIRS) database in the Environmental Protection Agency's Air Quality System (EPA-AQS) (http://www.epa.gov/ttn/airs/airsaqs/), which contains 24-hour average $PM_{2.5}$ mass concentrations measured every third day from 2001 to 2005 at 1402 monitoring sites

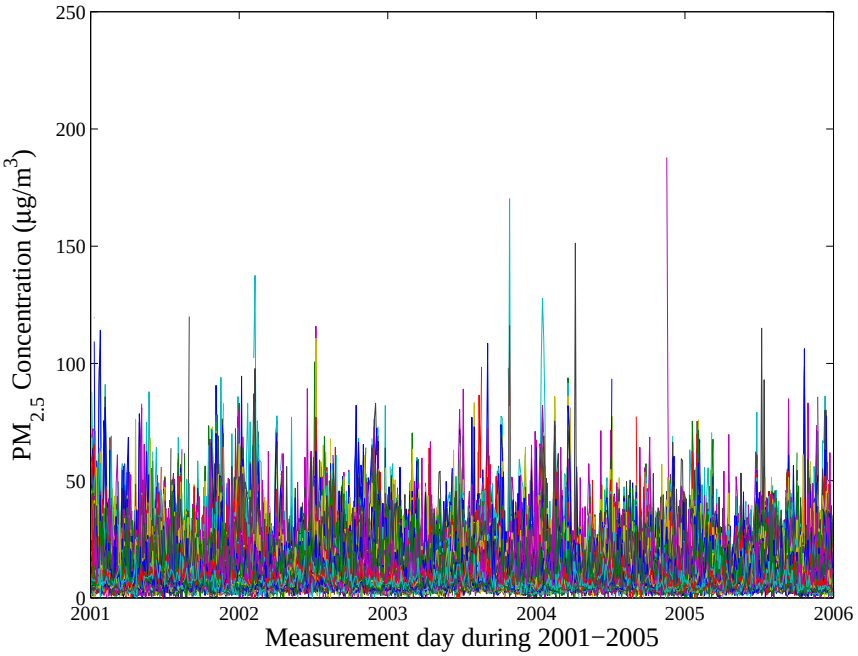


Figure 2.2. 24-hour average $PM_{2.5}$ concentrations measured every third day during five years (January 2001-December 2005) at 522 monitoring sites in the continental United States after removed the missing values and outliers.

in the continental United States. At each 24-hour average $PM_{2.5}$ mass monitoring site, 609 measurements were recorded between 2001 and 2005. Thus, the $PM_{2.5}$ concentration for monitoring site S_i at time T_j can be represented as follows:

$$Z\left(S_i, T_j\right) \quad \text{for} \quad \begin{cases} i &= 1, 2, ..., n, \\ j &= 1, 2, ..., m, \end{cases}$$

where n is the number of monitoring sites ($n = 1402$) and is the number of time points ($m = 609$).The database contains a number of missing values. Monitoring sites that had values missing for more than 50% of the observations or more than 10 consecutive missing values were excluded from the study. The database originally contains 1,402 monitoring sites. After excluding those sites, 522 monitoring sites remained. The remaining missing observations in the dataset were replaced with the interpolation of the nearby values, on the assumption that those were the result of measurement errors or instrument mal-

functions. In addition, we found one observation (October 27, 2003 in California) that had a much higher concentration (239.2 μg/m^3) than the values in its neighborhood. We considered this as an outlier and replaced it with an interpolated value. The remaining 522 sites include both the urban and rural sites. Temporal profile of the preprocessed dataset is shown in figure 2.2. In the present study, we combined the urban and rural sites in the analysis because we are more interested in analyzing an overall spatial and temporal pattern of $PM_{2.5}$ concentration in the continental U.S. rather than addressing questions related to levels of pollutants around specific commercial, industrial, residential, or agricultural sites. Also, we should point out that $PM_{2.5}$ speciation data can be useful for characterizing the patterns of components of total $PM_{2.5}$ mass concentration. However, because the numbers of monitoring sites where the speciation data are available are very limited and the present study seeks to characterize regions of homogenous $PM_{2.5}$ concentrations across the entire continental United States (regional scale), we focused on the analysis of total $PM_{2.5}$ mass concentrations.

2.2 Interpolation Technique to Impute Missing Observations and Outliers

Missing observations and outliers were replaced with interpolated values using an inverse distance squared weighted method [6]. Inverse Distance Weighted (IDW) is one of numerous interpolation techniques to impute missing observations and outliers. By compute a weighted average, missing observations and outliers were replaced with interpolated value. The weights for non-missing observations are the distance function based on Minkowski distance matrix as follows:

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - yk|^r \right)^{1/r}, \tag{2.1}$$

where n is total number of dimensional spaces, and r is the distance parameter. For example, $r = 1$ represents City Block (Manhattan) distance and $r = 2$ represent Euclidean

distance (L_1 norm). The interpolated value for site S_i at time T_j, $I(S_i, T_j)$ is computed as follows:

$$I(S_i, T_j) = \frac{\sum_{k=1, k\neq i}^{m} Z(S_k, T_j) \cdot \omega_k}{\sum_{k=1, k\neq i}^{m} \omega_k} \tag{2.2}$$

where m is the number of monitoring sites and ω_k is calculated as follows:

$$\omega_k(S_i) = \begin{cases} \frac{1}{d(S_i, S_{k,i\neq k})^p} & \text{if} \quad d(S_i, S_{k,i\neq k}) \leq L\, km \\ 0 & \text{if} \quad d(S_i, S_{k,i\neq k}) > L\, km, \end{cases} \tag{2.3}$$

where *Euclidean* distance is selected as distance measure with $r = 2$, inverse distance squared weighted method implies that $p = 2$, and L is a cutoff distance set to 180 km. Thus, $I(S_i, T_j)$ in (2.2) is the weighted average pollutant concentration observed in the surrounding m sites. The weights are determined by the way that observations in close spatial proximity are given more weight than those that are spatially separated. In this paper, L in (2.3) was set to 180 km. Based upon our own analysis, using a different L did not lead to significantly different results for interpolation.

Other approaches for interpolating outliers and missing values include functional, maximum likelihood imputation schemes, and Bayesian modeling. Polynomial functions and splines can be used to interpolate regularly-spaced data. Maximum likelihood or Bayesian modeling, which typically requires high computation, uses an iterative approach based on model parameter estimation. Examples of this approach include Expectation-Maximization [7], radial basis function [37], Bayesian hierarchical model [38, 39], and kriging [8].

2.3 Analytical Approach

2.3.1 k-means Clustering Analysis

k-means clustering analysis is one of the clustering analysis techniques which systematically partitions the dataset by minimizing within-group variation and maximizing between-group variation, and then assigning a cluster label to each observation.24 Clustering analysis has been widely used to facilitate the extraction of implicit patterns and to test the validity of the groupings obtained by visualization methods such as principal components analysis. Variation can be measured based on a variety of distance metrics, e.g. Minkowski distance, Cosine distance , or Correlation distance between observations in a dataset. The brief summary of the k-means clustering algorithm is as follows: Given k seed points, each observation is assigned to one of the k seed points close to the observation, which creates k clusters. Then, seed points are replaced with the mean of the currently assigned clusters. This procedure is repeated with updated seed points until the assignments do not change. The results of the k-means clustering algorithm depend on the distance metrics, the number of clusters (k), and the location of seed points.

For the distance metric, the correlation distance that measures the similarity in patterns between the two temporal profiles from each monitoring site was used. More precisely, for the monitoring sites x and y, the correlation distance between two temporal profiles that consist of a series of m time points can be computed as follows:

$$D_{[Z(S_x),\, Z(S_y)]} = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{Z(S_x,T_j)-\bar{Z}(S_x)}{\sigma_{Z(S_x)}}\right)\left(\frac{Z(S_y,T_j)-\bar{Z}(S_y)}{\sigma_{Z(S_y)}}\right), \tag{2.4}$$

where,

$$\bar{Z}(S_i) = \frac{1}{m}\sum_{j=1}^{m} Z(S_i,T_j), \tag{2.5}$$

and

$$\sigma_{Z(S_i)} = \left[\frac{1}{m}\sum_{j=1}^{m}\left[Z(S_i,T_j)-\bar{Z}(S_i)\right]^2\right]^{\frac{1}{2}}. \tag{2.6}$$

In contrast to Euclidean distance that measures the difference of each time point over the monitoring period, the correlation distance allows us to measure the similarity in shape between the two temporal profiles observed at each monitoring site. In other words, the correlation distance focuses more on an overall pattern rather than scale-difference between the profiles.

To determine the number cluster (k), a heuristic approach was used based on the assumption that we do not have explicit knowledge of expected $PM_{2.5}$ concentration changes in the continental United States. To be specific, we applied the k-means clustering algorithm to our dataset with k values ranging from 5 to 15 for 20 replications. We then selected the final k so that the average value of the standard deviation of k groups (for k = 5, 6,,15) reaches the first minimum. To determine the location of seed points, we used a *sample* method available in MATLAB (MathWorks Inc., Natick, MA).

A previous study applied the k-means clustering algorithm with Euclidean distance to SO_2 data from 30 sites in the eastern United States.8 The study obtained six clusters in which the sites within the cluster had a similar pattern of meteorological factors and ozone levels. The study determined the number k based on geographical and climatological characteristics and estimated the location of seed points using the centroid values of each region. In contrast to Holland et al [23], our study relied solely on statistical methods to determine the number k and the location of seed points. This is a reasonable approach because one of the main purposes of this study is to examine the feasibility of using only temporal patterns of $PM_{2.5}$ concentrations for characterizing spatial correlations. To facilitate the interpretation of temporal patterns, we applied robust locally weighted polynomial regression (LOWESS) [40]. For more mathematical details, see Cleveland [41, 42].

2.3.2 A Rotated Principal Components Analysis Technique

A rotated principal components analysis (RPCA) approach has been used to characterize spatio-temporal patterns of air pollution and meteorological fields‘[43, 13]. We begin with a brief introduction to a traditional PCA approach. Principal component analysis (PCA) is a multivariate data analysis technique primarily for dimensional reduction and visualization a data set via an orthogonal linear transformation [9]. The uncorrelated transform variable represented by the principal component (PC) is a linear combination of all the original variables. For example, the i^{th} PC can be expressed as follows:

$$\mathrm{PC}_i \quad = \mathbf{x}_1 k_{i1} + \mathbf{x}_2 k_{i2} + ... + \mathbf{x}_N k_{ip} \quad = \mathbf{X}\mathbf{k}_i \quad i = 1, 2, ..., p, \tag{2.7}$$

where p is the total number of variables in the original dataset. A set of coefficients is given by the eigenvector with the corresponding i_{th} largest eigenvalue of the covariance matrix of the original dataset where the eigenvalue represent the amount of variability accounted in each PC_i. Because the contribution of each variable to form a PC can be represented by each component of the eigenvector, this vector is often called a *loading vector*. For example, k_{i1} in (2.7) indicates the degree of importance of the first variable in the i_{th} PC domain. In general, the first PC (PC_1) is the most important PC accounted for the maximum variability and the last PC (PC_p) is the least important PC accounted for the minimum variability of the entire dataset. Thus, only first few PCs represent the lower dimensional space can explain most of the variability of the original dataset ($\mathbf{X}$).

Suppose that m PCs can accounted for most of the variable if the original dataset, Rotated Principal Component Analysis (RPCA) attampts to rotate the m loading vectors of the traditional PCA in order to facilitate the spatial interpretation. Typically there are two types of rotation: orthogonal rotation where the new axes are orthogonal to each other and oblique rotation where the new axes are not orthogonal to each other. Among

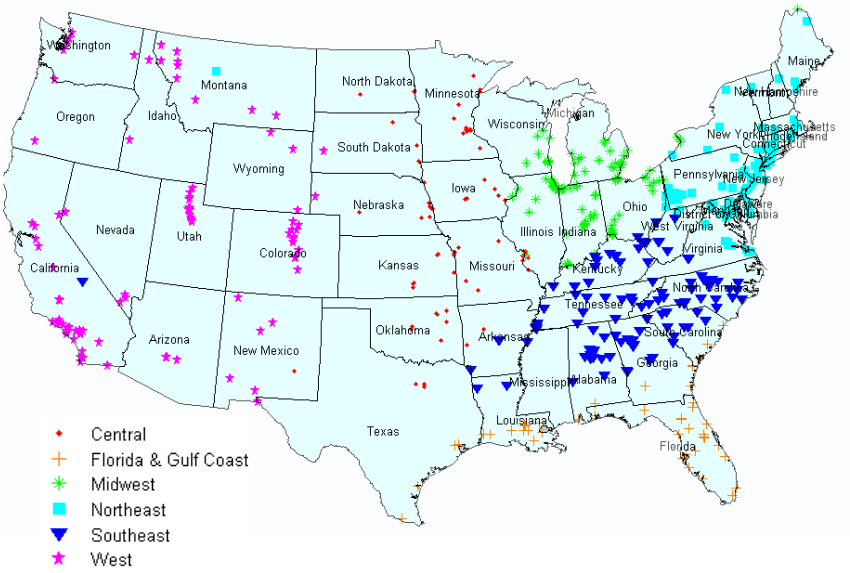


Figure 2.3. k-means clustering results for the continental United States.

the many options for rotation has been widely used, a varimax rotation is selected in this analysis. Varimax rotation is one of the orthogonal rotation that maximizes [9]

$$Q = \sum_{j=1}^{m} \left[\sum_{i=1}^{p} k_{ij}^4 - \frac{1}{p} \left(\sum_{i=1}^{p} k_{ij}^2 \right)^2 \right], \tag{2.8}$$

the sums of the variances of the squared components of loading vector from the traditional PCA where p is total number of variables and m is total number of factors to be rotated.

2.4 Results

2.4.1 Spatial Patterns of $PM_{2.5}$ Concentrations

The k-means clustering algorithm using the correlation distance was performed on the dataset of 522 monitoring sites, each of which had 609 time points. Based on the heuristic method described in previous section, the optimal number for k is six. The results of six-means clustering analysis on temporal profiles are displayed on the U.S. map (Figure 2.3). It is seen that the monitoring sites in close spatial proximity are grouped

together, demonstrating the identification of spatially homogeneous regions solely based on the temporal patterns of PM2.5 concentrations. To further characterize the spatial regions, the clustered sites can be grouped according to the following ad-hoc categories chosen by geographical locations, with the number of monitoring sites in each cluster indicated in parentheses: (i) Central (68); (ii) Florida & Gulf Coast (44); (iii) Midwest (103); (iv) Northeast (104); (v) Southeast (111); and (vi) West (92). Table 2.1 shows a list of states in the United States in each clustered region.

Main factor analysis that compares the mean $PM_{2.5}$ concentrations for each clustered region showed that mean $PM_{2.5}$ concentrations vary regionally from year to year although the degree of difference was not significant (Figure 2.4). In general the highest mean $PM_{2.5}$ concentrations occurred at sites in the Midwest, followed by the Southeast and the Northeast (Figure 2.4). This may be because of the high SO_2 emissions generated within the Ohio River Valley in the Midwest region [30, 44] The mean $PM_{2.5}$ concentration in the Midwest in 2001 (15.02 $\mu g/m^3$) and 2005 (15.56 $\mu g/m^3$), in particular, exceeds the annual federal standard of 15 $\mu g/m^3$ (Figure 2.4). Lower mean concentrations are observed in the West, Florida & Gulf Coast, and Central. It appears from figure 2.4 that the mean $PM_{2.5}$ concentrations have a downward trend from 2001 to 2004 but increase in 2005, except for the West, which exhibits a decreasing trend over the time period from 2001 to 2005.

2.4.2 Comparison with Rotated Principal Components Analysis

A RPCA approach was applied to the same dataset used in k-means clustering analysis. A set of ordered eigenvalue-eigenvector pairs was computed from a 522 by 522 covariance matrix containing the pair-wise covariance of the 522 monitoring sites. Usually, only a small number of PCs is needed to explain the variability in the original dataset. There is no definitive answer to determine an appropriate number of PCs to

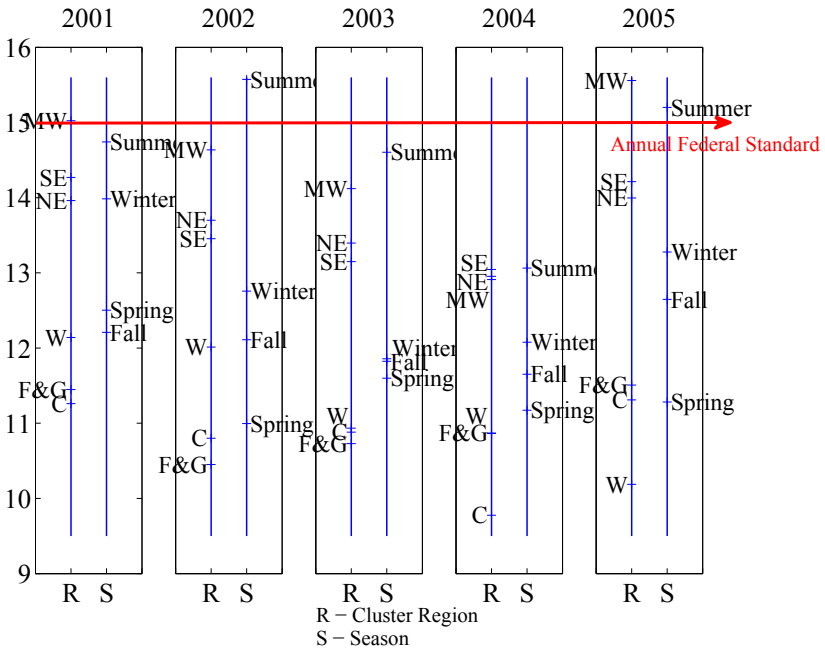


Figure 2.4. A design plot to compare the yearly mean values of $PM_{2.5}$ concentrations by region and season from 2001 to 2005.

retain [14]. One popular method is to use the property that the proportion of variability explained by each PC can be expressed by the eigenvalues. For example, the proportion of variability explained by the i^{th} PC ($V(PC_i)$) can be calculated from the following equation:

$$V(PC_i) = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}, \tag{2.9}$$

where λ_i is the i^{th} eigenvalue, and p is the total number of original variables. The idea of this method is to plot the ordered V(PC) against its rank and determine an appropriate number of PCs. This graphical method is rather subjective since the decision involves a visual inspection. The general recommendation is to find an elbow in the plot. In the present study, we found that the elbow point was observed around five, six, and seven PCs. Of these, we decided to retain the six PCs in order to ensure the comparability to the six clusters obtained from the clustering analysis in previous section.

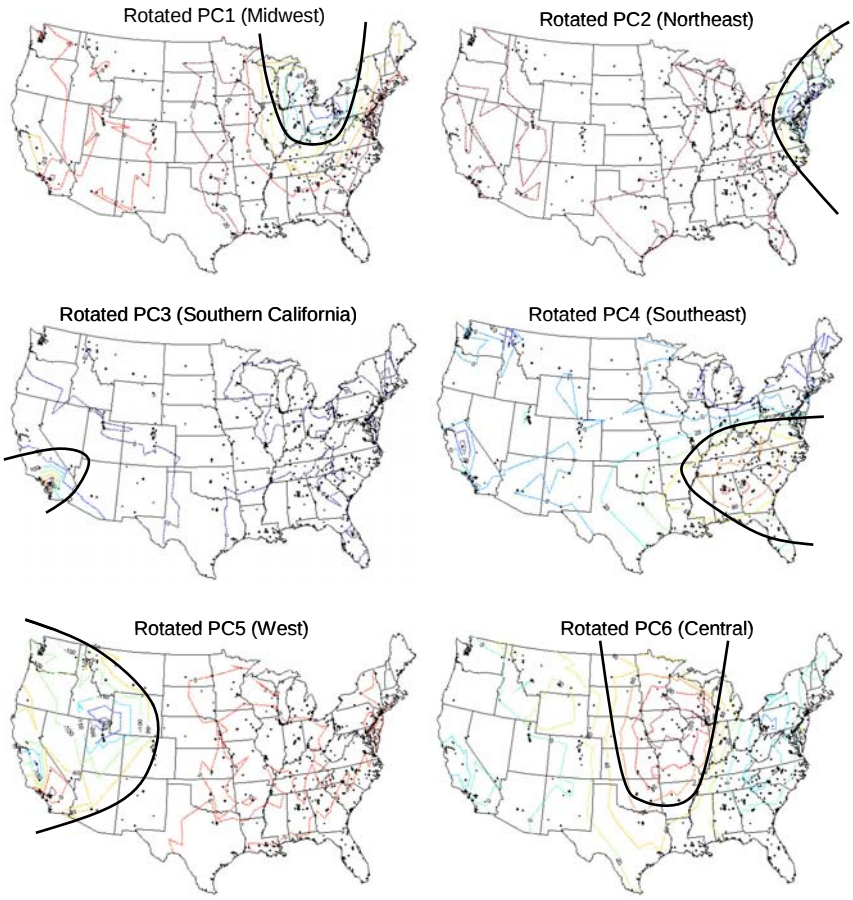


Figure 2.5. Contour plots of loadings from each of six RPCA.

Note that six PCs accounted for 65% of the variability of the entire dataset. A varimax rotation of the six PCs was performed. The components in the loading vectors of each of the six rotated PCs were displayed by contour plots on U.S. maps (Figure 2.5). The regions with higher loading values were highlighted. The first RPCA loading contour plot identified the monitoring sites in the Midwest. The second, third, fourth, fifth, and sixth RPCA loading contour plots identified the monitoring sites in the Northeast, Southern California, Southeast, West, and Central, respectively.

It is somewhat difficult to make a direct comparison between RPCA and k-means clustering analysis because of their different ways of determining the spatial groups of

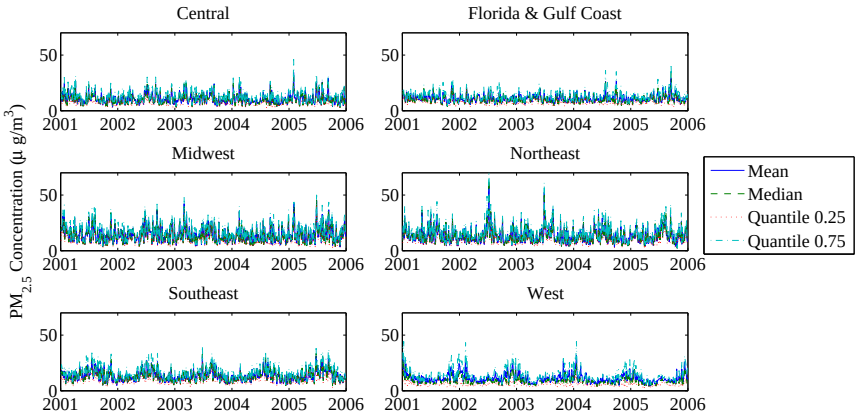


Figure 2.6. Mean, median, 25^{th} percentile, and 75^{th} percentile of temporal profiles for each clustered region.

homogeneous $PM_{2.5}$ concentrations. RPCA relies on a graphical interpretation of the contour plot of RPCA loadings, while k-means clustering analysis assigns a group label to each monitoring site. Note that Figure 2.3 is a plot of group labels from k-means clustering analysis. Nevertheless, identified homogeneous regions from RPCA and k-means clustering analysis seem similar. The main difference is that RPCA did not identify the sites in the Florida & Gulf Coast as a separate group but identified sites in Southern California. Both the RPCA and k-means clustering analysis are unsupervised learning techniques, in that they depend only on input variables (explanatory variables) but do not take into account the information from the response variable. However, from the mathematical point of view, RPCA and k-means clustering are different. RPCA identifies a new coordinate system that maximizes the variability of the original dataset through an orthogonal linear transformation, while k-means clustering analysis does not use any transformation processes but iteratively partitions the observations by minimizing within-group distances and maximizing between-group distances, then assigning a cluster label to each observation.

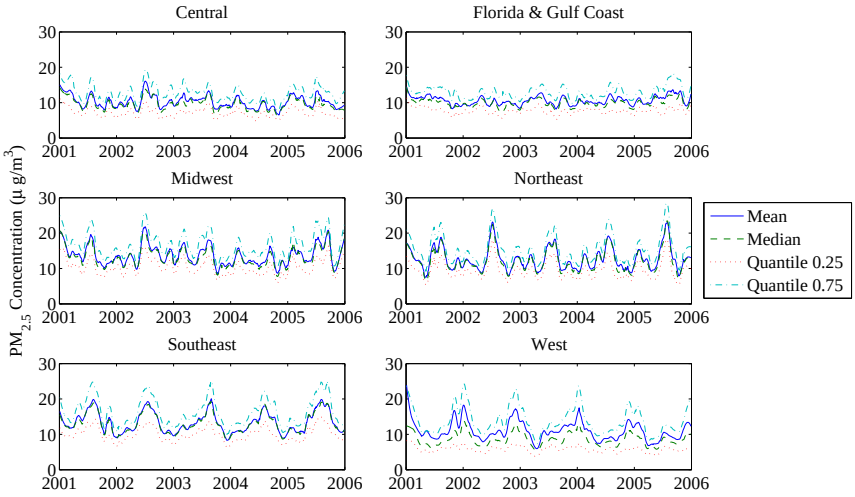


Figure 2.7. Smoothed mean, median, 25^{th} percentile, and 75^{th} percentile temporal profiles for each clustered region.

RPCA renders a graphical result, efficient in facilitating the visualization of a high-dimensional space. However, similar to other graphical methods, the interpretation of RPCA results can be subjective, with different analyzers drawing different conclusions. On the other hand, k-means clustering analysis provides a group label for each observation, and thus, the interpretation of results is more objective than RPCA. However, the k-means clustering results may vary with different choices of the starting means. No consensus exists about which is the better method (RPCA or clustering analysis) to satisfy all conditions. We believe that visualization methods, such as RPCA, can elicit the natural groupings of the observations, and clustering analysis can test the validity of the groupings obtained by RPCA. The following section discusses temporal and seasonal patterns of $PM_{2.5}$ concentrations according to k-means clustering results.

2.4.3 Temporal and Seasonal Patterns of $PM_{2.5}$ Concentrations

The temporal pattern and smoothed temporal pattern of each spatially homogeneous region identified via six-means clustering analysis over a time period from 2001 to 2005 is summarized using mean, median, 25^{th} percentile, and 75^{th} percentile profiles (Figure 2.6 and Figure 2.7). The rloess method with a span of 0.05 was used for smoothing the original time patterns. The similarity between the 25^{th} and 75^{th} percentile profiles confirms that there are no significant outliers in the dataset. A distinct temporal pattern was observed in each region. For ease of interpretation of temporal patterns and to explore seasonal variations, we defined the four seasons in a standard way: spring (March, April, May), summer (June, July, August), fall (September, October, November), and winter (December, January, February). Figure 2.4 shows the comparison of mean $PM_{2.5}$ concentrations for the four seasons. It can be seen that the highest mean concentration value was observed in summer, followed by winter for the period between 2001 and 2005. In particular, in 2002 and 2003, the mean concentrations in summer exceed the annual federal standard of 15 $\mu g/m^3$. The lowest mean concentration was observed in spring, except 2001. The results from Tukey's pair-wise comparisons test showed that the mean concentrations in every season were significantly different from each other (p-value $<$ 0.01).

It is important to observe from the box plots shown in Figure 2.8 that $PM_{2.5}$ concentrations between regions and seasons have interaction effects in that each clustered region differs in each of the four seasons (Figure 2.8). In the box plots, the lines in the middle of the boxes represent the median, and the distance between the top and bottom of the boxes represents the range from the 25^{th} to the 75^{th} percentiles (i.e., interquartile range). The plus sign at the top of the plot is an observation that is more than 1.5 times the interquartile range away from the top or from the bottom of the box.

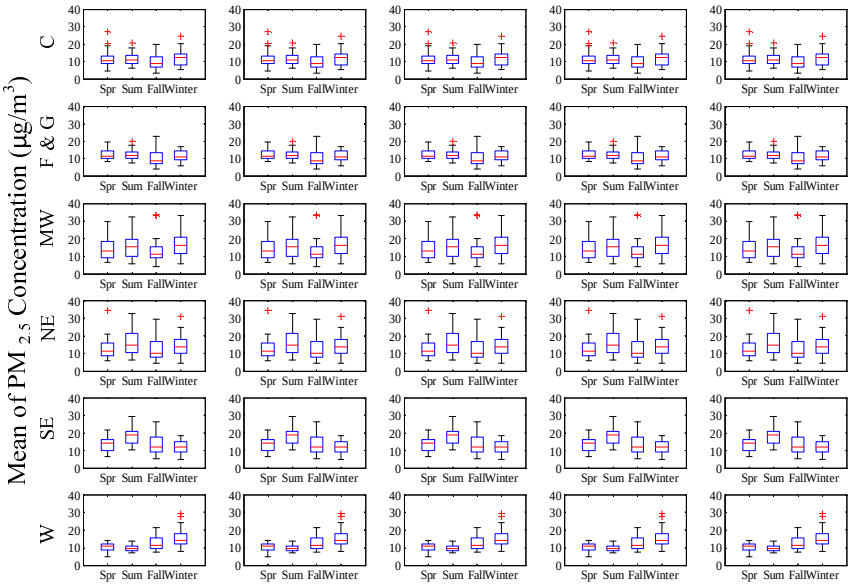


Figure 2.8. Box plots of the seasonal mean $PM_{2.5}$ concentrations in each region over the four seasons from 2001 to 2005.

According to Figure 2.8, the West region has the highest level of $PM_{2.5}$ in winter, likely because of the increase in NO_3^- and organic carbon during winter months. Major sources of NOx include transportation, industrial operations, electricity production, and non-industrial fuel burning. Quasi-equilibrium favors the particulate species under cool, moist conditions [45, 46]. This significant increase in the level of NO_3^- in the western United States in winter likely offsets the slight seasonal reduction of $SO_4^=$. A major source of organic carbon during wintertime in the western United States includes fireplace burning [47].

$PM_{2.5}$ concentrations tend to be higher in summer in many parts of the nation's northeastern and southeastern sections (Figure 2.8). Sulfate is produced from sulfur dioxide, which is prevalent in the East because of the relatively abundant coal-fired power plants [47]. Higher insolation and humidity during summer months enhance both homogeneous and heterogeneous reactions that produce secondary sulfate particles, one of the major components in $PM_{2.5}$ mass concentrations [48, 49].

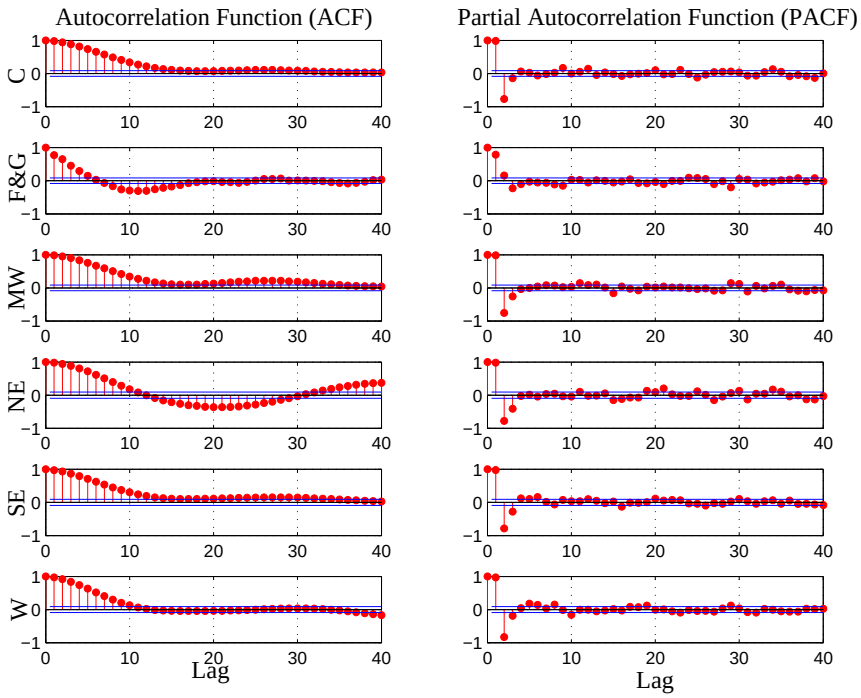


Figure 2.9. Autocorrelation and partial autocorrelation functions of the mean of smoothed time-series data (from 2001 to 2005) for each clustered region.

Midwest, Central, and Florida & Gulf Coast show comparable $PM_{2.5}$ levels during the four seasons, although the Midwest tends to show higher within-season variability than the Central and Florida & Gulf Coast regions.

To be able to predict $PM_{2.5}$ concentration as a function of time in each clustered region, time-series models were developed using the mean of smoothed time-series data (see Figure 2.7). The original time series shows a yearly or seasonal trend that causes a non-stationary time series. We subtracted the mean of each time series and used differencing to remove these trends and make the series stationary. To determine the time-series model, we used the Box-Jenkins graphical approach [50], which relies on the patterns of the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. Figure 2.9 shows ACF and PACF of the time-series data in each spatially homogeneous region. ACF slowly decays with either an exponential curve or sine waves,

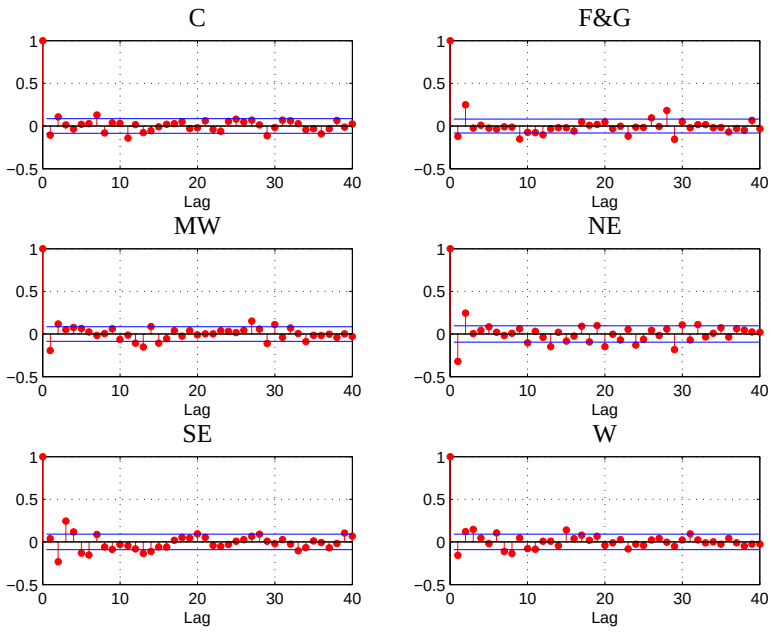


Figure 2.10. Autocorrelation of the residuals from time-series models.

while PACF has a large value for the first or second lag and becomes small (close to zero) for higher order lags. These patterns suggest that a first-order or second-order autoregressive (AR) model might be a good choice [50].

Table 2.2 summarizes time-series models with the estimated parameters for each clustered region. AR models consider a linear combination of past values and a Gaussian white noise term. AR(1) and AR(2) models are of the forms $Y_t = \phi_1 Y_{t-1} + Z_t$ and $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + Z_t$, respectively. Y_t is the $PM_{2.5}$ concentration at time t, ϕs are the parameters of the model, and Z_t is a Gaussian white noise series with mean zero and variance σ^2_{WM}. The parameters of the AR models can be estimated by the maximum likelihood estimation technique, available in many standard computer packages. In the present study, we used S-PLUS 6 (Insightful Corporation, Seattle, WA). To test the adequacy of the time-series model derived, the autocorrelation functions of the estimated residual values (e.g., $Y_t - \hat{\phi}_1 Y_{t-1}$ or $Y_t - \hat{\phi}_1 Y_{t-1} - \hat{\phi}_2 Y_{t-2}$) were generated (Figure 2.10).

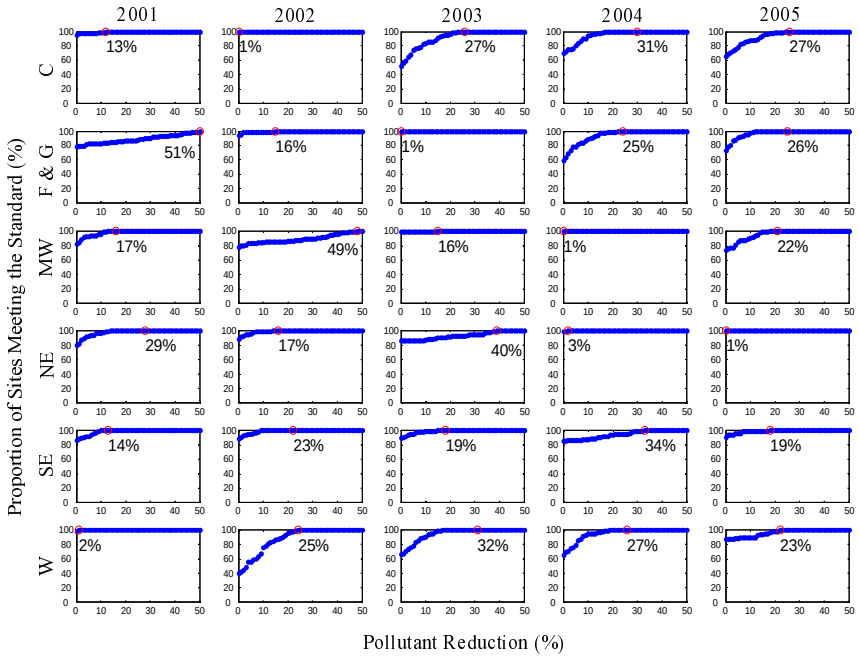


Figure 2.11. Percentage of sites meeting the federal standard for annual $PM_{2.5}$ levels.

Results show that only a few points out of 40 fall outside the bound, indicating that our derived time-series models fit the data well.

2.4.4 Comparison of Annual $PM_{2.5}$ Level of Each Spatially Homogeneous Region with the Federal Standard

Annual mean $PM_{2.5}$ concentrations for each clustered region were compared with the annual federal standard of 15.0 $\mu g/m^3$ (Figure 2.11). The x-axis shows the percent reduction required to meet the standard. For example, in 2005, in the Central region, 61 of 68 sites (89.7 percent) satisfied the federal standard, which corresponds to the y-axis value when the x-axis value of the plot is zero (Figure 2.11). It also shows that all sites in the Central region will satisfy the federal standard if an 18 percent reduction in pollutants is achieved for all sites in the region. The same analysis was performed for the other five clustered regions. The results showed that in 2005, 97.7 percent (Florida & Gulf Coast),

39.8 percent (Midwest), 65.4 percent (Northeast), 64.9 percent (Southeast), and 87.0 percent (West) of sites met the federal standard. To achieve the federal standard for all sites in each clustered region in 2005 would require pollutant reductions, by region, of 1 percent (Florida & Gulf Coast), 24 percent (Midwest), 31 percent (Northeast), 26 percent (Southeast), and 22 percent (West).

An overall pattern of pollutant reductions required in each clustered region seems similar over a period from 2001 to 2005. One clear pattern that emerged is that there were a relatively large proportion of nonattainment sites in 2001 and 2005 compared to 2002, 2003, and 2004.

Interestingly, the regions with a large proportion of nonattainment sites did not always require large amounts of pollutant reduction to satisfy the federal standard. A comparison of the Midwest and Northeast regions in 2005 provides a good example. In the Midwest region, only 39.82 percent of sites met the federal standard, but 65.38 percent in the Northeast met the standard. However, more efforts seemed to be required in order to achieve the federal standard for all sites in the Northeast than in the Midwest region. This implies that the number of sites exceeding the federal standard does not correlate directly with the percent of pollutant reduction required. These results indicate that different pollutant management programs should be applied to specific times and regions.

2.5 Conclusion

The present study examines the temporal patterns of $PM_{2.5}$ concentrations over the period from 2001 to 2005 across the continental U.S., so as to characterize spatially homogeneous regions. The k-means clustering algorithm using the correlation distance enabled us to measure the similarity of overall temporal patterns among 522 monitoring sites. We believe k-means clustering analysis can be useful as an alternate approach

to test the validity of the groupings obtained by visualization methods, such as RPCA, which has been used for characterizing spatial patterns in air pollution and meteorological fields. The k-means clustering analysis grouped the sites in close spatial proximity. More precisely, the analysis resulted in six spatial regions that exhibit homogenous temporal $PM_{2.5}$ concentration patterns over multiple years: Central, Florida & Gulf Coast, Midwest, Northeast, Southeast, and West. In each spatially homogenous region, distinct temporal patterns were observed. In general, higher $PM_{2.5}$ concentrations occur in winter in the western part of the United States, but in summer in the northeastern and southeastern regions. These results are generally consistent with other existing studies indicating the higher levels of NO_3^- and organic carbon in the west during winter and $SO_4^=$ in the east during summer. The results also indicate that $PM_{2.5}$ concentrations vary from year to year. This may due to meteorological variations or consequences of major human- or nature-related activities. To obtain more understanding of the observed time-series patterns, we fit time-series models based on the Box-Jenkins' graphical approach. Time-series models with mean-centered and differenced data provided AR(1) or AR(2) model for each of six clustered (homogenous) regions. Residual analysis confirmed the adequacy of the derived models. These time series models can be used to predict the future $PM_{2.5}$ mass concentrations in a regional scale. Finally, we showed the amounts of pollutant reduction required to meet the federal standard for all sites in each clustered region from 2001 to 2005.

Table 2.1. A list of states in the United States in each clustered region

Clustered Region	**Number of States**	**States**
Central	12	North Dakota, South Dakota[a], Nebraska[a], Kansas, Oklahoma, New Mexico[a], Texas[a], Minnesota, Iowa[a], Missouri, Arkansas[a], Illinois[a]
Florida & Gulf Coast	6	Texas[a], Louisiana[a], Alabama[a], Georgia[a], South Carolina[a], Florida
Midwest	6	Iowa[a], Wisconsin, Illinois[a], Indiana, Michigan, Ohio[a], New York[a], Pennsylvania[a], Maine
Northeast	15	Ohio[a], West Virginia[a], Virginia[a], Pennsylvania[a], New Jersey, Delaware, Maryland, Connecticut, New York[a], Massachusetts, Rhode Island, Vermont, New Hampshire, Maine, Montana[a]
Southeast	11	Arkansas[a], Louisiana[a], Tennessee, Mississippi , Alabama[a], Georgia[a], South Carolina[a], Virginia[a], West Virginia[a], Kentucky, California[a]
West	14	Washington, Oregon, California[a], Nevada, Idaho, Montana[a], Wyoming, Utah, Arizona, Colorado, New Mexico[a], Texas[a], South Dakota[a], Nebraska[a]

[a]Sites in these states are split into more than one clustered region.

Table 2.2. Time-Series models with the estimated parameters in each clustered region

Clustered Region	Time-Series Model	$\hat{\phi}_1$	$\hat{\phi}_2$
Central	AR(2)	1.750	-0.775
Florida & Gulf Coast	AR(1)	0.783	-
Midwest	AR(2)	1.733	-0.757
Northeast	AR(2)	1.749	-0.777
Southeast	AR(2)	1.271	-0.434
West	AR(2)	1.796	-0.829

CHAPTER 3

SPATIAL PREDICTION OF THE OZONE CONCENTRATION PROFILES

3.1 Introduction

Ground level ozone is one of the major air pollutants in many urban areas. The major sources of the precursors of ground level ozone are emissions from industrial facilities, motor vehicle exhausts, and electric utilities, all of which emit volatile organic compounds (VOC) and oxides of nitrogen (NOx). In the presence of heat and sunlight, these precursors undergo a chemical reaction that results in the formation of ground level ozone. The ozone formed adversely affects ecosystems; its results show up in stunted growth and lessened survivability of plants and animals and in reduced plant yields. Furthermore, ozone is known to be associated with adverse health effects in humans such as acute respiratory problems, chest pain, asthma, and inflammation of lung tissue [51].

The U.S. Environmental Protection Agency (EPA) has National Ambient Air Quality Standards (NAAQS) for six pollutants that are known as criteria pollutants. Ozone is one of them. All U.S. states are mandated to comply with the standards set by EPA for these six criteria pollutants. Exceeding the NAAQS results in non-attainment status for a region for that specific pollutant. The eight-hour standard for ozone is 75 parts per billion (ppb), which is the three-year average of the fourth highest daily maximum eight-hour ozone concentration [52]. Because Dallas-Fort Worth (DFW) has non-attainment status for ozone, area officials have initiated a system to warn residents of high ozone levels so that residents can curtail their outdoor activities. The warnings are based on predicted and interpolated ozone levels in the atmosphere. Such an ozone warning pro-

gram can be extremely beneficial to the public, especially to sensitive populations, e.g., children and older people. Therefore, an appropriate ozone prediction model is necessary for public safety. In addition to warnings of high ozone levels, officials should consider implementation of voluntary programs to reduce emissions on high ozone days [53].

In general, statistical models for air pollution focus on two major aspects of the problem: time series prediction [54, 55, 56, 19] and spatial predictions [6, 18, 34]. Yi and Prybutok [54] compared the performance of an artificial neural network (ANN) and a Box-Jenkins auto regressive integrated moving average (ARIMA) model to predict the daily maximum summer ozone concentration in DFW. The result showed that an ANN outperformed the Box-Jenkins ARIMA model. [55] compared the capabilities of linear regression, a regression tree, and an ANN to predict hourly surface ozone concentrations. They concluded that the ANN outperformed the two other models. They attributed the ANN's superior performance to its allowing arbitrary interactions and nonlinear relationships between predictor variables. However, the physical relationship among predictors cannot be readily interpreted from ANN models. Later these same researchers also used ANN models to investigate whether any discernible temporal and spatial trends can be detected in response to changes in the amount precursor emissions. They found that since 1994 meteorologically adjusted summer daily maximum ozone concentrations have been in general decline in the United Kingdom [57]. [56] constructed ANN models based on principal components to predict the next day's hourly ozone concentrations in Oporto, Portugal. They found that the use of principal components as inputs reduced model complexity and eliminated data collinearity. [19] found that in Houston, Texas, where the effect of midday wind is critical but difficult to model parametrically, a loess/generalized additive model outperformed linear, nonlinear regression, and ANN models.

Spatial predictions require an understanding of the spatial correlations of pollutant concentrations. The main goal of such predictions is to predict the concentrations in

unsampled locations. A number of studies have been undertaken at various locations to understand of spatial correlations. [18] constructed a fixed-effect analysis of variance model as a way to identify potential monitoring sites in Seattle, Washington, that can represent the characteristics of particulate matter 2.5 (particulate matter that is 2.5 micrometers or smaller in size). [34] developed a statistical application to determine a representative monitoring station for air quality measurement in Taipei, Taiwan. [6] evaluated the performance of the Carnegie/California Institute of Technology (CIT) model and found that spatial inhomogeneity needed to be considered in the development of model performance guidelines.

Recently, air quality modeling has simultaneously taken into account both spatial and temporal variability. [35] used space-time correlations to compare the results of daily $SO_4^=$ levels of observed data and those of the community multiscale air quality (CMAQ) model. The CMAQ model matches the space-time correlation structure of the observed data; however, the CMAQ partially captures the time-lagged spatial variation of $SO_4^=$ concentrations. [39] exploited the hierarchical Bayesian approach to predict the cell-average ozone concentration across the European region and found that a relevant fraction of the model's bias can be explained by subgrid spatial variability. [58] constructed a spatio-temporal model to predict the ozone concentration in Mexico City. They first employed a univariate time series analysis within the Bayesian framework to forecast the temporal components. The forecasted temporal components were then used in a Markov Chain Monte Carlo method to predict the ozone concentrations of an unsampled location.

Although the researchers' proposed methods performed reasonably well in the situations studied, no consensus exists about which of them best satisfies all conditions encountered in various environmental problems ([59]). In particular, existing studies have attempted to predict pollutant concentrations at a particular location and a time. To the best of our knowledge, no study has been conducted to achieve spatial predic-

tion of the time-series profile of pollutant concentrations. Further, although efficiency and accuracy obviously can be improved by including meteorological variables that significantly affect pollutant concentrations, existing procedures do not capitalize on this information. The present study proposes a statistical procedure that uses multiscale and functional modeling of available meteorological information as well as ozone concentrations to improve the spatial prediction of the ozone concentration profiles in the DFW area. It should be noted that even though the proposed procedure focuses on ozone concentrations, our procedure could also be applied to other application areas with spatial and temporal monitoring data.

3.2 Data

Monitoring data were obtained from the database maintained by the Texas Commission on Environmental Quality (TCEQ). The database (www.tceq.state.tx.us) contains daily maximum eight-hour ozone concentrations and the following six meteorological variables: (1) daily maximum temperature, (2) daily maximum solar radiation, (3) daily average wind gusts, (4) daily average resultant wind direction, (5) daily average resultant wind speed, and (6) daily average wind speed. This study focused on 14 monitoring sites in the DFW area (Figure 3.1) from September 10, 2003, to June 30, 2006 (1,024 time points).

Missing observations and outliers were replaced with interpolated values by using an inverse distance weighted (IDW) method ([6]). The interpolated value for site S_i at time T_j, $I(S_i, T_j)$ was computed as follows:

$$I(S_i, T_j) = \frac{\sum_{k=1, k\neq i}^{m} Z(S_k, T_j) \cdot \omega_k}{\sum_{k=1, k\neq i}^{m} \omega_k} \tag{3.1}$$

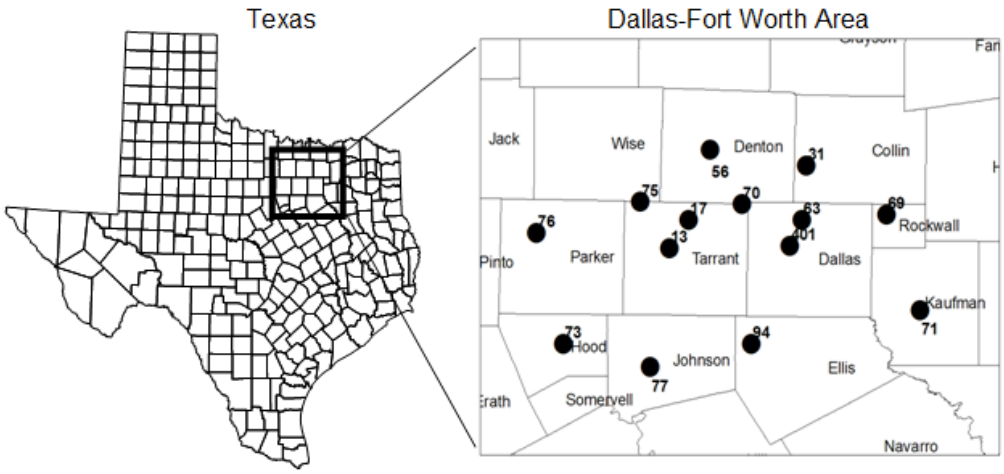


Figure 3.1. Locations of 14 ozone monitoring sites in the DFW area.

where m is the number of monitoring sites and ω_k is calculated as follows:

$$\omega_k(S_i) = \frac{1}{r_k^2}, \tag{3.2}$$

where r_k^2 is the Euclidean distance from monitoring site S_i to S_k at time T_j. Thus, $I(S_i, T_j)$ in (3.1) is the weighted average pollutant concentration observed in the surrounding m sites. The weights are determined by the way that observations in close spatial proximity are given more weight than those that are spatially separated.

3.3 Analytical Approaches

3.3.1 Overview

In the present study we propose a mutiscale and functional modeling procedure that takes advantage of available meteorological variables for spatial prediction of the ozone concentration profiles in the DFW area. An overview of the proposed procedure is shown in Figure 3.2. The procedure starts with stepwise regression to select the important meteorological variables to include in the model. Wavelet transformation decomposes the selected meteorological variables and the ozone concentration variables

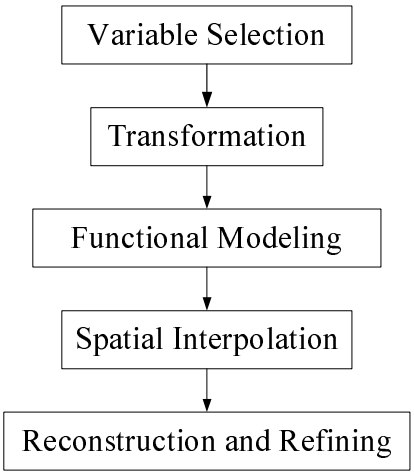


Figure 3.2. Overview of the analytical procedure.

into a multiscale of wavelet coefficients. The functional coefficients are obtained from the regression analysis of wavelet coefficients. Kriging is then used to predict the regression coefficients of unsampled locations. Predicted regression coefficients are reconstructed and refined to obtain the final ozone concentration profile of the unsampled location.

3.3.2 Wavelet Transforms

Wavelet transforms have the advantage of localizing the analysis to handle multi-scale information efficiently. Wavelet transforms analyze the profile by dividing it into segments of scale and by finding the correlation among these segments and the scale-dependent finite energy functions in which the maximum number of scales depends on the availability of data. The discrete wavelet transform (DWT) of a profile x is defined as

$$y_l(k) = \sum_{t=-\infty}^{\infty} x(t)\phi(t-k), \tag{3.3a}$$

$$y_{h_j}(k) = \sum_{t=-\infty}^{\infty} 2^{j/2} x(t)\psi(2^j t-k), \tag{3.3b}$$

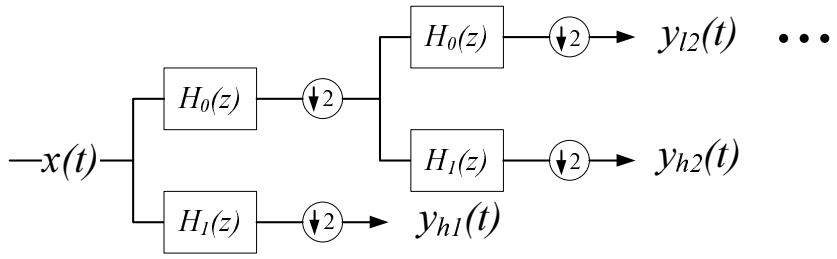


Figure 3.3. Diagram of discrete wavelet transformation. ↓ 2: downsampling by a factor of two.

where $x(t)$ represents the original profile in time domain, $y_l(k)$ represents the scaling coefficients of DWT, y_{h_j} represents wavelet coefficients of level j. $\phi(t)$ and $\psi(t)$ are the scaling and wavelet functions, and j is the level of decomposition. Figure 3.3 displays a diagram of DWT, showing that the wavelet coefficients (e.g., y_{l_j} and y_{h_j}) are obtained through a series of lowpass (e.g., $H_0(z)$) and highpass filters (e.g., $H_1(z)$). As in most practical multi-resolution signal analysis, we chose a factor of two ($\downarrow$ 2) in all the downsampling operations for the convenience of implementation and easy indexing.

Although a number of studies of wavelet analysis have been conducted in the general field of signal and image processing ([60, 61]), wavelets have not been thoroughly studied for application to air pollution. [62] applied wavelet analysis with Morlet mother wavelets to characterize the variation of total ozone concentrations and solar radio emissions. [63] applied wavelet analysis to isolate intermittent turbulent bursts within the vertical velocity of an ozone time series. [64] also employed the Morlet wavelet spectra and Mexican hat wavelet spectra, based on zonal averaged total ozone content, to study solar rotational activity effects on ozone. Further, [65] constructed wavelet networks based on Gaussian wavelets for short-term prediction of maximum ozone concentrations. Wavelet networks can be explained as neural networks in which each neuron is replaced by wavelets in which the translation and dilation parameters are iteratively adjusted.

However, all the studies mentioned earlier used the continuous wavelet transform (CWT) in which the number of wavelet coefficients is larger than the number of original time signals. Some information from CWT may be unnecessary for reconstructing the original time signal, which implies that the CWT is redundant after all. To avoid this redundancy, we propose the use of Daubechies wavelets, one of the DWTs. Daubechies is one of the most efficient discrete wavelets of those that provide the maximally flat frequency response for a given vanishing moment. The family of Daubechies wavelets (db) can be classified according to their designed vanishing moments, e.g., db4 has four vanishing moments. Figure 3.4 shows Daubechies' scaling and wavelet functions with different vanishing moments, illustrating that the more vanishing moments a wavelet has, the more frequency selective it is and the greater its computational complexity.

[66] proposed an approach for air pollution modeling based on DWT in which Daubechies filters with one and six vanishing moments were applied to construct a model to simulate transport and photochemical reactions in the atmosphere. However, no single filter performs best for all applications ([67]). Choosing a filter requires in-depth knowledge of the intended application and data characteristics. In the present study we propose to use Daubechies with four vanishing moments (db4) that offer reasonable frequency response and efficient computational complexity as a way to model ozone concentrations in the DFW area. The original data in the time domain, $x(t)$, can be reconstructed using inverse discrete wavelet transformation (IDWT). The process for reconstructing original time series from the wavelet coefficients is illustrated in Figure 3.5. It can be seen that the y_{h_j} and y_{l_j} are fused together using a series of $G_1(z)$ and $G_0(z)$ filters until the original time series $x(t)$ is obtained.

Perfect reconstruction can be computationally expensive. Thresholding, one of the data reduction methods, can parsimoniously represent the original profile while using only a small number of wavelet coefficients. The basic idea of thresholding is to zero out

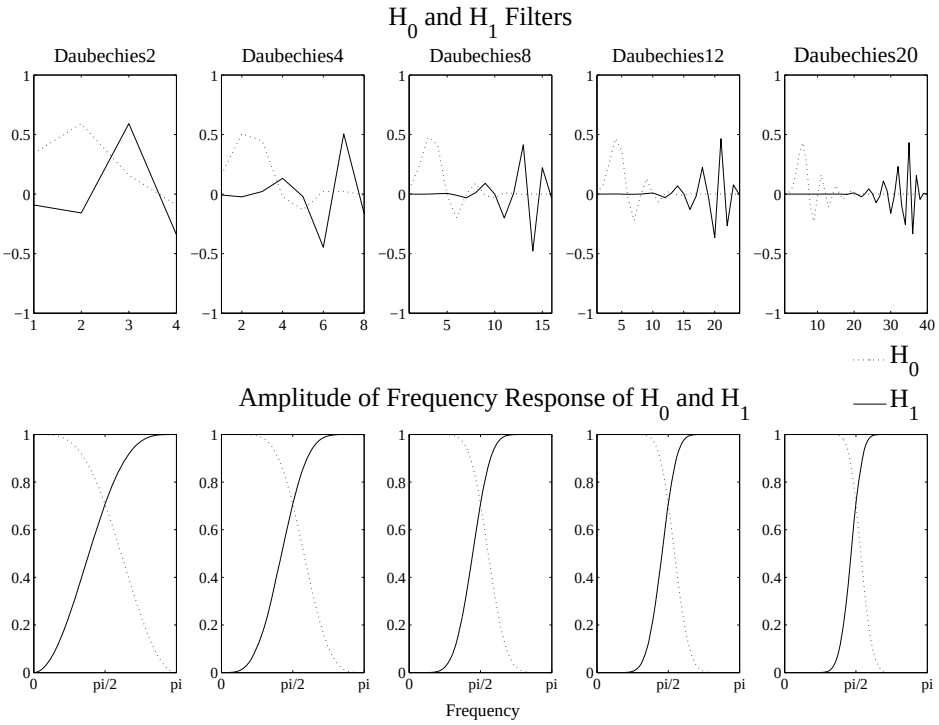


Figure 3.4. Daubechie scaling (H_0) and wavelet (H_1) functions.

the small magnitude wavelet coefficients in the wavelet transform domain. The main challenge with thresholding is how to determine the threshold value at which the coefficients will be discarded. A very large thresholding constant makes it difficult for a coefficient to be included in the profile reconstruction, which results in an over-smoothing of the profile. On the other hand, choosing a very small thresholding constant value allows many coefficients to be included in the reconstruction, yielding a result close to the original noisy profile. The literature contains many wavelet model selection procedures that are based on the idea of selecting important wavelet coefficients. These include VisuShrink ([68]), SureShrink ([68]), and AMDL ([69]). Although the thresholding algorithms can be efficient for reconstruction of an original profile with a few coefficients, these coefficients may not always yield good predictive accuracy. Indeed, in our problem, better predictive

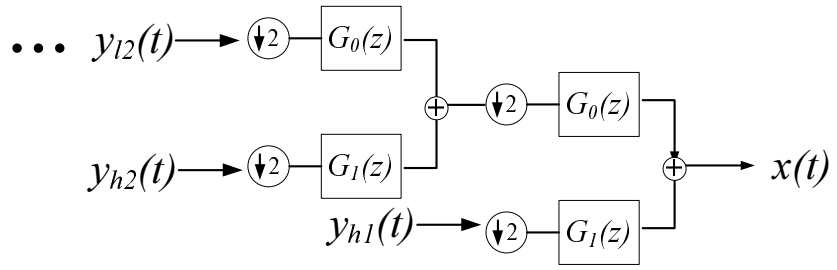


Figure 3.5. Diagram of inverse discrete wavelet transform, ↓ 2: downsampling by a factor of two.

accuracy is more important than aggressive data reduction. In the present study we propose to use an explanatory thresholding compromise between predictive accuracy and data reduction to find an appropriate set of wavelet coefficients. More details about such an explanatory thresholding method are presented in Section 3.4.7.

3.3.3 Functional Data Analysis for the Ozone-Concentration Profiles

Regression analysis was used to model a functional response ([70]). Multiple linear regression analysis is employed to model the wavelet coefficients of the ozone concentration profile as a function of the wavelet coefficients of the profiles of meteorological variables. The functional response of wavelet coefficients can be represented in the form of a traditional regression model as follows:

$$y_{(n\times 1)} = X_{(n\times p)}\beta_{(p\times 1)} + \epsilon_{(n\times 1)}, \tag{3.4}$$

$$\hat{y} = X\hat{\beta}_{(p\times 1)}, \tag{3.5}$$

where y is a functional response of wavelet coefficients. X is a matrix of wavelet coefficients of predictors. β is a vector of regression parameters $[\beta_0, \beta_1, ..., \beta_{p-1}]^T$. ϵ is a vector of normal independent random variables with expectation $E[\epsilon] = 0$ and a constant variance-covariance matrix. $\hat{y}$ denotes a vector of the fitted values of y, and $\hat{\beta}$ is a vector of least squares estimated regression coefficients β, $[\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{p-1}]^T$. These coefficients

characterize and summarize the relationship between the ozone concentration and meteorological variables and would be used for spatial prediction as described in the following section.

3.3.4 Spatial Prediction of the Ozone-Concentration Profiles

Kriging is one of the most widely used spatial prediction algorithms ([71, 72]). Typically, the kriging weights assigned to the surrounding data points are estimated as an inverse function of the distance of these points from the unsampled location; this result in the data points close to the unsampled location carrying more weight than those at remote points. The major advantage of kriging is that if some data points cluster together, kriging attempts to view that cluster as a single point ([73]). The basic form of kriging can be defined as

$$f^*(i) = \sum_{s=1}^{n(s)} \lambda(i_s) f(i_s), \tag{3.6}$$

where $f^*(i)$ is the predicted value, and $\lambda(i_s)$ is the weight corresponding to the available data point $f(i_s)$. In general, (3.6) can be explained by the standard form of kriging as follows:

$$Z^*(i) - \mu(i) = \sum_{s=1}^{n(s)} \lambda(i_s) \left[Z(i_s) - \mu(i_s) \right], \tag{3.7}$$

where $Z^*(i)$ is realization of an unsampled location. $\mu(i)$ and $\mu(i_s)$ are, respectively, the mean values of $Z(i)$ and $Z(i_s)$. $n(s)$ is the number of available locations. $Z(i_s)$ are the weight and the realization at location i_s. i_s represents the location vectors, in this case the latitude and longitude, from available monitoring sites, and i represents the location vectors on a random field in the kriging model. Thus, $Z(i)$ can be considered as a random field with a mean of $\mu(i)$.

Because our study attempted to construct a predictive model for an unsampled location, we may assume that the mean component of the realization $Z(i)$ is constant, $\mu(i) = \mu$. Therefore, (3.7) can be written as

$$Z^*(i) - \mu = \sum_{s=1}^{n(s)} \lambda(i_s) \left[Z(i_s) - \mu \right], \tag{3.8}$$

and the error variance is of the form

$$\begin{aligned} \sigma_E^2(i) &= \mathrm{Var}[e^*(i)] + \mathrm{Var}[e(i)] - 2\mathrm{Cov}[e^*(i), e(i)] \\ &= \sum_{s=1}^{n(s)} \sum_{u=1}^{n(s)} \lambda(i_s)\lambda(i_u) C(i_s - i_u) + C(0) - 2 \sum_{s=1}^{n(s)} \lambda(i_s) C(i_s - i). \end{aligned} \tag{3.9}$$

To obtain the kriging weights (λ), it is essential to minimize the error variance by taking the derivative of (3.9) with respect to the kriging weights and setting it to zero, then solving the system of equations.

3.4 Results

3.4.1 Variable Selection

An analysis without a variable selection process may contain redundant predictor variables that have the potential to cause deterioration of the accuracy of the prediction. Consequently, only those predictor variables with a high contribution to predict the response variable but less correlation among predictor variables should be selected for subsequent analyses. A stepwise regression approach searches different subsets of predictors to find the best regression model. The selection criteria for stepwise regression can be F-tests, t-tests, adjusted R-square, Akaike information criterion (AIC), Bayesian information criterion (BIC), and Mallows' Cp. The present study uses a stepwise regression approach based on AIC and selects the following four meteorological variables that are most predictive of the given response variable (site-average ozone concentrations):

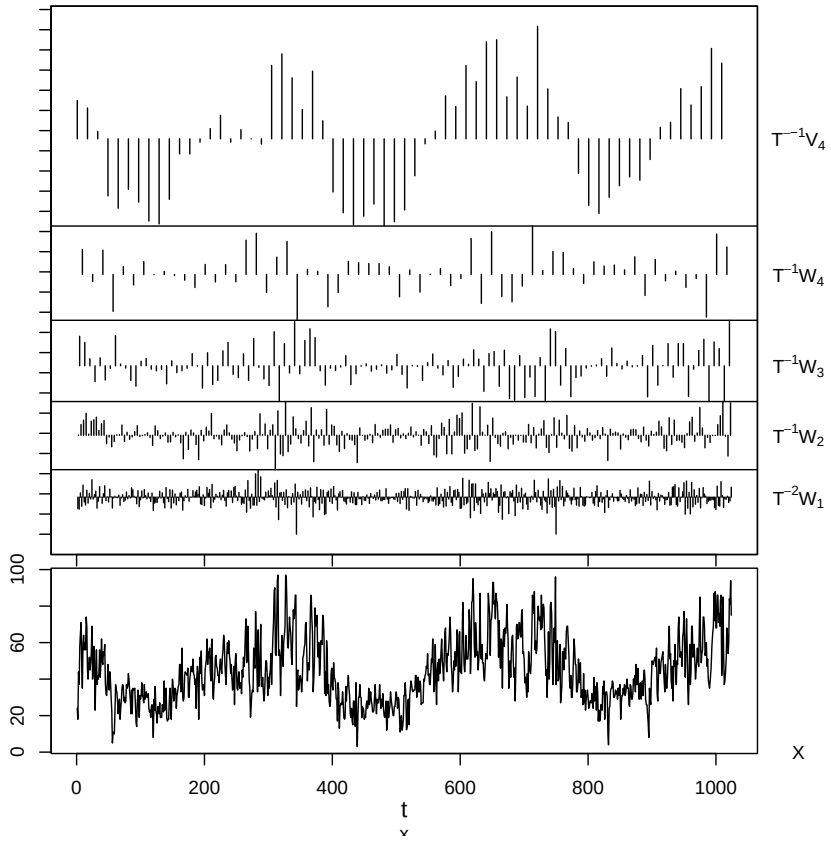


Figure 3.6. Wavelet decomposition of the ozone-concentration profile at CAM63.

(1) daily maximum temperature, (2) daily maximum solar radiation, (3) daily average resultant wind direction, and (4) daily average wind speed.

3.4.2 Wavelet Transforms

Ozone concentrations and selected meteorological predictors were decomposed to obtain five levels of wavelet coefficients, ranging from the finest to the coarsest levels. In Figure 3.6, the lowest plot shows an example of the original ozone concentration profile of CAM63. The second lowest plot represents the finest scale of wavelet decomposition up to the highest plot, which represents the coarsest scale of wavelet decomposition. The finest scale of wavelet coefficients can be interpreted as frequently occurring instances

such as noise, and the coarser scale of wavelet coefficients are those instances that occur less frequently. The coarsest scale of wavelet coefficient in this paper can be viewed as a trend pattern of profiles, either of ozone concentrations or of meteorological parameters.

3.4.3 Regression Analysis for Functional Modeling

Having found the wavelet coefficients, these coefficients were used to construct the regression model of ozone concentrations as a function of meteorological variables. With 13 training sites, $s = 1, 2, ..., 13$, and five levels of wavelet decomposition, $j = 1, 2, ..., 5$, (3.4) and (3.5) can be rewritten as follows:

$$y_{sj} = X_{sj}\beta_{sj} + \epsilon_{sj}, \tag{3.10}$$

$$\widehat{y_{sj}} = X_{sj}\hat{\beta}_{sj}, \tag{3.11}$$

where y_{sj} is a functional response of wavelet coefficients. X_{sj} is a matrix of wavelet coefficients from the selected meteorological variables. β_{sj} is a vector of regression parameters. ϵ_{sj} is a vector of normal independent random variables. $\widehat{y_{sj}}$ denotes a vector of the fitted value of y_{sj}, and $\hat{\beta}_{sj}$ is a vector of least squares estimated regression coefficients β_{sj}. Instead of a constructed model based on the original data points $y_{sj}(k_{sj})$ that incurred 13,312 data points, regression modeling reduced the number of data points to only 325 by storing the information of all 13,312 wavelet coefficients into 325 least squares estimated regression coefficients ($\hat{\beta}_{sj}$), which represents 97.5% data compression.

3.4.4 Kriging Model for Spatial Prediction

Kriging models were constructed to obtain regression coefficients of an unsampled location. With five scales of wavelet decomposition, (3.8) can be rewritten as

$$b_j^* = \mu_j + \sum_{s=1}^{n(s)} \lambda(i_s, j)\left[b_{sj} - \mu_j\right], \tag{3.12}$$

where b_j^* represents a vector containing least squares estimated regression coefficients of an unsampled location from wavelet level j. μ_j represents the mean of b_j^*. $\lambda(i_s, j)$ represents the kriging weights of training site i_s, and b_{sj} represents a vector containing least squares estimated regression coefficients from training sites.

3.4.5 Reconstruction of the Ozone-Concentration Profiles

In order to understand the physical meaning of the result from the spatial prediction, it is necessary to reconstruct the original ozone concentration profile of an unsampled location in the time domain. From (3.11), X_{sj} and b_{sj} can be replaced by X_j^* and b_j^*, which are the wavelet coefficients matrix of predictors and the predicted regression coefficients, to obtain the wavelet coefficients of an unsampled location, $\widehat{y_j^*}$, as follows:

$$\widehat{y_j^*} = X_j^* b_j^*. \tag{3.13}$$

Once sets of predicted wavelet coefficients at an unsampled location were obtained, the predicted ozone concentration profile at an unsampled location, $x^*(t)$, can be reconstructed using IDWT. For mathematical convenience, we represent $x^*(t)$ in wavelet form as follows:

$$x^*(t) = \sum_{k=-\infty}^{\infty} \hat{y_l^*}(k)\phi(t-k) + \sum_{j=0}^{\infty}\sum_{k=-\infty}^{\infty} \hat{y_{h_j}^*}(k)2^{j/2}\psi(2^j t - k). \tag{3.14}$$

where $\hat{y^*}$ is the predicted wavelet coefficients of ozone concentration at an unsampled location.

Figure 3.7 shows five levels of predicted wavelet coefficients at CAM63, which is considered as an unsampled site. Because there are 14 monitoring sites, we left one site out in each experiment to compute predictive accuracy. To be specific, one monitoring site was reserved for testing, and the remaining 13 sites were used for training. This process was repeated 13 more times, alternating the testing sites, to obtain the predictive

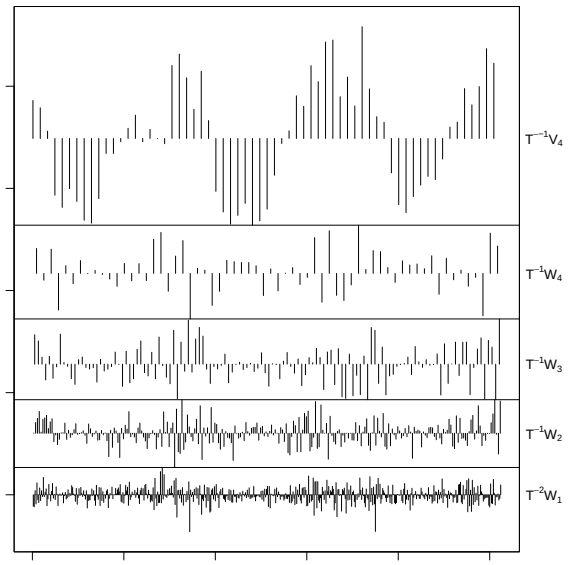


Figure 3.7. Predicted wavelet coefficients of each of five levels in CAM63.

accuracy for each site. The final predictive results from the 14 different testing samples were then averaged to obtain the overall accuracy of the proposed procedure. The predicted ozone concentration profiles of 14 monitoring sites are shown in Figure 3.8 and Figure 3.9, showing that the predicted profiles of ozone concentration performed excellently in representing the overall behavior of ozone concentrations. However, it could not reflect the high variability of ozone concentrations during the ozone season in the DFW area (from May to October).

3.4.6 Model Refining

In order to achieve higher predictive accuracy during the ozone season, it is necessary to further refine the procedure. To that end, we first calculated the difference between the predicted ozone concentration profile of an unsampled location and the actual ozone concentration profiles of training sites, $\varepsilon(t, s)$ as follows:

$$\varepsilon(s,t) = x(s,t) - x^*(t), \tag{3.15}$$

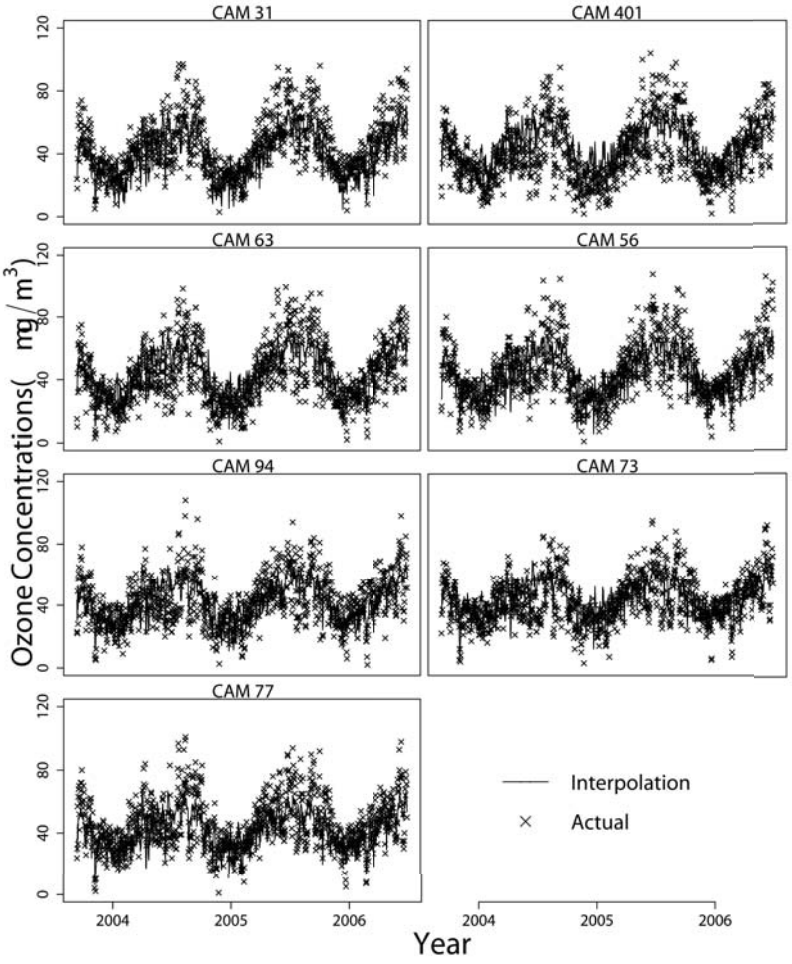


Figure 3.8. Actual vs. predicted ozone-concentration profiles of the monitoring sites in the DFW area (CAM31, CAM63, CAM94, CAM77, CAM401, CAM56, CAM73).

where $x(t, s)$ is the actual ozone concentration of training sites ($s = 1, 2, \cdots, 13$) and $x^*(t)$ is the predicted ozone concentration of an unsampled location. Figure 3.10 shows the 13 profiles (overlayed) of the resulting differences.

Because the variance of $\varepsilon(s, t)$ is not constant over time, we grouped the $\varepsilon(s, t)$ according to their variability. The $\varepsilon(t, s)$, which falls between the solid lines (Figure 3.10), illustrates constant variance with low variability, and the $\varepsilon(s, t)$, which falls between the dashed lines, illustrates constant variance with high variability. Other $\varepsilon(t, s)$ that do not

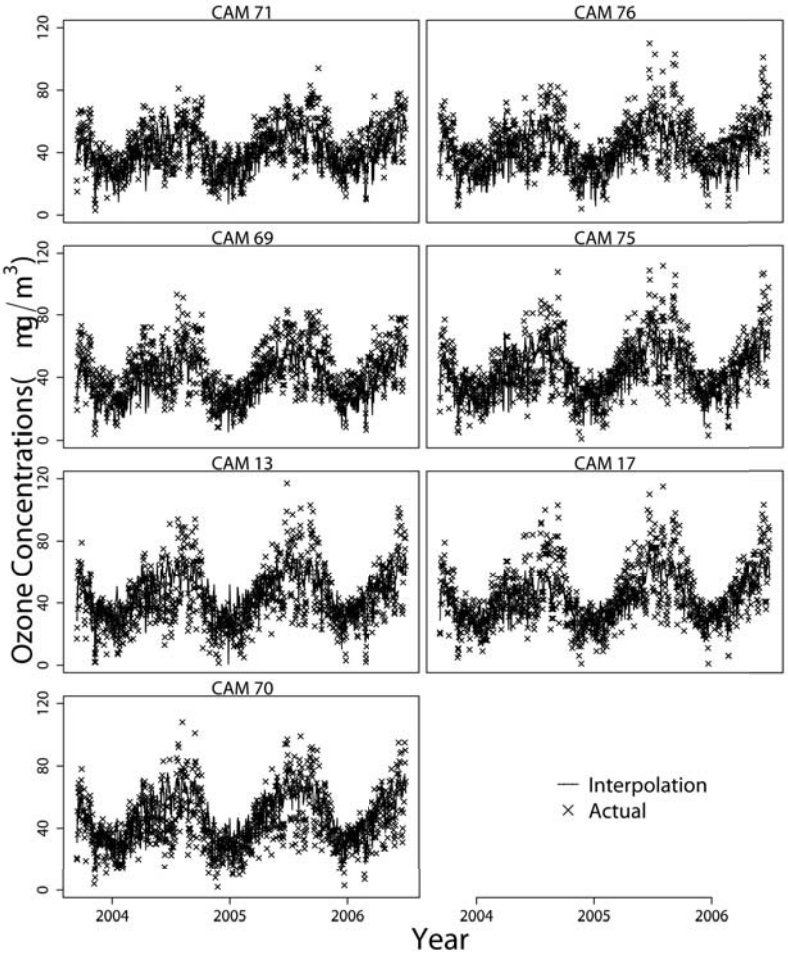


Figure 3.9. Actual vs. predicted ozone-concentration profiles of the monitoring sites in the DFW area (CAM71, CAM69, CAM13, CAM70, CAM76, CAM75, CAM17).

fall between any lines appear to be the transition periods between the times of high and low variability.

To effectively analyze the profiles of $\varepsilon(t, s)$, they were decomposed into three-level wavelet decomposition under the same wavelet functions previously used (i.e., db4). Those three levels of wavelet decomposition were intended to represent the three groups of variability: low, transitional, and high. Figures 3.11 and 3.12 show examples of the decomposed wavelet coefficients of $\varepsilon(s, t)$ before and after thresholding. To achieve the high-

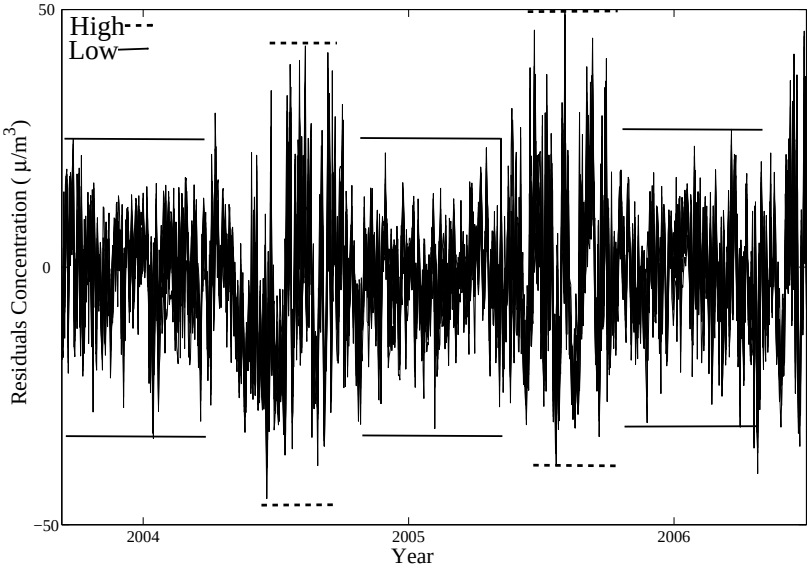


Figure 3.10. The difference between the predicted ozone-concentration profile of an unsampled location and the actual ozone-concentration profile of training sites.

est accuracy from the wavelet decomposition, we used the information from all wavelet coefficients by calculating the site-average wavelet coefficients, $\hat{u}_j(t)$,

$$\hat{u}_j(t) = \frac{1}{n(s)} \sum_{s=1}^{n(s)} u_{sj}(t), \tag{3.16}$$

where $u_{sj}(t)$ is the wavelet coefficients of the difference between the predicted ozone concentration of an unsampled location and the actual ozone concentration at time t from training sites $s = 1, 2, ..., 13$ and wavelet level $j = 1, 2, 3$. The complementary residuals from multiscale data analysis, $\hat{\varepsilon}^*(t)$, can be reconstructed from (3.14) and was added to the predicted ozone-concentration profile, $x^*(t)$, to obtain the following final predicted profile of ozone concentrations, $\chi^*(t)$:

$$\chi^*(t) = x^*(t) + \hat{\varepsilon}^*(t). \tag{3.17}$$

A refining step helps improve predictive accuracy, especially during the ozone season. Figure 3.13 and Figure 3.14 compare the predicted ozone concentrations after the

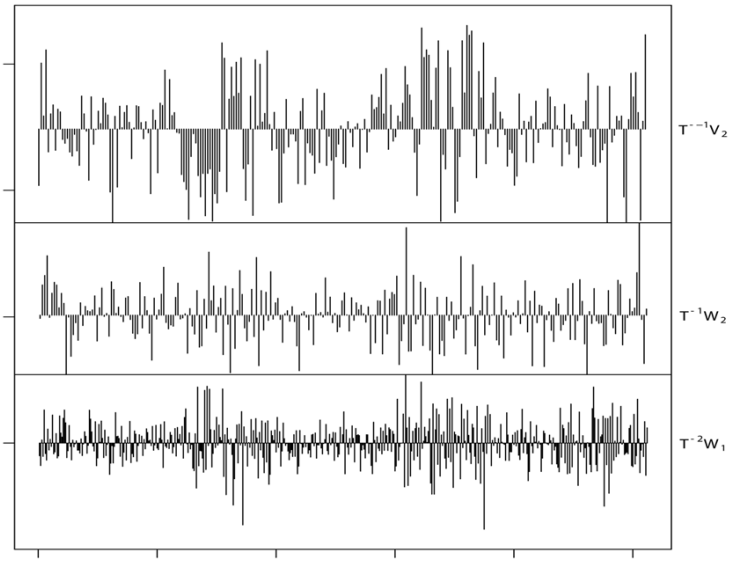


Figure 3.11. Wavelet decomposition of $\varepsilon(t, s)$ before thresholding.

refining step, predicted ozone concentrations before the refining step, and actual ozone concentrations from 14 monitoring sites across the DFW area. To clearly see the effect of the refining step, the prediction result from CAM63 is shown in Figure 3.15. The actual ozone concentration is represented by a cross mark. Predicted ozone concentration is represented by a black dot. The final predicted ozone concentration after the refining step is represented by a solid line. The predicted ozone concentrations can capture the overall characteristics of the actual ozone concentration, but they could not accurately predict the actual highly variable ozone concentrations that occur during the ozone season. Once the complementary residuals calculated in the refining step have been added to the predicted ozone concentration, our research, which is the final predicted ozone concentration, can better capture the behavior of ozone during the ozone season. This can be seen in Figure 3.15, which shows that most of the lines touch the cross marks, but the black dots cluster around the centerline.

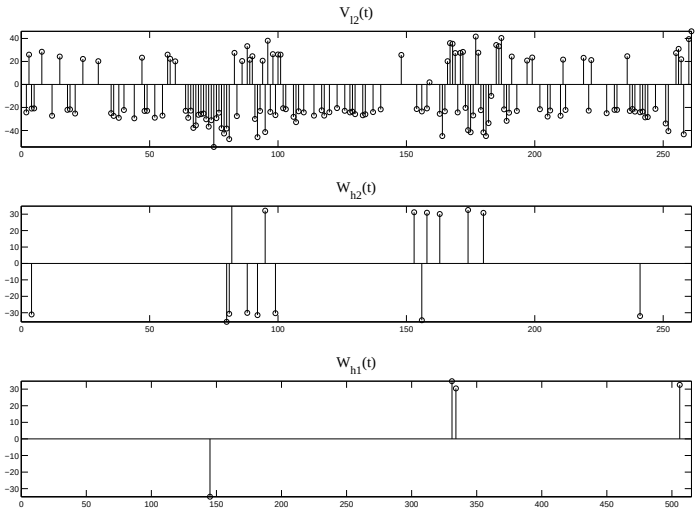


Figure 3.12. Wavelet decomposition of $\varepsilon(t, s)$ after thresholding.

3.4.7 Model Comparison

Table 3.1 shows the predictive accuracy and model complexity in various situations. The Baseline Case is the predicted ozone concentration without the refining step. The All Case represents when all site-average wavelet coefficients were used for reconstruction of complementary residuals. Other cases include the predicted ozone concentration after thresholding, using either VisuShrink or our exploratory thresholding method. Our exploratory thresholding method is based on a semi-comprehensive search of thresholds within the user-specified bounds. To be specific, we searched the thresholds between 10 and 80 in the coarsest scale, 10 and 40 in the middle scale, and 10 and 30 in the finest scale, each with a 10-increment unit. The second column shows the number of wavelet coefficients left after thresholding. The third column shows the percentage of data reduction as a result of the thresholding. The fourth and the fifth columns show the cross-validated mean absolute error (CV-MAE) and the CV-MAE for days when the

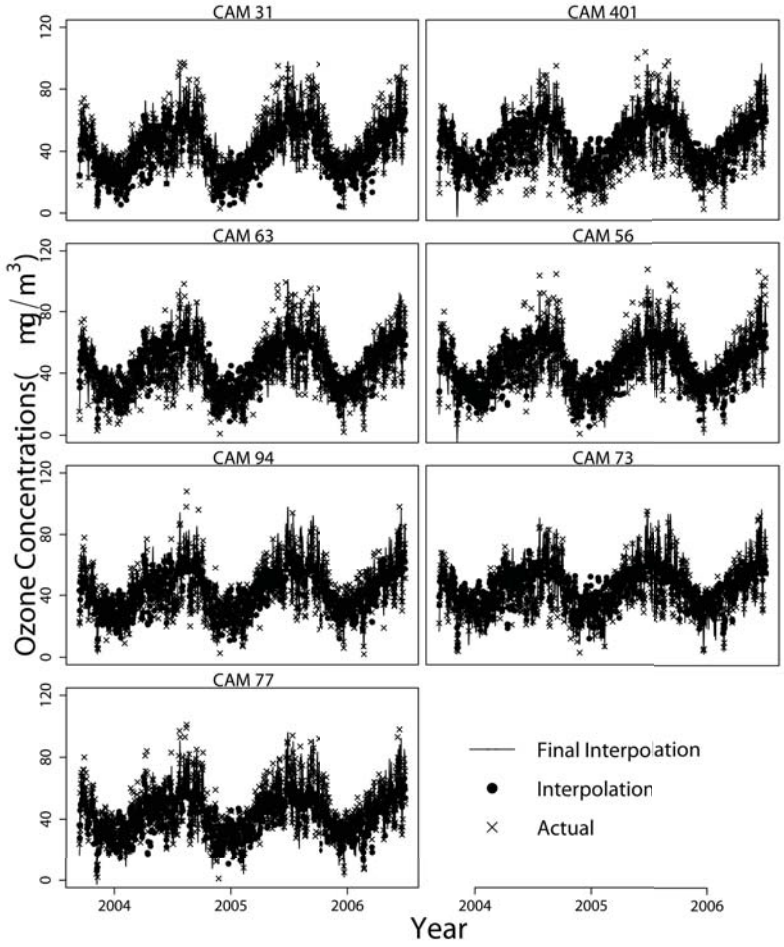


Figure 3.13. Actual vs. predicted (before refining) vs. predicted (after refining) ozone-concentration profiles (CAM31, CAM63, CAM94, CAM77, CAM401, CAM56, CAM73).

actual ozone concentration exceeded 75 ppb (CV-MAE$_{75}$), where MAE of each round of cross validation can be calculated as follows:

$$\text{MAE} = \frac{1}{m \times n} \sum_{j=1}^{n} \sum_{i=1}^{m} |x(s_i, t_j) - \chi^*(s_i, t_j)|, \tag{3.18}$$

$$\text{MAE}_{75} = \frac{1}{m} \sum_{i=1}^{m} \frac{\sum_{j=1}^{n} I(x_{ij}) |x(s_i, t_j) - \chi^*(s_i, t_j)|}{\sum_{j=1}^{n} \zeta(x_{ij})}, \tag{3.19}$$

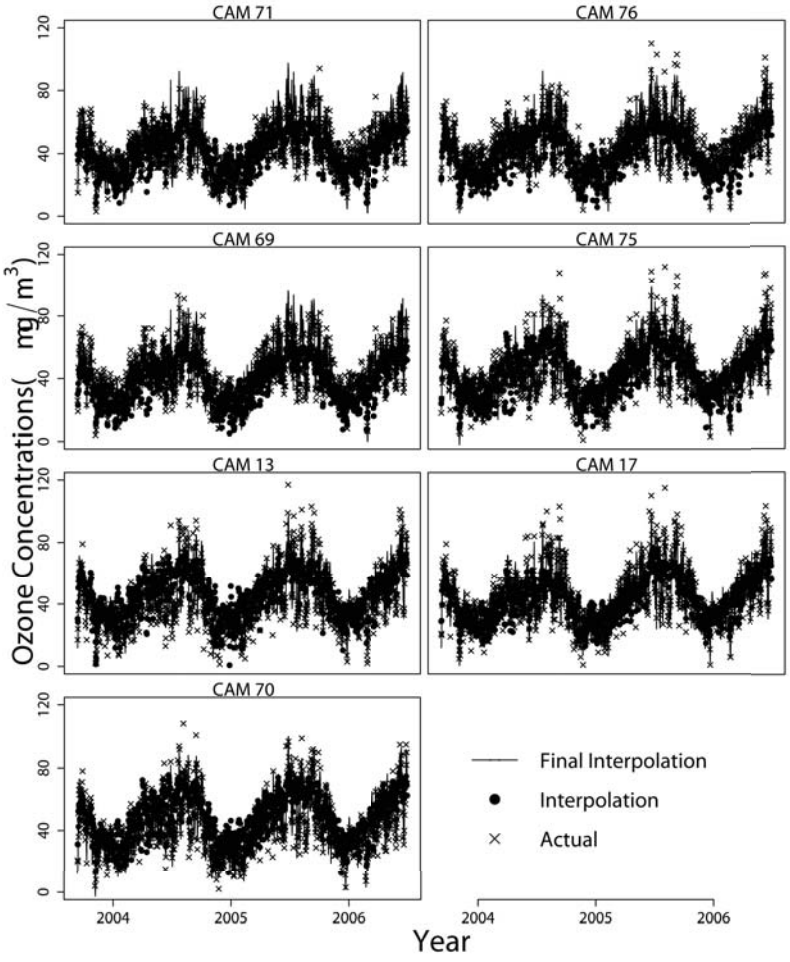


Figure 3.14. Actual vs. predicted (before refining) vs. predicted (after refining) ozone-concentration profiles (CAM71, CAM69, CAM13, CAM70, CAM76, CAM75, CAM17).

$$I(x_{ij}) = \begin{cases} 1, & x(s_i, t_j) \geq 75, \qquad i = 1, ..., m,\ j = 1, ..., n, \\ 0, & \text{otherwise.} \end{cases}$$

$x(s_i, t_j)$ represents actual ozone concentrations, and $\chi^*(s_i, t_j)$ represents the predicted ozone concentrations from monitoring site s_i at time t_j. m is the total number of monitoring sites, and n is the total time points.

Before the refining step, our procedure yields predictive accuracy of 9.17 ppb in CV-MAE and 21.60 ppb in CV-MAE$_{75}$. By adding all the wavelet coefficients from

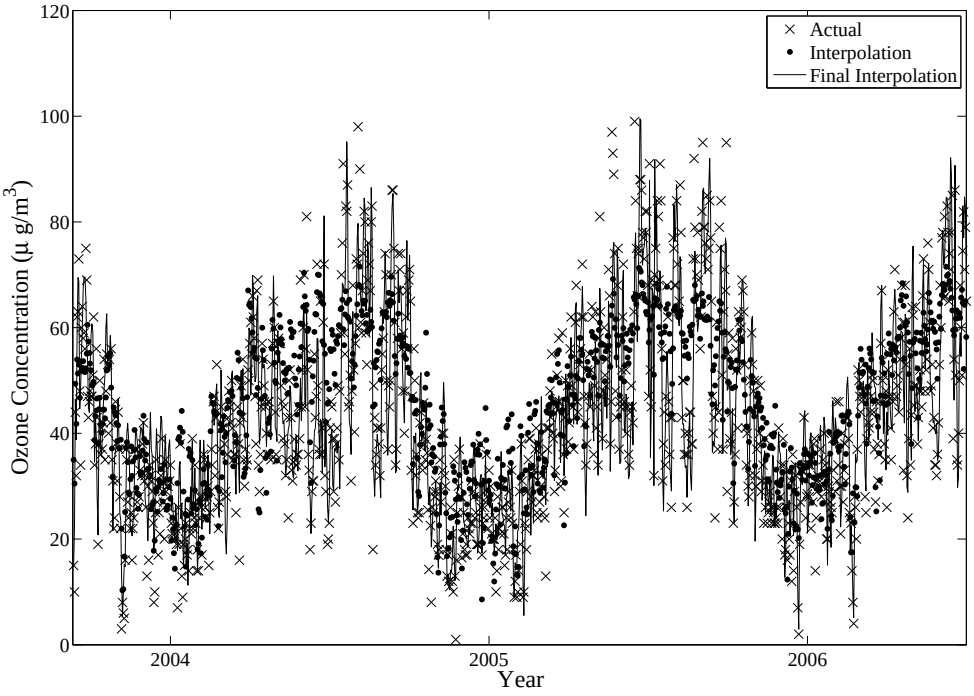


Figure 3.15. Actual vs. predicted (before refining) vs. predicted (after refining) ozone concentrations profiles of CAM63.

complementary residuals (All Case), the CV-MAE is reduced to 3.95 ppb and the CV-MAE_{75} is reduced to 8.48 ppb, which demonstrates the significance of the refining step.

Thresholding was performed to reduce the complexity of the multiscale model. The existing thresholding method, VisuShrink, and our exploratory thresholding technique were performed. VisuShrink selected 20 wavelet coefficients, leading that CV-MAE equals 9.00 ppb and CV-MAE_{75} equals 17.31 ppb. VisuShrink tends to reduce more wavelet coefficients, but achieves only a small improvement in CV-MAE compared with the Baseline Case. Consequently, VisuShrink is not recommended in our problem because higher predictive accuracy is more important than aggressive data reduction.

Cases T1 to T7 show the predictive results from our exploratory thresholding methods in which the larger the case number, the smaller the number of wavelet coefficients

Table 3.1. Comparison of prediction accuracy

Case	No. of Wavelet Coefficients	Data Reduction (%)	CV-MAE (ppb)	CV-MAE$_{75}$ (ppb)
Baseline	-	-	9.17	21.60
All	1024	0	3.95	8.48
VisuShrink	20	98.04	9.00	17.31
T1	10	99.02	9.05	18.05
T2	51	95.02	8.79	15.67
T3	63	93.85	8.60	14.57
T4	78	92.38	7.76	11.08
T5	90	91.21	7.58	10.34
T6	173	83.11	6.99	9.76
T7	643	37.21	5.68	9.04

used for reconstruction. In Case T1, only 10 wavelet coefficients were used to reconstruct the complementary residuals. Thus, CV-MAE (9.05 ppb) and CV-MAE$_{75}$ (18.05 ppb) in Case T1 are even higher than in the case of VisuShrink. The results show that higher predictive accuracy can be achieved by using a larger number of wavelet coefficients. Case T7 achieves the best predictive accuracy of any case but yields only a slight CV-MAE and CV-MAE$_{75}$ reduction compared with Cases T4, T5, and T6. Further, Case T7 used 470 more coefficients that Case T6. Overall, Cases T4, T5, and T6 represent well a trade-off between model accuracy and model complexity. It is noteworthy that these three cases significantly reduced CV-MAE$_{75}$ compared with the Baseline Case.

3.5 Conclusion

We have proposed a statistical procedure that takes advantage of available meteorological predictor variables for spatial prediction of the ozone concentration profiles in the DFW area. Because of the characteristics of ozone concentration profiles, we believe that it is necessary to disintegrate data into multiscales, then concentrate on each scale simultaneously. However, multiscale data analysis yielded a large amount of data that posed a great challenge to analytical and computational capabilities. Regression analysis

and thresholding were used to compress and reduce data points. Kriging was used for spatial prediction. Finally, a model refining step was performed to improve the accuracy of spatial prediction, especially for application in ozone season. The experimental results with real data demonstrated that the proposed procedures achieved an acceptable accuracy in spatial prediction.

To improve overall predictive accuracy even further, several modifications can be considered. First, the number of predictors can be increased to extend coverage as well to the chemical precursors of the ozone formation reaction. Nevertheless, considerable caution will still be required because the chemical reactions involved in ozone formation are extremely complicated. Second, an ordinary regression model can be substituted for other statistical models or the interaction terms can be allowed to better represent the characteristics of ozone concentration. Finally, an in-depth study of the thresholding constant is tremendously important because the thresholding constant affects both predictive accuracy and model complexity. We hope that our approach used in this study will be useful for air quality management and thus stimulate further investigation in air pollution modeling.

CHAPTER 4

AN EFFICIENT STRATEGY FOR CLASSIFICATION OF PROSTATE CANCER IN NEAR INFRARED SPECTRA

4.1 Introduction

Development of advanced sensing technology has multiplied the volume of spectral data, one of the most common types of data found in many research disciplines where advanced statistical methods are combined with highly efficient computation. Examples of the fields in which spectral data abound include NIR, mass spectroscopy (MS), and nuclear magnetic resonance (NMR) spectroscopy. Of these, NIR has advantages over other analytical tools because it is noninvasive, requires minimal sample preparation, and yields a response in real time. Various analytical studies of NIR spectra have been conducted during the past decade. Applications of NIR spectroscopic data can be found in medical and biomedical studies [74, 75, 76], pharmaceutical study [77, 78, 79], food science [80, 81, 82, 83], forestry [84], and petroleum [85]. Analysis of NIR spectra usually involves a combination of multiple samples, each of which has a large number of correlated features. A variety of data mining algorithms have been introduced to reduce the complexity that such large amounts of data present and thus help identify meaningful patterns in NIR spectra. Wu and Massart [79] developed artificial neural networks (ANN) to use NIR spectra to classify the different strengths of drugs and different qualities of solvents and polymers. The authors proposed a data pretreatment method that combined principal component analysis (PCA) and Fisher Transformation (FIT). The result of this study showed that ANN with a PCA/FIT model achieved higher predictive accuracy than ANN without data pretreatment. Dou et al. [77] constructed ANN mod-

els in term of first derivative NIR spectra, second derivative NIR spectra, and standard normal variate spectra in order to predict the components of compound paracetamol and diphenyldramine hydrochloride powdered drugs. The ANN models were then compared with partial least squares (PLS) models for three types of spectra. This study revealed that the ANN model of the first derivative NIR spectra yielded higher predictive accuracy than the others. Fernandez-Novales[80] used NIR spectroscopic data to conduct PCA and PLS analysis to identify the wavelengths important to the improvement of the sensitivity of white wine to volumic mass change and sugar reduction during the formation of its alcohol content. Uddin et al. [83] constructed a linear discriminant analysis (LDA) model based on reduced dimensions of NIR spectra, as determined by PCA, that were then used to distinguish between fresh and thawed red sea ream. The model demonstrated 100% classification accuracy. Chaychard et al. [86] compared the performance of a least-square support vector machine (LS-SVM), partial least-square regression, and multiple linear regression (MLR) in terms of their capability to use NIR spectra to accurately predict the acidity of three different grape varieties. They found that LS-SVM regression produced higher predictive accuracy than the others. Balabin et al. [87] predicted the properties of gasoline such as density and boiling points using various data mining algorithms, including MLR, PCR, PLS, Poly-PLS, spline-PLS, and ANN, and compared them in terms of predictive accuracy, computational time, and ease of use. Candolfi et al. [88] compared the performance of LDA, quadratic discriminant analysis (QDA), and k-nearest neighbor (kNN) methods in order to classify samples of clinical study lots (a tablet dataset and a capsule dataset). They concluded that it might be necessary to propose a two-step procedure for a classification model, first to discriminate between given classes, and second to apply a method that allows positive identification.

Despite such extensive research on data mining algorithms to characterize NIR spectral patterns, few efforts have been made to develop methods to address situations

of NIR spectral analysis in which the number of samples from the normal group greatly exceeds those from the abnormal group - a class imbalance problem - or when patterns between the normal and abnormal groups are not clearly distinguishable - a problem of overlapping classes. The main objective of this paper is to develop an efficient classification strategy that addresses these problems, both of which are encountered often in biomedical applications. Our prime example of these problems uses NIR spectra taken from ex vivo human prostate glands. The goal is to classify, with a high degree of accuracy, these spectra as either normal or cancerous. Our experimental data are imbalanced and overlapped between the normal and malignant groups.

4.2 Background

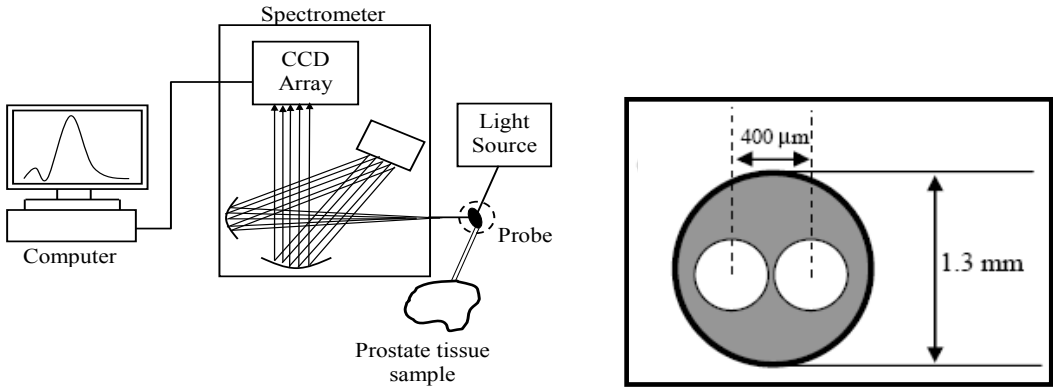
Prostate cancer is a common cancer associated with elderly men all over the world and is the second leading cause of cancer death in the United States. The prostate specific antigen (PSA) test, digital rectal examination or trans-rectal ultrasound are typically used to screen for this cancer. Because prostate cancer is asymptomatic and difficult to detect, a biopsy can be performed as a follow-up to suspicious screening results. A number of men report discomfort during and after prostate biopsy [89] because the biopsy can cause minor bleeding, infection, and difficulty in urination. Once prostate cancer is diagnosed, common treatments usually are watchful waiting [90], prostatectomy [91, 92], radiation therapy [93, 94], hormonal therapy [95], cryosurgery [96], High Intensity Focused Ultrasound [97], or a combination of all of these. Because each treatment has different side effects, no consensus exists on the best treatment for prostate cancer. Usually, the treatment options depend on the stage of the cancer, the Gleason score, the PSA level, the patient's health, and age [98]. One of the common treatment methods for prostate cancer is a radical prostatectomy, which severs nerves and blood vessels, damages tissues, and reduces the survivor's quality of life.

4.3 Data

4.3.1 Data Collection

Fig. 4.1 shows the experimental setup used for optical spectroscopic measurements. It consists of a spectrometer (USB 2000, Ocean Optics, Inc., FL) with a wavelength range from 400 nm to 1000 nm, a tungsten-halogen light source (HL-2000, Ocean Optics, Inc.), a laptop computer equipped with LabView interface software (National Instruments, Austin, TX) for collecting and displaying the optical reflectance curves in real time, and a fiber-optic probe. The optical measurements were taken at the University of Texas Southwestern Medical Center, and the data processing was done at the University of Texas at Arlington.

After approval by our Institution Review Board, we prospectively collected optical spectroscopic measurements (OSM) on consecutive prostate specimens removed through laparoscopic radical prostatectomy because cancer was confirmed. Immediately after its extraction, each specimen was stored on ice and transferred to a pathology facility located next to the operating room. Then, the prostate sample was bihalved so that the outside fibrous prostatic tissue and capsules were bypassed. Direct contact with the inside tissue through the optical probe was assured in order to initially obtain true cancer signatures. After the prostate sample was bihalved, it was often noticed that in many cases there were no abnormalities that could be detected by the naked eye. In this study, the fiber-optic probe contained two 400-μm diameter fibers for light delivery and light collection (Fig. 4.1(b)). Initial measurement was performed on 12 prostate samples, left and right peripheral regions (Fig. 4.2), with a maximum of eight measured-locations per each subject for a total of 97 spectra (Fig. 4.3(a)).



(a) (b)

Figure 4.1. (a) Schematic diagram representing the experimental set up for optical spectroscopic measurements, (b) The schematic cross section of the 400-μm fiber probe.

4.3.2 Data Description

Each NIR spectrum contains 1,312 features that represent the different wavelengths in the NIR bandwidth. Pathologists' reports and optical results were used to initially identify the spectra into 82 normal and 15 tumor spectra. This typifies the imbalance property commonly found in the biomedical data.

To ensure comparability between spectra, normalization was done by dividing each spectral point by the area of the total intensity of the spectrum. Two potential outliers were identified from each class and were removed from the subsequent analysis. This reduced the number of spectra being studied to 95 spectra(81 normal, 14 tumor). All 95 normalized spectra are displayed in Fig. 4.3(b) in which light grey lines represent the normal spectra. The display clearly shows that the normal and tumor spectra are highly overlapped.

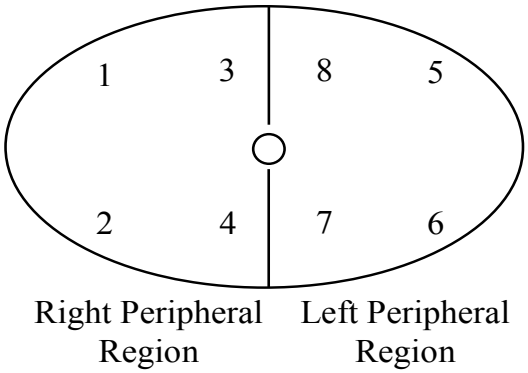


Figure 4.2. Eight sample locations on prostate gland.

4.4 Methods

4.4.1 Overview

As stated earlier, the main purpose of this study is to develop an efficient data mining algorithm that discriminates between normal spectra and cancerous spectra in the NIR data. Fig. 4.4 shows an overview of the proposed approach that consists of three main steps: (1) clustering analysis, (2) first-stage classification, and (3) second-stage classification. In the clustering analysis, k-means clustering analysis was applied to determine the pattern of the 95 NIR spectra based on their characteristics without prior knowledge of their preexisting class label (normal and tumor). Using a k-nearest neighbor (kNN) algorithm for the first-stage classification, we classified all spectra based on the new class labels obtained from the preceding clustering analysis. If the spectra belong to a pure cluster containing only normal spectra, then no further classification is required. Then we proceed to the second-stage classification for the spectra in a mixed cluster. Second-stage classification involves feature selection and classification. A classification tree algorithm was used to select important features and perform classification

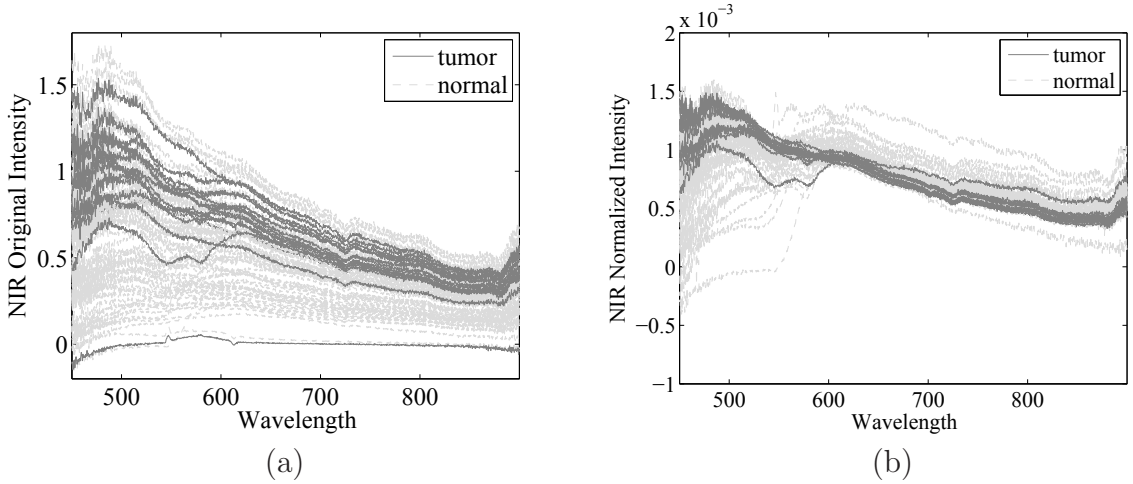


Figure 4.3. Plot of NIR spectra: (a) 97 original spectra (b) 95 normalized spectra after elimination of outliers.

analysis for the second stage. All approaches used in this study were implemented using MATLAB [99].

4.4.2 Clustering Analysis

Clustering analysis was conducted to determine patterns of the spectra based upon their characteristics while withholding preexisting class information (normal and cancer). To be precise, we applied k-means clustering analysis to 95 spectra in order to group them into new clusters instead of separating them into normal and cancerous clusters. k-means clustering analysis systematically partitions the dataset by minimizing within-group variation and maximizing between-group variation. The spectra with similar patterns are grouped together under the same cluster label, while spectra with different patterns are isolated into different clusters. A brief summary of the k-means clustering algorithm is as follows: Given k seed points, each observation is assigned to one of the k seed points close to the observation, which creates k clusters. Then, seed points are replaced with the mean of the currently assigned clusters. This procedure is repeated with updated

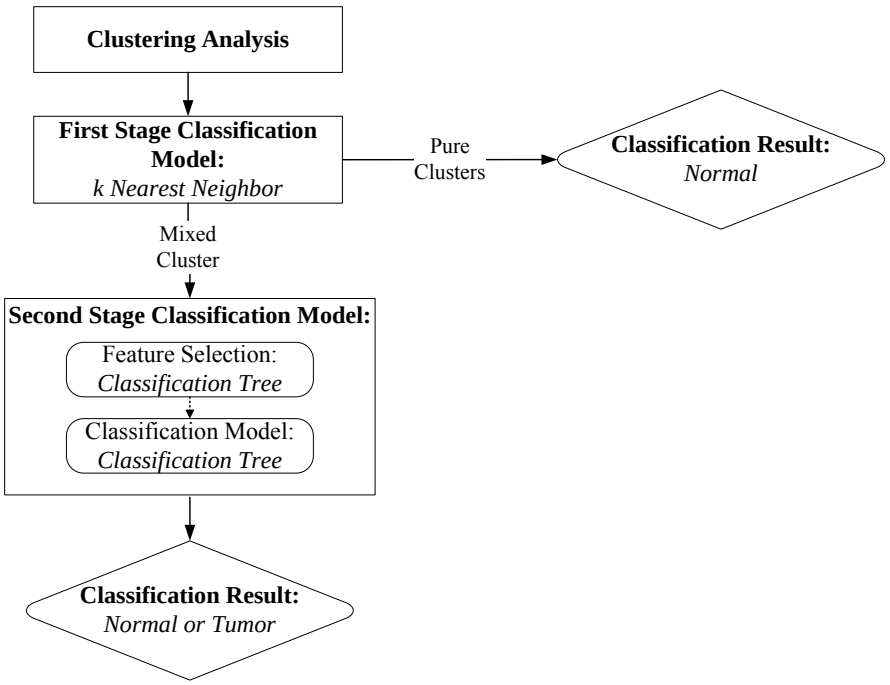


Figure 4.4. Overview of 2-stages classification algorithm.

seed points until the assignments do not change. The results of the k-means clustering algorithm depend on distance metrics and the number of clusters (k).

In this study the following correlation coefficient (D) between the two spectra is used as the distance metric:

$$D_{[x_a,\,x_b]} = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{x_a(w_j)-\bar{x}_a}{\sigma_{x_a}}\right)\left(\frac{x_b(w_j)-\bar{x}_b}{\sigma_{x_b}}\right), \tag{4.1}$$

where

$$\bar{x} = \frac{1}{m}\sum_{j=1}^{m} x(w_j), \tag{4.2}$$

and

$$\sigma_x = \left[\frac{1}{m}\sum_{j=1}^{m}[x(w_j)-\bar{x}]^2\right]^{\frac{1}{2}}. \tag{4.3}$$

$x(w_j)$ represents normalized NIR intensity corresponding to wavelength w_j from the spectra x. $\bar{x}$ and σ_x, respectively, represent the mean and standard deviation of

normalized NIR spectra. In contrast to Euclidean distance that measures the difference of each spectrum over the different wavelengths, the correlation distance allows us to measure the similarity in shape between the two NIR spectra.

Although a variety of methods are available to determine the number of clusters, no consensus exists about which one best satisfies all conditions. Here we applied the k-means clustering algorithm with k values ranging from two to 10. We then selected the final k so that the number of pure clusters reaches the first maximum. As a consequence, we used k=3.

4.4.3 Classification

4.4.3.1 k Nearest Neighbor Algorithm

A kNN algorithm was used to classify each spectra based on the class labels obtained from the preceding k-means clustering analysis. In order to construct a classification model that achieves high classification accuracy, compatibility between the clustering algorithm and the classification algorithm should be considered. Because a correlation coefficient was employed as a distance measure in the k-means clustering analysis, a classification algorithm that classified spectra based on the preceding class labels was recommended so, as to use the same distance metric. Although a variety of classification algorithms are available, because a kNN algorithm can be constructed with a variety of distance metrics, one of which is a correlation coefficient, kNN was selected in this study. kNN is a widely used classification algorithm that does not require a rigid model structure. Instead, it decides the class of an object by analyzing its k nearest neighbors within the training data.

A kNN algorithm first calculates the distance between an unknown data point and a training dataset, then ranks the training dataset based on the calculated distance

from minimum to maximum. The assigned class is calculated from a majority voting scheme of the first k^{th} training data. In other words, the class with the highest number of observations out of k^{th} training data is assigned as the predicted class of an unknown data point. Another parameter that needs to be specified in a kNN algorithm is the number of nearest neighbors (k). Because the majority voting scheme is achieved through local information, the model with a small k is more responsive but also sensitive to noise and outliers, but the model with a large k is less responsive. To determine the number of nearest neighbors, we tested different values of k ($k = 2, \cdots, 15$), then selected the k that produces the first minimum misclassification rate and found that k=6 produces the minimum misclassification rate. Thus, k was set as 6 in this study.

4.4.3.2 Classification Tree

A classification tree is one of the widely used classification methods that partitions the input (feature) space into disjointed hyper-rectangular regions according to performance measures such as misclassification errors, the Geni index, and cross-entropy and then fits a constant model in each disjointed region [100]. The number of disjointed regions (equivalent to the number of terminal nodes in a tree) should be determined appropriately because a very large tree overfits the training set, but a small tree cannot capture important information in the data. In general, there are two approaches to determining tree size. The first approach is the direct stopping method that attempts to stop tree growth before the model overfits the training set. The second approach is tree pruning that removes the leaves and branches of a full-grown tree to find the right size of the tree. To determine tree size, we stopped the growth of a tree when the number of data points in the terminal node reached 10, and the Geni index was used as a performance measure.

To obtain classification accuracy, all models were constructed based on a leave-one-out cross validation technique in which 94 spectra were employed in model training, and the remaining one spectrum was reserved for model testing for a total of 95 spectra. This process was repeated 94 more times with alternation of the testing spectrum. The final classification results from the 95 testing spectra were then averaged to obtain the cross-validated error rates of the classification models.

4.5 Result

4.5.1 Classification with Original Class Labels

We first attempted to conduct a classification analysis using the original class labels: normal and cancer. Based on our study with the correlation distance as a distance measure, kNN identified that $k = 11$ produces a minimum misclassification rate of 14.74%.

Although a classification model with the original class labels perfectly identified 81 normal spectra, it nevertheless misclassified all 14 tumor spectra as normal (Table 4.2). This result implies that the classification model with the original class labels lacks the ability to discriminate between normal and cancer spectra, and hence, inspires the need to develop models that perform better.

4.5.2 Clustering

Because the classification model that used original class labels was unable to successfully discriminate between normal spectra and tumor spectra, we performed the clustering analysis with the goal of extracting the pure normal spectra that can be used for the subsequent classification analysis. k-means clustering analysis was performed on 95 NIR spectra. Our analysis indicated that when the number of clusters (k) is more than three, k-means clustering analysis attempts to break the spectra in the higher number of mixed clusters, i.e., $p = 2, m = k - p$ for $k = 3, 4, \cdots, 10$, where p represents the number

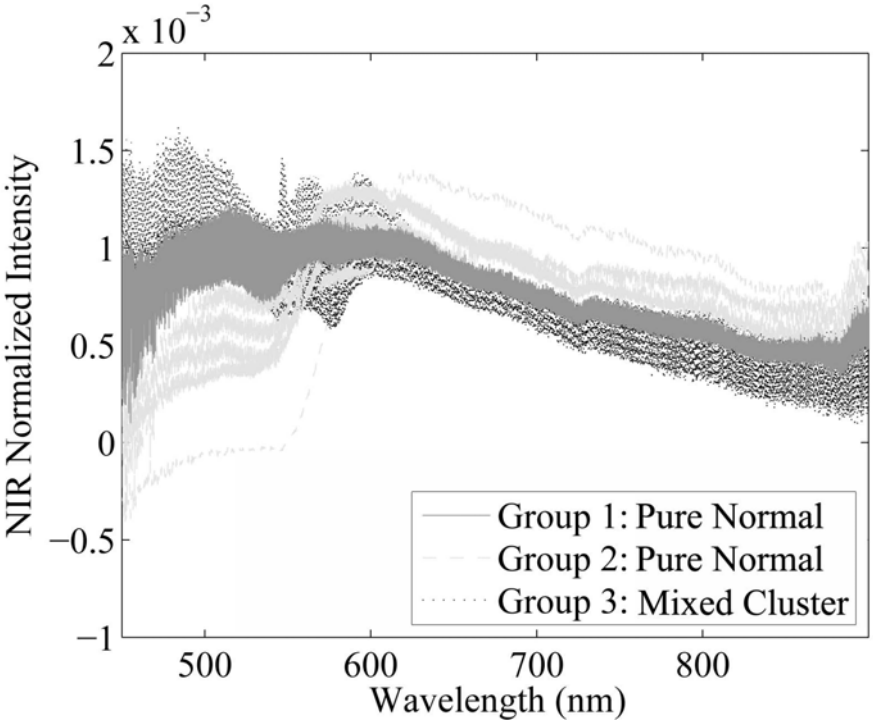


Figure 4.5. Result from 3-means clustering analysis.

of pure clusters and m represents the number of mixed clusters. In other words, k=3 achieves the first maximum of two pure clusters.

Fig. 4.5 shows the resulting 3-means clustering analysis in which 22 normal spectra were clustered into Group 1, 8 normal spectra were clustered into Group 2, and the remaining 51 normal spectra and 14 tumor spectra were clustered into Group 3. It can be seen that the pattern in Group 2 is distinctly different from Group 1 and Group 3, but some spectra from Group 1 and Group 3 are partially overlapped.

4.5.3 First Stage Classification

Using the class labels obtained from the preceding k-means clustering analysis, first-stage classification was conducted using a kNN algorithm with a correlation distance in which $k = 6$ (Fig. 4.6) reached the first minimum misclassification rate of 0.0211 for a

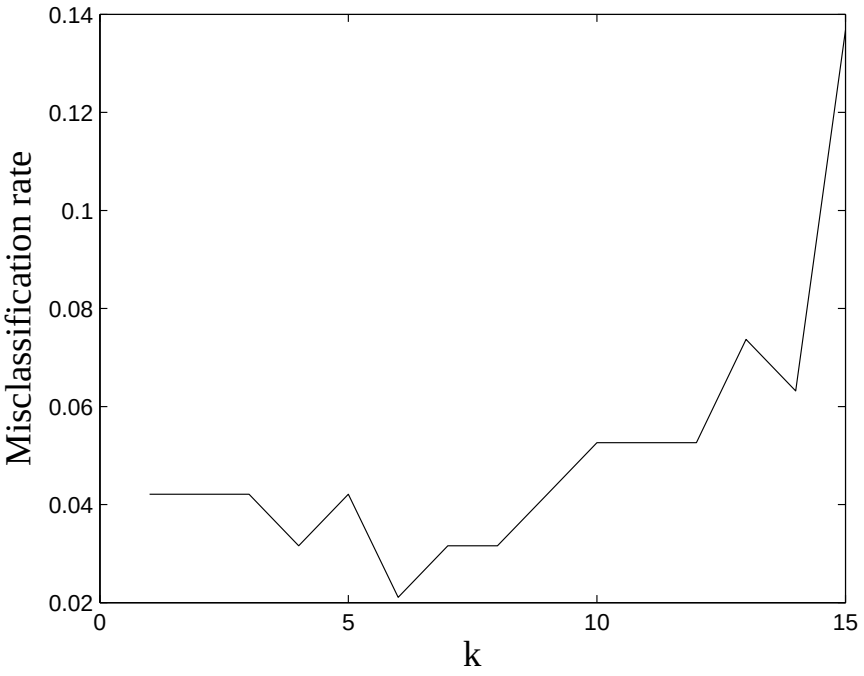


Figure 4.6. Determine the number of k by selecting k that reach the first minimum misclassification rate (k=6).

three-class classification problem. It can be observed that kNN successfully predicted most of the normal spectra in groups 1 and 2 with 93.33% accuracy. In other words, kNN correctly identified 28 normal spectra in groups 1 and 2 from a total of 30 spectra. Because Groups 1 and 2 belong to pure clusters, these 28 spectra do not require further analysis, and therefore, it can be concluded that they are normal spectra. On the other hand, the spectra that were assigned to the mixed cluster (Group 3) should proceed to second-stage classification for further analysis.

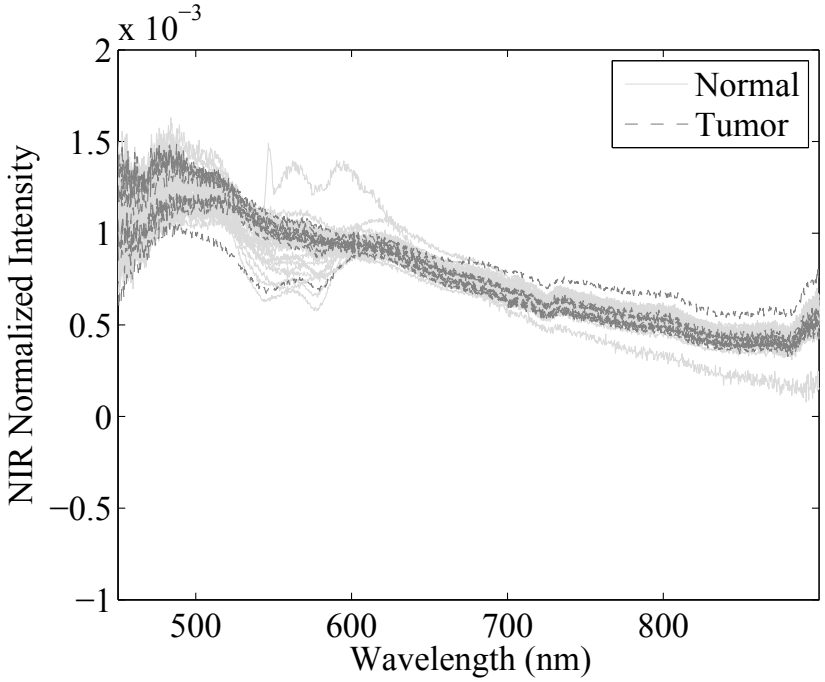


Figure 4.7. Plot of spectra in Group 3 which contains both normal spectra and tumor spectra.

4.5.4 Second Stage Classification

4.5.4.1 Second Stage Classification without Feature Selection

Figure 4.7 displays the spectra in the mixture cluster that can be hardly distinguished between the two classes. Various classification algorithms including classification trees, support vector machines, and kNN were performed to obtain the accurate classification result. Instead of comparisons between the shapes of these spectra, a classification tree, which is one of the nonlinear and nonparametric models, was selected. In general, a classification tree examines all input (feature) space in order to construct a nonlinear decision structure for a given dataset. The second-stage classification without feature selection constructed decision trees based on all 1,312 features to discriminate between spectra in the mixed cluster. Although the misclassification rate (25.26%) of this model is higher than the classification model based on the original class labels(14.74%), we

think that a classification tree has the capability to discriminate between normal spectra and tumor spectra because it correctly identified 60 spectra as normal and 11 spectra as tumor (Table 4.3). Despite this success, this model raised the question of whether to include all features in the analysis, a question that will be examined in the next section.

4.5.4.2 Second Stage Classification with Feature Selection

Because the two-stage classification model without feature selection misclassified more spectra than expected, the question must be posed of whether to include all features in the analysis. By carefully investigating the plot of spectra in the mixed cluster (Figure 4.7(a)), we assume hypothetically that some features might contribute more to the classification model than some others. To select the important features, we developed a heuristic approach, based on the classification tree structure, to assign an overall weight to each feature. Features with high overall weight imply high importance, while features with low overall weight are considered less significant. This heuristic approach was constructed based on the idea that features selected at a higher level in the tree structure received greater weight than features selected at a lower level. In addition, features that were frequently selected received greater weight than those that were infrequently selected. To determine what weights to assign to selected features, we first calculated the weight corresponding to each decision node, $wt(i, f)$, by taking the inverse of the level as follows

$$wt(i, f) = \frac{1}{l}, \tag{4.4}$$

where l represents the level that a particular node belongs to, i represents only the decision node, and f represents the selected feature. For a given feature, we obtained the total weight, $TW(f)$, by calculating the summation of weight as

$$TW(f) = \sum_{i=1}^{n} wt(i, f). \tag{4.5}$$

Because the summation of total weight for a given tree structure should be equal to one, we normalized the total weight by dividing each total weight by the summation of total weight of a particular tree as follows

$$NM(f) = \frac{TW(f)}{\sum_{f=1}^{N} TW(f)}, \tag{4.6}$$

where $NM(f)$ represents the normalized weight. If the cross validation scheme is determined, the average weight for a given feature, $\bar{w}_f$, is calculated by taking the summation of normalized weight from every fold of the cross validation, then dividing by the total number of cross validation folds (V). It should be noted that some features were not selected in every fold of the cross validation. Therefore, the unselected features were assigned a normalized weight of zero for that particular fold

$$\bar{w}_f = \frac{\sum_{v=1}^{V} NM(f, v)}{V}. \tag{4.7}$$

For a simple interpretation, we once normalized the $\bar{w}_f$ by the maximum weight of $\bar{w}_f$. Thus, the overall weight, $W(f)$, can be calculated as:

$$W(f) = \frac{\bar{w}_f}{\max(\bar{w}_f)}. \tag{4.8}$$

Our classification tree-based feature selection approach identified 40 features out of 1,312 as important. Table 4.1 lists the important features and their corresponding weights. Among these 40 features, the five features (466.085, 521.835, 603.881, 780.576, and 782.563) received considerably higher weight compared with others. Therefore, in our proposed procedure, construction of the second-stage classification model would be based on these five selected features.

The classification results from our proposed procedure (classification model with feature selection) reveal a misclassification rate of 10.53%, which is significantly lower

Table 4.1. Important Wavelength

Wavelength (nm)	Overall Weight(%)	Wavelength (nm)	Overall Weight(%)
450.247	0.52	595.066	0.71
460.2	2.80	**603.881**	**16.21**
461.672	0.91	630.875	1.99
466.085	**47.88**	644.11	0.42
469.023	0.42	645.152	0.42
469.757	0.33	649.665	0.42
470.858	1.83	667.648	0.42
486.964	0.33	670.405	0.42
503.356	1.54	674.88	0.42
516.411	0.51	720.948	1.53
520.028	0.42	773.607	0.42
521.835	**22.02**	**780.576**	**100.00**
527.251	0.50	**782.563**	**20.74**
533.737	4.71	783.888	0.76
541.648	0.50	804.991	1.87
555.265	0.84	805.32	3.06
563.48	0.49	847.029	0.22
587.641	0.45	853.165	1.53
589.764	3.13	856.388	0.85
592.593	0.90	886.492	1.50

than either the classification model constructed with the original class labels (14.74%) or the classification model without feature selection (25.26%) were selected. It can be observed from Table 4.4 that out of 81 normal spectra, our two-stage classification model with feature selection correctly identifies 76 spectra as normal and only misidentifies 5 spectra as tumors. In contrast, our model correctly identifies 9 spectra as tumors while misidentifying 5 of 14 tumor spectra as normal. Our proposed procedure yields higher sensitivity (0.94) than the classification model constructed with the original class labels but is comparable to the classification model without feature selection. How-

Table 4.2. Prediction results of original class labels

Actual	Predicted		
	Normal	Tumor	Total
Normal	81	0	81
Tumor	14	0	14
Total	95	0	95
Sensitivity	0.85		
Specificity	0.00		

Table 4.3. Prediction results of two-stage classification method without feature selection

Actual	Predicted		
	Normal	Tumor	Total
Normal	60	21	81
Tumor	3	11	14
Total	63	32	95
Sensitivity	0.95		
Specificity	0.34		

ever, the specificity (0.64) is significantly higher than with both the classification model constructed with the original class labels and the classification model without feature selection.

4.6 Conclusion

This study proposes an effective classification approach to discriminate between NIR spectra that represent normal prostate tissue and those that represent prostate cancer. In order to efficiently handle the imbalance of data and its significant overlapping, we propose first to perform clustering analysis to obtain new class labels that improve

Table 4.4. Prediction results of the propose two-stage classification method with feature selection

Actual	Predicted		
	Normal	Tumor	Total
Normal	76	5	81
Tumor	5	9	14
Total	81	14	95
Sensitivity	0.94		
Specificity	0.64		

classification accuracy. Then, using the class labels obtained from the previous clustering analysis, we undertook a second step of two-stage classification: The first of these two stages is an effort to construct a classification model, and the second stage focuses on the group of mixed classifications created in the first stage. To increase accuracy, the second classification model was built based on selected features that capture important characteristics of the spectral data. Our proposed procedure produced higher classification accuracy than the classification model with the original labels.

CHAPTER 5

FUTURE WORKS

5.1 Spatial Interpolation of the Ozone Concentration Profiles

Based on our previous work "Spatial Prediction of the Ozone Concentration Profiles", the purpose of our study was to propose a procedure for general prediction. However, without taken into account of the boundary of the location space, though our model achieved high accuracy, the result might not be hold in some specific regions. To further the development of this model, we would like to extend our research by identify the convex hull of the study region in order to achieve higher interpolation accuracy for the temporal profile. Then the model can be broaden to cover wider area of interest, e.g., the state of Texas and the continental United States, in which the spatial homogeneity and heterogeneity across the regions should be taken into consideration.

5.2 Classification Tool for Prostate Cancer Detection in Near Infrared Spectra

In order to implement the spectroscopic probe as a guidance tool in in vivo patients, the model could be improved to achieve higher classification accuracy. Because the imbalance property of the data set, our future research direction is to explore the one-class classification data mining algorithm. The advantage of one-class classification is such that it tries to identify one class of objects and distinguish it from all other objects. According to our clustering analysis, the NIR spectra can be grouped into three clusters. This dataset's characteristic combined with the discrimination ability of the one-class classification motivated our research to study the feasibility of applying the one-class

classification data mining technique as a tool to increase the classification accuracy of NIR spectra.

REFERENCES

[1] K. Wark, C. F. Warner, and W. T. Davis, *Air Pollution, Its Origin and Control*. Melon Park, CA: Addison-Wesley, 1998.

[2] M. Lippmann, "Health effects of ozone. a critical review," *Journal of Air Pollution Control Association*, vol. 39, no. 5, pp. 672–675, 1989.

[3] R. Bobbink, "Impacts of tropospheric ozone and airborne nitrogenous pollutants on nature and semi-nature ecosystems: a commentary," *New Phytologist*, vol. 139, pp. 161–168, 1998.

[4] W. L. Chameides and P. S. Kasibhatla, "Growth of continental-scale metro-agro-plexes, regional ozone pollution, and world food production," *Science*, vol. 264, no. 5155, pp. 74–77, 1994.

[5] C. A. Pope, R. Burnett, N. J. Thun, E. E. Calle, D. Krewskik, K. Ito, and G. D. Thurston, "Lung cancer, cardiopulmonary mortality, and long term exposure to fine particulate air pollution," *Journal of the American Medical Association*, vol. 287, pp. 1132–1141, 2002.

[6] L. A. McNair, R. A. Harley, and A. G. Russell, "Spatial inhomogeneity in pollutant concentrations, and their implications for air quality model evaluation," *Atmospheric Environment*, vol. 20, pp. 4291–4301, 1996.

[7] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. Boca Raton, Florida: Chapman Hall/CRC, 1997.

[8] M. L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1997.

[9] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., ser. Springer Series in Statistics. New York, NY: Springer-Verlag, 2002.

[10] S. Tilmes and J. Zimmermann, "Investigation on the spatial scales of the variability in measured near-ground ozone mixing ratios," *Geophysical Research Letters*, vol. 25, no. 20, pp. 3827–3830, 1998.

[11] S. A. Abdul-Wahab and S. M. Al-Alawi, "Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks," *Environmental Modelling and Software*, vol. 17, no. 3, pp. 219–228, 2002.

[12] A. Lengyel, K. Heberger, L. Paksy, I. O. Banhid, and R. Rajko, "Prediction of ozone concentration in ambient air using multivariate methods," *Chemosphere*, vol. 57, no. 8, pp. 889–896, 2004.

[13] J. Lehman, K. Swinton, S. Bortnick, C. Hamilton, E. Baldridge, B. Eder, and B. Cox, "Spatio-temporal characterization of tropospheric ozone across the eastern united states," *Atmospheric Environment*, vol. 38, pp. 4357–4369, 2004.

[14] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Prentice Hall, 2002.

[15] A. Piechocki-Minguy, H. Plaisance, C. Schadkowski, I. Sagnier, J. Y. Saison, J. C. Galloo, and G. R., "A case study of personal exposure to nitrogen dioxide using a new high sensitive diffusive sampler," *Science of the Total Environment*, vol. 336, pp. 55–64, 2006.

[16] J. P. Shi and R. M. Harrison, "Regression modeling of hourly nox and no_2 concentrations in urban air in london," *Atmospheric Environment*, vol. 31, no. 24, pp. 4081–4094, 1997.

[17] A. B. Chelani and S. Devotta, "Air quality forecasting using a hybrid autoregressive and nonlinear model," *Atmospheric Environment*, vol. 40, no. 10, pp. 1774–1780, 2006.

[18] E. Goswami, T. Larson, T. Lurnley, and L. J. S. Liu, "Spatial characteristics of fine particulate matter: Identifying representative monitoring location in seattle, washington," *Journal of the Air and Waste Management Association*, vol. 52, pp. 324–333, 2002.

[19] J. M. Davis and P. Speckman, "A model for predicting maximum and 8h average ozone in houston," *Atmospheric Environment*, vol. 33, pp. 2487–2500, 1999.

[20] W. Wang, W. Lu, X. Wang, and A. Y. Leung, "Prediction of maximum daily ozone level using combined neural network and statistical characteristics," *Environment International*, vol. 29, no. 5, pp. 555–562, 2003.

[21] E. Gramsch, F. Cereceda-Balic, P. Oyola, and D. Von Baer, "Examination of pollution trends in santiago de chile with cluster analysis of pm_10 and ozone data," *Atmospheric Environment*, vol. 40, no. 28, pp. 5464–5475, 2006.

[22] N. T. K. Oanh, P. Chutimon, W. Ekbodin, and W. Supat, "Meteorological pattern classification and application for forecasting air pollution episode potential in a mountain-valley area," *Atmospheric Environment*, vol. 39, no. 7, pp. 1211–1225, 2005.

[23] D. M. Holland, P. P. Principe, and J. E. Sickles, "Trends in atmospheric sulfur and nitrogen species in the eastern united states for 1989-1995," *Atmospheric Environment*, vol. 33, no. 1, pp. 37–49, 1998.

[24] Environment Protection Agency. (2007, May) Understanding the clean air act. U.S. Environment Protection Agency. Research Triangle Park, NC. [Online]. Available: http://epa.gov/air/caa/peg/understand.html

[25] ——. (2007, Dec.) PM standards. U.S. Environment Protection Agency. Research Triangle Park, NC. [Online]. Available: http://www.epa.gov/oar/particlepollution/standards.html

[26] J. Schwartz, D. W. Dockery, and L. M. Nwas, "Is daily mortality associated specifically with fine particles?" *Journal of the Air and Waste Management Association*, vol. 46, pp. 927–939, 1996.

[27] A. D. Shendriker and W. K. Steinmetz, "Integrating nepholometer measurements for air-borne fine particulate matter $pm_{2.5}$ mass concentration." *Atmospheric Environment*, vol. 37, pp. 1383–1392, 2003.

[28] M. Russell, D. T. Allen., D. R. Collins, and M. P. Fraser, "Daily, seasonal, and spatial trends in $pm_{2.5}$ mass and composition in southeast texas," *Aerosol Science and Technology*, vol. 38, no. S1, pp. 14–26, 2004.

[29] P. Paatero, P. K. Hopke, J. Hoppenstock, and S. I. Berly, "Advance factor analysis of spatial distributions of $pm_{2.5}$ in the eastern united states," *Environmental Science and Technology*, vol. 37, pp. 2460–2476, 2003.

[30] W. C. Malm, B. A. Schichtel, R. B. Ames, and K. A. Gebhart, "A 10-year spatial and temporal trend of sulfate across the united states," *Journal of Geophysical Research (Atmospheres)*, vol. 107, no. D22, pp. ACH11.1–ACH11.20, 2002.

[31] R. J. Farber, L. C. Murray, and W. A. Moran, "Exploring spatial patterns of particulate sulfur and omh from the project mohave summer intensive regional network using analyses of variance techniques and meteorological parameters as sort determinants," *Journal of Air and Waste Management Association*, vol. 50, pp. 724–732, 2000.

[32] K. A. Gebhart and W. C. Malm, "Spatial and temporal patterns in particle data measured during the mohave study," *Journal of Air and Waste Management Association*, vol. 47, pp. 119–135, 1997.

[33] W. C. Malm, "Characteristics and origins of haze in the continental united states," *Earth Science Reviews*, vol. 33, pp. 1–36, 1992.

[34] C.-C. Chan and J.-S. Hwang, "Site representativeness of urban air monitoring stations," *Journal of the Air and Waste Management Association*, vol. 46, no. 8, pp. 755–760, 1996.

[35] M. Jun and M. L. Stein, "Statistical comparison of observed and cmaq modeled daily sulfate levels," *Atmospheric Environment*, vol. 38, pp. 4427–4436, 2004.

[36] S.-K. Park, C. E. Cobb, K. Wade, J. Mulholland, Y. Hu, and A. Russell, "Uncertainty in air quality model evaluation for particulate matter due to spatial variation in pollutant concentrations," *Atmospheric Environment*, vol. 40, pp. 563–573, 2006.

[37] S. B. Phillips and P. L. Finkelstein, "Comparison of spatial patterns of pollutant distribution with CMAQ predictions," *Atmospheric Environment*, vol. 40, pp. 4999–5009, 2006.

[38] J. L. Swall and J. M. Davis, "A bayesian statistical approach for the evaluation of CMAQ," *Atmospheric Environment*, vol. 40, pp. 4883–4893, 2006.

[39] A. Riccio, G. Barone, E. Chianese, and G. Giunta, "A hierarchical bayesian approach to the spatio-temporal modeling of air quality data," *Atmospheric Environment*, vol. 40, pp. 554–556, 2006.

[40] J. Fox, *Nonparametric simple regression: smoothing scatterplots*. Thousand Oaks, CA: SAGE, 2000.

[41] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979.

[42] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, 1988.

[43] B. K. Eder, J. M. David, and P. Bloomfield, "A characterization of the spatiotemporal variability of non-urban ozone concentration over the eastern united state," *Atmospheric Environment*, vol. 27A, no. 16, pp. 2645–2668, 1996.

[44] R. E. Baumgardner, S. S. Isil, J. J. Bowser, and K. M. Fitzgerald, "Measurement of rural sulfur dioxide and particle sulfate: Analysis of CASTNet data, 1987 through 1996," *Journal of Air and Waste Management Association*, vol. 49, pp. 1266–1279, 1999.

[45] P. McMurry, M. Shepherd, and J. Vickery, *Particulate Matter Science for Policy Makers: A NARSTO Assessment*. Cambridge, UK: Cambridge University Press, 2004.

[46] J. Seinfeld and S. Pandis, *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. New York, NY: A Wiley-Interscience Publication, 2000.

[47] Environment Protection Agency. (2002) National emissions inventory data & documentation. U.S. Environment Protection Agency. Washington, DC. [Online]. Available: http://www.epa.gov/ttn/chief/net/2002inventory.html

[48] W. C. Malm, B. A. Schichtel, M. L. Pitchford, L. L. Ashbaugh, and R. A. Eldred, "Spatial and monthly trends in speciated fine particle concentration in the united states," *Journal of Geophysical Research*, vol. 109, 2004.

[49] W. C. Malm, J. F. Sisler, D. Huffman, R. A. Eldred, and T. A. Cahill, "Spatial and seasonal trends in particle concentration and optical extinction in the united states," *Journal of Geophysical Research*, vol. 99, pp. 1347–1370, 1994.

[50] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 3rd ed. Englewood Clifs, NJ: Prentice Hall, 1994.

[51] Environment Protection Agency. (1997, July) Health and environmental effects of ground-levels ozone. U.S. Environment Protection Agency. Research Triangle Park, NC. [Online]. Available: http://www.epa.gov/ttn/oarpg/naaqsfin/o3health.html

[52] ——. (2008, May) Ozone air quality standards. U.S. Environment Protection Agency. Research Triangle Park, NC. [Online]. Available: http://www.epa.gov/air/ozonepollution/standards.html

[53] W. G. Cobourn, D. Leslie, F. Mark, and H. Milton, "A comparison of nonlinear regression and neural network models for ground-level ozone forecasting," *Journal of Air & Waste Management Association*, vol. 50, pp. 1999–2009, 2000.

[54] J. Yi and V. R. Prybutok, "A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area," *Environmental Pollution*, vol. 92, pp. 349–357, 1996.

[55] M. W. Gardner and S. R. Dorling, "Statistical surface ozone models: an improved methodology to account for non-linear behaviour," *Atmospheric Environment*, vol. 34, pp. 21–34, 2000.

[56] S. I. V. Sousa, F. G. Martins, M. C. M. Alvim-Ferraz, and M. C. Pereira, "Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations," *Environmental Modelling & Software*, vol. 22, pp. 97–103, 2007.

[57] M. W. Gardner and S. R. Dorling, "Meteorologically adjusted trends in uk daily maximum surface ozone concentrations," *Atmospheric Environment*, vol. 34, pp. 171–176, 2000.

[58] G. Huerta, B. Sans, and J. R. Stroud, "A spatiotemporal model for mexico city ozone levels," *Journal of the Royal Statistical Society: Series C*, vol. 53, no. 2, pp. 231–248, 2004.

[59] M. L. Thompson, J. Reynolds, L. H. Cox, P. Guttorp, and P. D. Sampson, "A review of statistical mehods for the meteorological adjustment of tropospheric ozone," *Atmospheric Environment*, vol. 35, pp. 617–630, 2001.

[60] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 41, no. 7, pp. 909 – 996, 1988.

[61] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[62] R. A. Sych, G. K. Matafonov, A. J. Belinskaya, and N. J. Ferreira, "The periodic spatial-temporal characteristics variations of the total ozone content," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 67, no. 17-18, pp. 1779–1785, 2005.

[63] J. Salmond, "Wavelet analysis of intermittent turbulence in a very stable nocturnal boundary layer: implications for the vertical mixing of ozone," *Boundary-Layer Meteorology*, vol. 114, pp. 463–488, 2005.

[64] R. Warner, "The latitudinal ozone variability study using wavelet analysis," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 70, no. 2-4, pp. 261–267, 2008.

[65] E. S. Garcia-Trevino, V. Alarcon-Aquino, and M. A. Herrrera-Garcia, "Wavelet-networks for prediction of ozone levels in puebla city mexico," in *Proc. on Electronics, Communications and Computer'07*, 2007.

[66] A. Heidarinasab, B. Dabir, and M. Sahimi, "Multiresolution wavelet-based simulation of transport and photochemical reactions in the atmosphere," *Atmospheric Environment*, vol. 38, no. 37, pp. 6381–6397, 2004.

[67] C. Hogrefe, S. Vempaty, S. T. Rao, and P. S. Porter, "A comparison of four techniques for seperating different time scales in atmospheric variables," *Atmospheric Environment*, vol. 37, pp. 313–325, 2003.

[68] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[69] N. Saito, "Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion," in *Wavelets in Geophysics*, E. Foufoula-Georgiou and P. Kumar, Eds. New York, NY: Academic Press, 1994, p. 299324.

[70] J. J. Faraway, "Regression analysis for a functional response," *Technometrics*, vol. 39, no. 3, pp. 254–261, 1997.

[71] H. Bayraktar and F. S. Turalioglu, "A kriging-based approach for locating a sampling site—in the assessment of air quality," *Stochastic Environmental Research and Risk Assessment*, vol. 19, no. 4, pp. 301–305, 2005.

[72] X. Emery, "Ordinary multigaussian kriging for mapping conditional probabilities of soil properties," *Geoderma*, vol. 132, no. 1-2, pp. 75–88, 2006.

[73] O. Schabenberger and C. A. Gotway, *Statistical Methods for Spatial Data Analysis*, ser. Text in Statistics Science, B. P. Carlin, C. Chatfield, M. Tanner, and J. Zidek, Eds. Boca Raton, Florida: Chapman Hall/CRC, 2005.

[74] M. John and C. G. amd H. Liu, "Determination of hemoglobin saturation in blood-perfused tissues using reflectance spectroscopy with small source-detector seperations," *Applied Spectroscopy*, vol. 55, pp. 1686–1694, 2001.

[75] D. L. Peswani, "Detection of positive cancer margins intra-operatively during nephectomy and prostatectomy using optical reflectance spectroscopy," Master's thesis, Biomedical Engineering, The University of Texas at Arlington, 2007.

[76] M. U. Utzinger, E. Silva, D. Gershenson, R. C. B. Jr., M. Follen, and R. Richards-Kortum, "Reflectance spectroscopy for in vivo characterization of ovarian tissue," *Lazers in Surgery and Medicine*, vol. 28, pp. 56–66, 2001.

[77] Y. Dou, Y. Sun, Y. Ren, and Y. Ren, "Artificial neural network for simutaneous determinatio of two components of compound paracetamol and diphenhydramine hydrochloride powder on nir spectroscopy," *Analytica Chimica Acta*, vol. 528, pp. 55–61, 2005.

[78] L. Zhao, Y. Gao, Y. Dou, B. Wang, H. Mi, and Y. Ren, "Application of artificial neural networks to the nondestructive determination of ciprofloxacin hydrochloride

in powder by short-wavelength nir spectroscopy," *Journal of Analytical Chemistry*, vol. 62, no. 12, pp. 1156–1162, 2007.

[79] W. Wu and D. Massart, "Artificial neural networks in classification of nir spectral data: Selection of the input," *Chemometrics and Intelligent Laboratory Systems*, vol. 35, pp. 127–135, 1996.

[80] J. Fernndez-Novales, M.-I. Lpez, M.-T. Snchez, J.-A. Garca, and J. Morales, "A feasibility study on the use of a miniature fiber optic nir spectrometer for the prediction of volumic mass and reducing sugars in white wine fermentations," *Journal of Food Engineering*, vol. 89, p. 325329, 2008.

[81] M. Huang, Y. Bao, and Y. He, "Discrimination of rapeseed and weeds under actual field conditions based on principal component analysis and artificial neural network by vis/nir spectroscopy," vol. 6788, 67882S. Proc. of SPIE, 2007.

[82] F. Liu and Y. He, "Classification of brands of instant noodles using vis/nir spectroscopy and chemometrics," *Food Research International*, vol. 41, p. 562567, 2008.

[83] M. UDDIN, E. OKAZAKI, S. TURZA, Y. YUMIKO, M. TANAKA, and U. FUKUDA, "Non-destructive visible/nir spectroscopy for differentiation of fresh and frozen-thawed fish," *JOURNAL OF FOOD SCIENCEVol. 70, Nr. 8, 2005*, vol. 70, no. 8, pp. C506–C510, 2005.

[84] A. Alves, A. Santos, D. da Silva Perez, J. Rodrigues, H. Pereira, R. Simoes, and M. Schwanninger, "Nir plsr model selection for kappa number prediction of maritime pine kraft pulps," *Wood Sci Technol (2007) 41:491499*, vol. 41, p. 491499, 2007.

[85] R. M. Balabin, R. Z. Safieva, and E. I. Lomakina, "Wavelet neural network (wnn) approach for calibration model building based on gasoline near infrared (nir) spectra," *Chemometrics and Intelligent Laboratory Systems*, vol. 93, p. 5862, 2008.

[86] F. Chauchard, R. Cogdill, S. Roussel, J. Roger, and V. Bellon-Maurel, "Application of ls-svm to non-linear phenomena in nir spectroscopy: development of a robust and portable sensor for acidity prediction in grapes," *Chemometrics and Intelligent Laboratory Systems*, vol. 71, p. 141 150, 2004.

[87] R. M. Balabin, R. Z. Safieva, and E. I. Lomakina, "Comparison of linear and nonlinear calibration models based on near infrared (nir) spectroscopy data for gasoline properties prediction," *Chemometrics and Intelligent Laboratory Systems 88 (2007)*, vol. 88, p. 183188, 2007.

[88] A. Candolfi, W. Wu, D. Massart, and S. Heuerding, "Comparison of classification approaches applied to nir-spectra of clinical study lots," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 16, p. 13291347, 1998.

[89] M. Essink-Bot, H. de Koning, H. Nijs, W. Kirkels, P. van der Maas, and F. Schroder, "Short-term effects of population-based screening for prostate cancer on health-related quality of life," *Journal of The National Cancer Institute*, vol. 90, pp. 925–931, 1998.

[90] H. Wu, L. Sun, J. W. Moul, H. Wu, D. G. McLeod, C. Amling, R. Lance, L. E. O. Kusuda, T. Donahue, J. Foley, A. Chung, W. Sexton, and D. Soderdahl, "Watchful waiting and factors predictive of secondary treatment of localized prostate cancer," *The Journal of Urology*, vol. 171, no. 3, pp. 1111–1116, 2004.

[91] A. Bill-Axelson, L. Holmberg, M. Ruutu, M. Haggman, S.-O. Andersson, S. Bratell, A. Spangberg, C. Busch, S. Nordling, H. Garmo, J. Palmgren, H.-O. Adami, B. J. Norlen, and J.-E. Johansson, "Radical prostatectomy versus watchful waiting in early prostate cancer," *The New England Journal of Medicine*, vol. 352, no. 19, pp. 1977–1984, 2005.

[92] J. A. Smith Jr, R. C. Chan, S. S. Chang, S. D. Herrell, P. E. Clark, R. Baumgartner, and M. S. Cookson, "A comparison of the incidence and location of positive surgical

margins in robotic assisted laparoscopic radical prostatectomy and open retropubic radical prostatectomy," *The Journal of Urology*, vol. 178, no. 6, pp. 2385–2390, 2007.

[93] A. V. D'Amico, J. Manola, M. Loffredo, A. A. Renshaw, A. DellaCroce, and P. W. Kantoff, "6-month androgen suppression plus radiation therapy vs radiation therapy alone for patients with clinically localized prostate cancer," *Journal of American Medical Association*, vol. 292, no. 7, pp. 821–827., 2004.

[94] C. A. Perez, G. E. Hanks, S. A. Leibel, A. L. Zietman, Z. Fuks, and W. R. Lee, "Localized carcinoma of the prostate (stages t1b, t1c, t2, and t3). review of management with external beam radiation therapy," *Cancer*, vol. 72, no. 11, pp. 3156 – 3173, 2006.

[95] D. A. Loblaw, D. S. Mendelson, J. A. Talcott, K. S. Virgo, M. R. Somerfield, E. Ben-Josef, R. Middleton, H. Porterfield, S. A. Sharp, T. J. Smith, M. E. Taplin, N. J. Vogelzang, J. L. W. Jr, C. L. Bennett, and H. I. Scher, "American society of clinical oncology recommendations for the initial hormonal management of androgen-sensitive metastatic, recurrent, or progressive prostate cancer," *Journal of Clinical Oncology*, vol. 22, no. 14, pp. 2927–2941, 2004.

[96] B. Rubinsky, G. Onik, J. J. Finkelstein, D. Neu, and S. Jones, "Cryosurgical system for destroying tumors by freezing," US Patent 5334181, 1994. [Online]. Available: http://www.patentstorm.us/patents/5334181.html

[97] T. A. Gardner and M. A. Koch, "Prostate cancer therapy with high-intensity focused ultrasound-comprehensive review," *Clinical Genitourinary Cancer*, vol. 4, no. 3, 2005.

[98] (2008) American Cancer Society. [Online]. Available: http://www.cancer.org

[99] (1994-2008) The MathWorks, Inc. [Online]. Available: http://www.mathworks.com/

[100] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and Regression Trees*. 1984Boca Raton, FL: Chapman & Hall/CRC, 1984.

BIOGRAPHICAL STATEMENT

Chivalai Temiyasathit was born in Bangkok, Thailand, in 1980. She received her Bachelor degree in Environmental Engineering from Chulalongkorn University, Bangkok in 2001, followed by her Master degree in Industrial and Manufacturing System Engineering from The University of Texas at Arlington (UTA) in 2003. Before joining UTA, she had an internship at the Petroleum Authority of Thailand. During her doctorate years, she worked as a graduate research assistant at the Center on Stochastic Modeling, Optimization and Statistics (COSMOS). Her research interests are in the area of statistical data mining, functional data analysis, pattern recognition, biomedical signal processing, and bioinformatics. She is a member of the Institute For Operations Research and the Management Sciences (INFORMS) and Institute of Industrial Engineers (IIE).