NONPARAMETRIC MODELING OF THE EFFECTS OF AIR POLLUTION ON PUBLIC HEALTH

PENG QIAO

NATIONAL UNIVERSITY OF SINGAPORE

2005

NONPARAMETRIC MODELING OF THE EFFECTS OF AIR POLLUTION ON PUBLIC HEALTH

PENG QIAO

(B.Sc. Peking University, China)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2005

ACKNOWLEDGEMENTS

For the completion of this thesis, I would like to express my heartfelt gratitude to my supervisor, Assistant Professor Xia Yingcun, for all his invaluable advice and guidance, endless patience, kindness and encouragement during the mentor period in the Department of Statistics and Applied Probability of National University of Singapore. I have learned many things from him, especially regarding academic research and character building. I truly appreciate all the time and effort he has spent on helping me to solve my problems even when he was in the midst of his work.

I also wish to express my sincere gratitude and appreciation to my other lecturers, namely Professors Bai Zhidong, Chen Zehua, and Loh Wei Liem, etc, for imparting

knowledge and techniques to me and their precious guidance and help in my study.

I would like to take this opportunity to record my thanks to my dear parents who have always been supporting me with their encouragement and understanding. And special thanks to all of my friends, who have contributed to my thesis in one way or another, for their concern and inspiration in my study and life during the past two years. It is a great experience to share those colorful days with them.

Finally, I would like to attribute the completion of this thesis to other members and staffs in our department for their help in various ways and providing such a pleasant studying and working environment.

Peng Qiao

August 2005

Contents

SUMMARY

This thesis aims to analyze the effects of exposure to air pollution on public health across 15 populous cities in the United States, based on daily observations from January 1987 to December 1998. In our analysis, the first step is to perform the Efficient Dimension Reduction (EDR) procedure to reduce the complexity resulting from high dimensionality involved in the air pollution problem. After obtaining the dimension and the directions of the EDR space for each study city, we then compare the cross-validatory (CV) values, which assess models in view of their forecasting performance, of a Generalized Additive Model (GAM) with those values of a general nonparametric regression model. The criterion is to choose the model with smaller CV-values. Finally, we need

to answer one important question: whether the commonly used GAM is acceptable to quantify the effects of air pollution on public health?

Our results show that air pollutants (PM_{10}, O_3, SO_2, NO_2 and CO) at current levels, acting with weather conditions (measured by temperature and humidity) together, have adverse effects on human health. The more influential hazards to death are O_3, PM_{10}, and weather variates. As for model selection, our results suggest that EDR via the rMAVE method proposed by Xia *et al*. (2002) is necessary to the original pollution data set, and that the general nonparametric regression model incorporating EDR outperforms GAMs. That is, GAMs are not desirable when considering the predictive ability, and hence they can be improved to better fit the air pollution data.

These results represent a starting point for refinement in the future analysis of the effects of air pollution on public health. It would seem appropriate then to investigate how to adjust the EDR space for proper usage of GAMs to gain a better forecasting performance and a deeper understanding of the link between air pollution and mortality rate for future work.

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Backgrounds on Air Pollution

Based on a series of infamous air pollution "disasters" (Meuse Vally, Belgium, 1930; Donora, Pennsylvania, United States, 1948; London, United Kingdom, 1952) (Lipfert, 1994), the link between air pollution at extremely high concentrations and acute increases in death was established by the 1980s. Those findings prompted serious consideration of ambient air quality standards and health guidelines around the world, such as the National Ambient Air Quality Standards (NAAQS) of America and the Air Quality Guidelines (AQG) of World Health Organization (WHO), to protect the public from air pollution. As a result, ambient air quality has been improved considerably in recent few

decades.

However, numerous studies published recently have reported that exposure to ambient air pollution, even at the levels commonly achieved nowadays in many cities in developed countries, is associated with various negative health outcomes, both acute and chronic, ranging from irritant effects to death (Dominici *et al*., 2000; Samet *et al*., 2000; WHO working group, 2003). Some studies have also indicated the most common and damaging air pollutants through epidemiological, toxicological and clinical approaches. Examples of potentially harmful air pollutants are respirable particulate matter (PM), ozone (O_3), sulphur dioxide (SO_2), nitrogen dioxide (NO_2) and carbon monoxide (CO). These pollutants have been recognized as respiratory irritants and can exacerbate illnesses in individuals with chronic cardiovascular and respiratory diseases (Lipfert, 1994; Pope III *et al*., 2002; WHO working group, 2003; Xia and Tong, 2005). Their effects could be more severe under certain temperature and humidity conditions (McGeehin and Mirabelli, 2001). In the following subsections we present a brief introduction to these common pollutants. (All the information refer to the following webpages:

1) Air Pollutants and Your Health (http://www.sbcapcd.org/sbc/pollut.htm);

2) Air Pollutants and Health Effects (http://www.stormfax.com/airwatch.htm); and

3) The Chemistry of Atmospheric Pollutants

(http://www.aeat.co.uk/netcen/airqual/kinetics).)

1.1.1 Particulate Matter (PM)

The term "particulate matter" refers to a complex mixture of organic and inorganic particles suspended in the air. They vary widely in physical and chemical composition, source and particle size. The primary sources of particulate matter are coal combustion processes and road traffic emissions. Ambient PM_{10} particles, which are less than 10 μm in diameter, are of currently major concern, since they can not only pass into the upper airways (nose and mouth) but also penetrate into the deepest and most sensitive areas of the lungs, and hence they are considered to be more hazardous than coarse particles. PM_{10} has been linked to numerous adverse health effects, including increased hospital admissions, exacerbation of chronic cardiovascular and respiratory diseases, and decreased lung function.

1.1.2 Ozone (O_3)

Ozone is formed as a secondary pollutant when nitrogen dioxide and volatile organic compounds chemically react in the presence of sunlight. O_3 displays strong seasonal and diurnal patterns. Some epidemiological studies have indicated that exposure to ground-level ozone air pollution, even at very low levels, can cause a number of adverse respiratory effects particularly over time. When people breathe in air polluted with ozone,

the lining of their lungs can become irritated and inflamed, causing coughs, chest discomfort and breathing difficulty. People with asthma and other respiratory diseases are particularly susceptible. Long-term exposure to ozone may lead to accelerated aging of the lungs, decreased lung function and capacity, bronchitis and emphysema. Additionally, it is reported that effects of ozone can be enhanced by particulate matter and vice versa.

1.1.3 Sulphur Dioxide (SO_2)

Sulphur dioxide is released into the air mainly from power plants, large industrial facilities, diesel vehicles and oil-burning home heaters. Sulphur dioxide is a poisonous gas that aggravates existing lung diseases especially bronchitis, constricts breathing passages in asthmatic people and causes shortness of breath. Long-term exposure to sulphur dioxide will lead to higher occurrence rates of respiratory illness. Sulphur dioxide also reacts with oxygen and rainwater to form sulphuric acid which is the major contributor to acidity in acid rain.

1.1.4 Nitrogen Dioxide (NO_2)

Dominant sources of nitrogen oxides are motor vehicles and power plants. Nitrogen dioxide is a respiratory irritant, which may exacerbate asthma and possibly increase

susceptibility to infections, especially in young children and people with existing respiratory illnesses. It disrupts and may even damage the cell membrane; it can cause acid induced irritation leading to or contributing to diminished pulmonary function and right heart stress under long-term exposure. Furthermore, nitrogen oxides is a precursor for a number of harmful secondary pollutants, so health risks of NO_2 may come from itself and its reaction products including ozone and secondary particles.

1.1.5 Carbon Monoxide (CO)

Carbon monoxide is a toxic gas which is emitted into the atmosphere as the result of combustion processes and also formed by the oxidation of hydrocarbons and other organic compounds. It is produced primarily from motor vehicles in urban cities. Carbon monoxide weakens heart contractions and lowers the amount of oxygen carried by the blood. It possibly causes nausea, dizziness and headaches and is fatal at very high concentration.

1.2 Quantification of Health Effects

As evidence of negative impacts of air pollution on public health has been accumulated, quantification of these impacts has increasingly become a critical concern. This

concern has led to several long-term research programs organized by government agencies to continuously monitor pollutant levels and regularly collect data on health outcomes in different areas, with the aim of analyzing public health-related effects. In fact, based on those systematic observations, many studies have been proposed to estimate the numbers of death attributable to air pollution (Schwartz *et al*., 1996; WHO working group, 2000), although these methods and estimates are rather different. In general, impact assessment studies follow at least three different strategies: the estimation of the exposure-response function for mortality is based on either 1) cohort studies, 2) time-series studies, or 3) an average estimate of time-series and cohort study results (Künzli *et al*., 2001). Cohort studies explore the association between measures of long-term cumulative exposure and time to death (Pope III *et al*., 2002; WHO working group, 2002). Some researchers argue that long-term exposure may be more important in view of overall public health. However, most of recent research have focused on effects of short-term exposures (several days up to a few weeks) which are the main content of time-series studies, as there are more observations available. Time-series studies explore the association between death probability and levels of air pollution shortly before the death, using mortality counts as the outcome measure. Our study is a time-series analysis.

One feature of time-series studies on heath effects of air pollution is that the probability of death is influenced not by a single hazard, but rather by a function of a whole

set of risk factors including weather conditions. Therefore, various complex statistical methods have been used to detect health-related impacts (Schwartz *et al*., 1996; Daniels *et al*., 2004). Among those methods, one commonly used approach involves a semi-parametric Poisson regression with daily mortality counts as the outcome, linear terms measuring the percentage increase in mortality associated with elevations in pollutant levels, and smooth functions of time, weather and other variables adjusting for the time-varying confounders,

$$\log \mathbb{E}(\text{daily death counts}_t) = \beta_1 \, \text{PM}_{10,\,t} + \beta_2 \, \text{O}_{3,\,t} + \text{confounders}.$$

See Schwartz *et al*. (1996). Other techniques under consideration to assess the adverse effects of air pollution include models with splines, thresholds or distributed lags.

During the last few years, Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1986) have become the most widely applied method, because it allows for highly flexible nonparametric fitting of seasonal and long-term time trends in air pollution as well as nonlinear associations with weather variables (Dominici *et al*., 2000, 2002, 2004; Lee *et al*., 2000; Xia and Tong, 2005). Furthermore, interpretation of GAMs is simpler and more intuitive when compared with a general multiple regression model. In statistical terminology, let Y and $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ be $\mathbb{R}$-valued and $\mathbb{R}^p$-valued random variables respectively, then a GAM is expressed as

$$Y = \mu(\boldsymbol{X}) + \varepsilon = g_1(X_1) + \ldots + g_p(X_p) + \varepsilon, \tag{1.1}$$

where $g_i(\cdot) : \mathbb{R} \to \mathbb{R},\ i = 1, \ldots, p$, are unknown functions and ε is a random term in $\mathbb{R}$.

Virtually, GAM (1.1) simplifies the multiple regression problem by restricting $\mu(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X})$ as a summation of several univariate functions. However, if there is significant nonlinear interaction among the predictors $\{X_1, \ldots, X_p\}$, the additive form in (1.1) will no longer hold. In such case, the estimator $\hat{\mu}$ based on GAMs need not be consistent. More importantly, the validity of using GAMs should be checked.

In reality, it is obvious that people cannot selectively inhale some air pollutants but not others. We also know that two or more pollutants and other hazards may involve in complicated reaction process in atmosphere to affect human health together. Therefore, human health effects should be a result of a complex of inhaled multi-pollutants under certain weather conditions. For example, nitrogen dioxide (NO_2) is oxidized to form nitric acid (HNO_3), which can be neutralized in the atmosphere. Secondary particles produced in this process are usually one dominant component of fine particulate matters (WHO working group, 2003). Hence, the question whether a GAM is valid for time-series air pollution data rises. To date, however, those reports using GAMs to model health impacts only discussed the estimates but not statistically justified the use of GAMs.

Is there any feasible method to assess the performance of GAMs on fitting the associations between mortality rates and air pollutant levels and weather conditions? Is there any improvement in statistical methodology to better estimate the link and to gain deeper understanding? We will discuss these issues in the following chapters.

1.3 Objectives and Organization

In this thesis, we propose a nonparametric approach to quantify the health effects of air pollution and check the performance of GAMs. Instead of directly applying GAMs to time-series air pollution data, we first use the adaptive Effective Dimension Reduction (EDR) method (rMAVE) of Xia *et al*. (2002) to reduce the high dimensionality for general multiple regression problems. By doing so, we preliminarily include interactions across pollutants and weather conditions in those "efficient directions", as well as solve the "curse of dimensionality problem". We then consider the regression problem in the reduced space, comparing a GAM with a general multiple model for the air pollution data. In other words, our approach can be viewed as a two-stages procedure. The first stage is to find the "canonical" variates to reduce the multi-predictor dimension from p to some much smaller integer D; the second stage is to check the validity of a GAM via a cross-validatory criterion which measures models' predictive performance, the regression being applied to the dimension-reduced data.

The rest of this thesis is organized as follows. In the next chapter, Chapter 2, we describe the sources and characteristics of the mortality and pollution data of America under our study. Chapter 3 introduces the nonparametric method involved in this study. One component of our approach is the "rMAVE" dimension reduction method based on a semi-parametric regression model to determine the EDR space; the other component is

the leave-one-out cross-validatory (CV) criterion to check the performance of regression models from their predictive abilities. To check the feasibility of our cross-validatory criterion for model selection, we have conducted some simulations and their typical results are reported in Chapter 4. In Chapter 5, we apply our algorithms to the practical air pollution data and present the results with some discussion. We end this thesis with concluding remarks in Chapter 6. Appendixes are included to illustrate the conditions of a theory and some figures mentioned in the thesis.

Chapter 2

Materials

2.1 Data Source

The data used in subsequent analysis come from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) database. The NMMAPS, sponsored by the Health Effects Institute (HEI), is a systematic investigation of the dependence of mortality rates on air pollution. The database includes various cause-mortality counts, weather conditions and air pollution data for the 108 largest cities in the United States for the 13-year period from January 1st, 1987 to December 31st, 2000.

The NMMAPS data on mortality, weather, census and air pollution were assembled

from publicly available sources. The daily cause-specific mortality counts were obtained from the National Center for Health Statistics and classified into three age groups ($\leq$65 years; 65-75 years; and $\geq$75 years). The daily values of temperature and humidity were obtained from the National Climatic Data Center EarthInfo CD-ROM. Census data about population etc. were drawn from the 2000 Census from the United States Census Bureau. The daily levels of air pollutants, such as PM_{10}, O_3, SO_2, NO_2 and CO, were supplied by the Aerometric Information Retrieval System (AIRS) and the AirData System database maintained by the United States Environmental Protection Agency. The iHASS website (http://www.ihapss.jhsph.edu) contains further detailed information about the NMMAPS database.

2.2 Data Descriptions

The NMMAPS database contains a considerable number of observations and there are many different choices for an interested variable. In our study, we selected the 24-hours mean of temperature and dew point temperature as measurements of meteorology. To measure air pollution levels, we used the 10% trimmed mean and added back yearly average adjustment for each pollutant. Weather conditions (temperature and humidity) and five air pollutants (PM_{10}, O_3, SO_2, NO_2 and CO) consist of our predictor set. As for the response variable, we chose to focus on cardiovascular and respiratory death

Figure 2.1 Locations of the fifteen study cities in United States.

counts for the elder population group (>75 years), since death of cardiovascular and respiratory diseases would be more relevant to a relatively longer exposure period (one month) and adverse health effects of exposure to air pollution would be more significant for the elders.

However, when examining the original data in NMMAPS database, we found that each city has missing values in daily observations. For example, in several locations, there are high percentages of days with missing values for PM_{10} because measurements have been required only once for every six days since 1987 by the Environment Protection Agency. As another example, in several less populous cities, the entire observations of some pollutants (e. g. O_3 and SO_2) are not available. Moreover, daily observations of weather conditions for all cities are only provided from January, 1987 to December, 1998. Therefore, we need to reorganize the original NMMAPS data for analysis.

For previous studies suggested that air pollution may affect mortality with some lags (several days up to a few weeks), we decided to use the original NMMAPS data on monthly basis. That is, we selected the monthly averages of the daily observations for death counts, weather conditions and all pollutants as our primary analytic variables. The missing values were ignored when calculating the monthly means for all variables of interest. After this adjustment and excluding cities which still contain missing values, we have fifteen cities left to be analyzed. Figure 2.1 shows the locations of these 15 cites. From this figure, we observe that most of the 15 cities are in the littoral areas. Note that our study include the three greatest cities: Los Angeles, New York and Chicago.

Chapter 3

Methodology

Essentially, quantification of health effects of air pollution can be viewed as a multiple regression problem with death counts as the response variable, the whole set of various air pollutants and weather variables as multi-predictors. Specifically, let Y and $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ be respectively $\mathbb{R}$-valued and $\mathbb{R}^p$-valued random variables and they are linked in an unknown form

$$Y = g(\boldsymbol{X}) + \varepsilon = \mathbb{E}(Y|\boldsymbol{X}) + \varepsilon,$$

where $\varepsilon \in \mathbb{R}$ is the random error. Then our goal is to approximate $g(\cdot)$ by a function having a simplified structure which makes efficient estimation and meaningful interpretation possible. In recent epidemiological studies on the health impacts of air pollution, the regression function g is often modeled in a nonparametric fashion because of its flexibility in estimating the smooth components and capturing the nonlinear patterns contained in

the air pollution data. In this chapter, we describe the nonparametric method used in our study to explore the associations between mortality rate and air pollution. Our approach can be viewed as a two-stages procedure: 1) efficient dimension reduction through a semi-parametric regression and 2) model selection through a cross-validatory criterion. We will introduce them in the following subsections respectively.

3.1 Dimension Reduction Through Regression

The final goal of a multiple regression analysis is to understand how the conditional distribution of a univariate response Y given a vector $\boldsymbol{X}$ of p predictors depends on the value of $\boldsymbol{X}$. If the conditional distribution of $Y|\boldsymbol{X}$ was completely known for each value of $\boldsymbol{X}$ then the problem would be definitely solved. However, in practice, the study of $Y|\boldsymbol{X}$ is problematic since the dimension of $\boldsymbol{X}$ is quite high and this high dimensional nature makes the estimation challenging. Recent statistical efforts have been spent on efficiently finding the relationship between Y and $\boldsymbol{X}$, essentially via two approaches: one is largely concerned with function approximation and the other is mainly concerned with searching for an Effective Dimension Reduction (EDR) space. In this thesis, we consider an adaptive EDR approach recently proposed by Xia *et al*. (2002), the refined Minimum Average (conditional) Variance Estimation (rMAVE) method based on semi-parametric models. It is easy to implement and needs no strong assumptions on the

probabilistic structure of $\boldsymbol{X}$.

We briefly describe here the basic ideas and main steps of the rMAVE algorithm. Consider a semi-parametric regression-type model for dimension reduction

$$Y = g(B_0^T \boldsymbol{X}) + \varepsilon, \tag{3.1}$$

where $g(\cdot)$ is an unknown smooth link function, $B_0 = (\beta_1, \cdots, \beta_D)$ is a $p \times D$ orthogonal matrix ($B_0^T B_0 = I_D$) with $D < p$, and $\mathbb{E}(\varepsilon|\boldsymbol{X}) = 0$ almost surely. The last condition allows ε to be dependent on $\boldsymbol{X}$. In the terminology of Cook and Weisberg (1999), Model (3.1) implies that the distribution of $Y|\boldsymbol{X}$ is the same as that of $Y|B_0^T\boldsymbol{X}$. Therefore, the p-dimensional predictor $\boldsymbol{X}$ can be replaced by the D-dimensional predictor $B_0^T\boldsymbol{X}$ without loss of regression information and this replacement represents a potentially useful reduction in the dimension of the multi-predictor vector. The space spanned by the columns of B_0 can be uniquely defined under some mild conditions and is called the EDR space. Hence, we will refer to the column vectors of B_0 as the EDR directions, which are unique to the orthogonal transformations.

To estimate the EDR space, we need to estimate the directions B_0 as well as the dimension D. In fact, the direction estimation B_0 is a solution to the problem

$$\min_{B:B^TB=I_D} \mathbb{E}\left[Y - \mathbb{E}(Y|B^T\boldsymbol{X})\right]^2. \tag{3.2}$$

For any orthogonal matrix B, the conditional variance of (3.2) given $B^T\boldsymbol{X}$ is

$$\sigma_B^2(B^T\boldsymbol{X}) = \mathbb{E}\left\{\left[Y - \mathbb{E}(Y|B^T\boldsymbol{X})\right]^2 |B^T\boldsymbol{X}\right\}.$$

It follows that

$$\mathbb{E}\left[\sigma_B^2(B^T\boldsymbol{X})\right] = \mathbb{E}\left[Y - \mathbb{E}(Y|B^T\boldsymbol{X})\right]^2.$$

Therefore, B_0 that minimizes (3.2) is also a solution to

$$\min_{B:B^TB=I_D} \mathbb{E}\left[\sigma_B^2(B^T\boldsymbol{X})\right]. \tag{3.3}$$

This is where the "Minimum Average (conditional) Variance Estimation" or MAVE comes from.

To solve B_0 via (3.3), we have first to estimate the conditional variance function $\sigma_B^2(B^T\boldsymbol{X})$ given $B^T\boldsymbol{X}$. Let $g_B(\boldsymbol{\nu}) = \mathbb{E}(Y|B^T\boldsymbol{X} = \boldsymbol{\nu})$, where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_D)^T$. Given a sample $\{(\boldsymbol{X}_i, Y_i),\ i = 1, \ldots, n\}$, a local linear fit is applied to estimate $g_B(\cdot)$ at point $\boldsymbol{X}_0$. That is,

$$\mathbb{E}(Y_i|B^T\boldsymbol{X}_i) \approx a + \boldsymbol{b}^T B^T(\boldsymbol{X}_i - \boldsymbol{X}_0), \tag{3.4}$$

where $a = g_B(B^T\boldsymbol{X}_0)$ and $\boldsymbol{b}^T = (b_{(1)}, \ldots, b_{(D)})$ with

$$b_{(k)} = \frac{\partial g_B(\boldsymbol{\nu})}{\partial \nu_k}\Big|_{\boldsymbol{\nu}=\boldsymbol{X}_0}, \quad k = 1, \ldots, D.$$

Since the estimation of variance can be expressed as a weighted sum square of residuals, based on (3.4), the estimator of conditional variance σ_B^2 at $B^T\boldsymbol{X}_0$ is

$$\hat{\sigma}_B^2(B^T\boldsymbol{X}_0) = \min_{a,\boldsymbol{b}} \sum_{i=1}^{n} \omega_{i,0}\left(Y_i - \left[a + \boldsymbol{b}^T B^T(\boldsymbol{X}_i - \boldsymbol{X}_0)\right]\right)^2, \tag{3.5}$$

where $\omega_{i,0}, i = 1, \ldots, n$, are properly selected non-negative weight functions at $B^T\boldsymbol{X}_0$.

Now, according to (3.3)-(3.5), the EDR directions B_0 are estimated by solving the minimization problem

$$\min_{B,a_j,b_j} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \omega_{i,j} \left(Y_i - \left[a_j + \boldsymbol{b}_j^T B^T (\boldsymbol{X}_i - \boldsymbol{X}_j) \right] \right)^2 \right), \tag{3.6}$$

where B satisfies $B^T B = I_D$ and

$$\omega_{i,j} = \frac{K_h \left(B^T (\boldsymbol{X}_i - \boldsymbol{X}_j) \right)}{\sum_{l=1}^{n} K_h \left(B^T (\boldsymbol{X}_l - \boldsymbol{X}_j) \right)}$$

is multidimensional kernel weight. As for computation, we start with the identity matrix I_D as an initial estimator of B to be used in the kernel weights. Then iteratively we use the multidimensional kernel weights to obtain an estimator $\hat{B}$ by minimizing problem (3.6), refine the kernel weights with the updated value of $\hat{B}$ and iterate until convergence. The choices of the bandwidth h in kernel weights and the EDR dimension D are implemented through a cross-validatory technique. Moreover, Xia *et al*. (2002) showed that the dimension of the EDR space D can be consistently estimated under some restrictions.

In a word, the rMAVE method may be view as a simultaneous implementation of the EDR direction estimation and the nonparametric link function estimation by local polynomials, showing computational benefits.

3.2 Model Selection Through Cross-Validation

Once we have found the EDR space for a data set, we need to select an appropriate model from a potentially large class of plausible models. In particular to the studies about health effects of air pollution, there are many popular models used to quantify the link as we mentioned in Chapter 1. However, as far as we know, there is no justification for the use of these models, especially for GAMs. In this subsection, we introduce a nonparametric model selection criterion based on the Cross-Validatory (CV) values measuring the predictive performance of models. In the following discussions, we assume the actual dimension of the EDR space is D.

Model selection can be based on subjective judgements as well as on more objective methods. Often the two are combined. The objective methods for model selection have largely been based on either a testing approach or a prediction performance approach. In this study, we adopt the cross-validatory criterion which is a method of evaluating given models by their forecasting ability to choose a model with proper complexity. It is well-known that a cross-validatory approach penalizes the complexity of the model (Stone, 1976).

Cross validation, first suggested by Allen (1974), is a nonparametric model selection technique based on data resampling. It involves dividing the data into two subsamples,

using one (the training set) to estimate the underlying model, and using the other subsample (the validation set) to assess the given model's predictive performance. If the samples in the validation set are well-predicted from the other samples in the training set, it indicates that the model will have good forecasting ability for new samples of the same general population. In the simplest case, the validation set contains only one sample: this is so called the "leave-one-out cross validation" that is broadly used.

Specifically, consider the general framework of nonparametric regression

$$Y = g(\boldsymbol{X}) + \varepsilon, \tag{3.7}$$

where $g(\cdot)$ is an unknown function, $\mathbb{E}(\varepsilon|\boldsymbol{X}) = 0$ and $\mathbb{E}(\varepsilon^2|\boldsymbol{X}) = \sigma^2$ (> 0) almost surely. Assume that $\boldsymbol{X} = (X_1, \ldots, X_D)^T$ is a D (≥ 1) dimensional random vector with finite variance and continuous distribution. Let $\mathscr{P}$ denote the class of non-negative even functions $k(\cdot) : \mathbb{R}^1 \to \mathbb{R}^1$, satisfying

$$\int_{\mathbb{R}^1} k(u)du = 1 \quad \text{and} \quad \int_{\mathbb{R}^1} |u|k(u)du < \infty.$$

Then, for $k \in \mathscr{P}$ and $\boldsymbol{u} = (u_1, \ldots, u_l)^T \in \mathbb{R}^l$, we define

$$K_l(\boldsymbol{u}) = \prod_{i=1}^{l} k(u_i).$$

We attach an l to functions of $K_l : \mathbb{R}^l \to \mathbb{R}^1$ to emphasize the dimension of the kernel function. It is not essential for our results to have K_l in the form of a "product" kernel, so K_l could be any other multiple kernel functions.

Let f (or more precisely f_D) denote the density of $\boldsymbol{X}$, $\{(\boldsymbol{X}_i, Y_i),\ i = 1,\ldots,n\}$ denote the observations from (3.7) and h denote the bandwidth. For $\boldsymbol{x} \in \mathbb{R}^D$, the kernel estimator of the density function f is

$$\hat{f}_n(\boldsymbol{x}) = \frac{1}{nh^D} \sum_{i=1}^{n} K_D \left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h} \right), \tag{3.8}$$

and the kernel estimator for the regression function $g(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X})$ is

$$\hat{g}_n(\boldsymbol{x}) = \frac{1}{nh^D} \sum_{i=1}^{n} Y_i K_D \left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h} \right) \hat{f}_n(\boldsymbol{x})^{-1}. \tag{3.9}$$

Now define the (normalized) residual sum of squares, *RSS*, in our context by

$$RSS = \frac{1}{n} \sum_{j=1}^{n} \{Y_j - \hat{g}_n(\boldsymbol{X}_j)\}^2 \omega(\boldsymbol{X}_j), \tag{3.10}$$

where $\omega(\cdot)$ is a non-negative weight function. A statistic related to *RSS* is the cross-validated residual sum of squares (*CV*),

$$CV = \frac{1}{n} \sum_{j=1}^{n} \{Y_j - \hat{g}_{n,-j}(\boldsymbol{X}_j)\}^2 \omega(\boldsymbol{X}_j), \tag{3.11}$$

where $\hat{g}_{n,-j}(\boldsymbol{x})$ and $\hat{f}_{n,-j}(\boldsymbol{x})$ are as defined by (3.9) and (3.8) respectively, with the exception that now the summations are over $i = 1,\ldots,n$ but $i \neq j$ in each case and the divisor n is replaced by $(n-1)$ for obvious reasons. In fact, $\hat{f}_{n,-j}$ and $\hat{g}_{n,-j}$ are the leave-one-out cross-validatory estimators of f and g, with the observation $\boldsymbol{X}_j$ and the summand $\boldsymbol{X}_j K_D \left(\frac{\boldsymbol{x} - \boldsymbol{X}_j}{h} \right)$ left out respectively, for $j = 1,\ldots,n$.

To justify the use of *CV*-values for model selection, we would need to investigate its sampling properties. By analogy with the classical regression theory, it is expected

that the *RSS* in (3.10) will have an asymptotic bias as an estimator of σ^2 and so is *CV* in (3.11). The following theorem summarizes the main results about the asymptotic behaviors of *RSS* and *CV*. The complete proofs can be found in Cheng and Tong (1993).

Theorem 1 *Under conditions (A.1)-(A.15) which are listed in Appendix A,*

$$RSS = \sigma_n^2 \left(1 - \frac{(2\alpha - \beta)\gamma}{nh^D} + o_p(\frac{1}{nh^D})\right), \tag{3.12}$$

and

$$CV = RSS \left(1 + \frac{2\alpha\gamma}{nh^D} + o_p(\frac{1}{nh^D})\right), \tag{3.13}$$

where

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \omega(\boldsymbol{X}_i), \quad \alpha = K_D^{1/D}(0), \quad \beta = \int K_D^2(\boldsymbol{u})d\boldsymbol{u},$$
$$\textit{and} \quad \gamma = \int \omega(\boldsymbol{x})d\boldsymbol{x} / \int \omega(\boldsymbol{x}) f(x) d\boldsymbol{x}.$$

A direct corollary of this theorem is that

$$CV = \sigma_n^2 \left(1 + \frac{\beta\gamma}{nh^D} + o_p(\frac{1}{nh^D})\right). \tag{3.14}$$

Comparing (3.12) with (3.14), we observe that *RSS* has a negative relative bias $-\frac{(2\alpha-\beta)\gamma}{nh^D}$ while the *CV* has a positive relative bias $+\frac{\beta\gamma}{nh^D}$, although they have the same rate of convergence $(nh^D)^{-1}$.

GAMs are of special interest in the studies of air pollution. We will focus on this model. Now let us discuss the cross-validatory estimation for GAMs and its asymptotic

properties. The GAM is an approach to simplify the fitting of a general multivariate regression model by restricting the form of the regression function $g(\cdot)$ as

$$g(\boldsymbol{X}) = \sum_{j=1}^{D} g_j(X_j). \tag{3.15}$$

The estimation of (3.15) has been investigated extensively. The most popular method is the back-fitting method. The idea of this iterative procedure can be stated as follows:

1) Assign initial functions to each component $g_l(\cdot), l = 1, \ldots, D$. The initial functions can be obtained by spline method or multi-kernel smoothing method.

2) Calculate the estimated partial residual of the l^{th} additive component as

$$\hat{r}_l = Y - \sum_{j \neq l, j=1}^{D} g_j(X_j),$$

and then smooth $\hat{r}_l$ to update $\hat{g}_l(X_l)$ by

$$\hat{g}_l(u) = \frac{\sum_{i=1}^{n} k\left(\frac{X_{i,\,l}-u}{h}\right)\hat{r}_l}{\sum_{i=1}^{n} k\left(\frac{X_{i,\,l}-u}{h}\right)},$$

where u is in the neighborhood of $X_{i,l}$.

3) Repeat the above step until the *RSS* stabilizes.

Since in this estimation procedure each step involves only a univariate kernel estimation, then by Theorem 1 we have the following conjectures:

$$RSS = \sigma_n^2 \left(1 - \frac{c_1}{nh} + o_p(\frac{1}{nh})\right), \tag{3.16}$$

and

$$CV = \sigma_n^2 \left(1 + \frac{c_2}{nh} + o_p(\frac{1}{nh})\right), \tag{3.17}$$

where c_1 and c_2 are constant numbers in $\mathbb{R}^+$. To avoid confusion, denote *RSS* in (3.16) and *CV* in (3.17) for GAMs as RSS^A and CV^A respectively, and denote *RSS* in (3.12) and *CV* in (3.14) for general multivariate models as RSS^G and CV^G respectively.

Based on those notations and discussions, we now construct our model selection criterion across several candidates, particularly between a GAM and a general multivariate model. When comparing RSS^A with RSS^G, no matter whether g does satisfy the additive form (3.15) or not, it is observed that

$$RSS^A > RSS^G$$

always hold for sufficiently large n such that $h < 1$. Therefore, *RSS* does not have ability to differentiate a GAM from a general multiple model. Hence, *RSS* can not be used as a model selection criterion. However, when we compare CV^A with CV^G, the situation is completely different. If g satisfies the additive form (3.15), for sufficiently large n such that $h < 1$, we have

$$CV^A < CV^G.$$

On the other side, if the additive form (3.15) is not correct but we still use GAMs to fit the data, the kernel estimator $\hat{g}$ will have a fixed bias resulting in large *CV*-value. A natural conjecture for CV^A is that $CV^A \to \sigma_n^2(1+c_3)$, as $n \to \infty$, where c_3 is a positive real number. As a consequence, for sufficiently large n, we have

$$CV^A > CV^G$$

if the true model is not additive. Note that the general multivariate model is always true. In conclusion, CV-value has the capability to tell a GAM from a general multiple regression model.

Now let us summarize our model selection procedure based on the CV-value criterion for the given model's forecasting ability. Firstly, for each candidate model, we replace Y_j by $\hat{g}_{n,-j}$, the kernel estimator of the conditional mean function based on observations $\{(\boldsymbol{X}_i, Y_i),\ i = 1,\ldots,n,\ i \neq j\}$. Secondly, we evaluate the weighted cross-validatory residual sum of squares defined in (3.11), especially CV^G and CV^A. Then we minimize CV-values with respect to h over a suitably prefixed range. Finally, the CV-value of each candidate model is pooled for comparison. The model with the smallest CV-value is preferred.

All analysis were carried out using both Matlab (The MathWorks, Natick, Massachusetts) and R (http://www.r-project.org/; Version 2.0.0).

Chapter 4

Simulations

In this chapter, we carry out simulations to check the performance of the proposed cross-validatory criterion to select a model for its forecasting ability described in the previous chapter.

Consider the following model

$$Y = \lambda(X_1 + X_2) + (1-\lambda)X_1X_2 + \sigma\varepsilon, \tag{4.1}$$

where

1) $X_1,\ X_2$ are independent random variables with a uniform distribution over the interval [0, 1],
2) ε is a random variable with a standard normal distribution *Normal* (0,1) and independent with $X_1,\ X_2$,

3) λ is a constant number in the range of [0, 1], and

4) σ is a positive constant to adjust the effects of the error term which is additive to the link function.

Model (4.1) is a weighted average of the additive term $(X_1 + X_2)$ and the interaction term $(X_1 \times X_2)$. For different weights λ in [0,1], we shall obtain different models. By following the procedure described in previous chapter, we can calculate CV^A- and CV^G-values for GAMs and general multiple models respectively, and then compare the two CV-values under each model. The main idea here is that, by changing the weight λ from 0 to 1, the underlying model (4.1) is changing from a model with only an interaction term

$$Y = X_1 X_2 + \sigma\varepsilon, \tag{4.2}$$

to an pure additive model

$$Y = X_1 + X_2 + \sigma\varepsilon. \tag{4.3}$$

When λ is close to 1, the additive term plays a more important role than the interaction term in the underlying model (4.1). As a result, the calculated CV-values corresponding to GAMs should be consistently smaller than those for general multiple models, namely, $CV^A < CV^G$ in general. However, if λ approaches to 0, the situation would be entirely reversed. That is, the interaction term will have more significant effects on the underlying model (4.1) and the calculated CV-values for general multivariate models would ususally be smaller than those for GAMS, or $CV^A > CV^G$ generally. Therefore, we can count the number of smaller CV-values for GAMs or for general multi-models in many

Table 4.1 Number of $CV^A > CV^G$ in 100 replications for the model (4.1).

λ	σ=0.1				σ=0.3				σ=0.5			
	n=50	100	200	400	n=50	100	200	400	n=50	100	200	400
0.00	93	100	100	100	32	60	88	96	40	51	62	71
0.05	91	98	100	100	50	57	74	96	40	45	42	70
0.10	86	100	100	100	35	47	70	93	26	38	42	62
0.15	75	100	100	100	31	50	66	88	31	35	48	59
0.20	70	97	100	100	28	42	51	82	25	32	42	40
0.25	63	97	100	100	24	35	52	71	28	30	36	45
0.30	50	88	100	100	30	34	30	66	30	31	33	29
0.35	35	85	99	100	12	26	37	62	21	31	24	31
0.40	40	78	95	99	17	21	30	52	17	34	23	25
0.45	36	50	89	100	16	18	27	35	20	16	21	17
0.50	18	54	78	99	13	17	20	36	19	19	19	10
0.55	21	27	59	99	18	15	12	28	17	13	17	9
0.60	5	16	18	84	19	8	10	21	21	18	10	13
0.65	12	12	20	58	8	11	10	13	23	17	8	10
0.70	5	1	2	44	11	12	8	7	13	12	13	6
0.75	3	6	1	17	3	7	6	5	20	15	9	4
0.80	3	1	2	9	8	8	5	6	12	12	3	7
0.85	4	1	2	0	9	3	4	5	18	12	7	7
0.90	0	1	2	0	12	4	3	3	11	15	5	4
0.95	0	2	0	0	6	1	5	2	11	9	4	8
1.00	1	1	0	0	5	5	6	1	15	9	6	8

replications to check the performance of our model selection criterion. Additionally, it is noticeable that we have assumed an additive random error. Thus, it will also have effects on the simulation results.

In our simulations, we have set σ at 0.1, 0.3, and 0.5 with the sample size n = 50, 100, 200 or 400 respectively, and drawn 100 replications in each case. Let λ increase from 0 to 1 and count the number of $CV^A > CV^G$. Table 4.1 lists the simulation results. The results suggest that the number of $CV^A > CV^G$ decreases as λ increases in general. This observation indicates that, a general multiple model would outperform a GAM for

their predictive performance when λ is small, namely, the effects of additive terms can be ignored. From the other side, when λ becomes greater, the general model would underperform a GAM. This is what we have expected and suggests that our CV-value criterion is feasible. We also observed that, when the sample size n is increasing from 50 to 400, the difference between the frequency with smaller λ and that with larger λ in each column becomes greater for every fixed σ. It implies that our proposed CV-value model selection criterion has better and more stable performance with larger sample size. Moreover, the results show that the additive error term has impacts on the performance of our selection method. As σ becomes greater, the numbers of $CV^A > CV^G$ for small λ (close to 0) reduce significantly, which is the result of the additive error assumption.

All of these observations from Table 4.1 are consistent with what we have supposed before simulations. Therefore, we can conclude that our CV-value based model selection criterion could be used for detecting whether a model contains significant nonlinear interaction terms across the predictors. In particular, this criterion can pick a GAM out from general nonparametric models for sufficiently large samples.

Chapter 5

Results and Discussions

In this chapter, we present the results of performing the nonparametric methodology proposed in Chapter 3 to explore the association between exposure to air pollution and its effects on public health across 15 populous cities in the United States. The selection of these cities refers to Chapter 2. Our study is a city-specific analysis. That is, for each of the 15 cities with the data about air pollutants, weather conditions and mortality counts from the NMMAPS database, we first employ the rMAVE method to search for its EDR space, and then apply the cross-validatory criterion to compare a GAM with a general multiple model for their predictive performances in the EDR space.

5.1 Preliminary Analysis

Before we carry out nonparametric regression analysis, we perform preliminary statistic analysis to capture basic characteristics of the data that was reorganized on a monthly basis in Chapter 2. The following discussions present the results.

Table 5.1 provides elementary information about the 15 cities during the study period, including population sizes, means and standard errors of death counts, weather conditions, and levels of air pollutants. The monthly counts for death of cardiovascular and respiratory diseases among the 15 cities is about 18 persons per month on average. The lowest average monthly death rate (2 persons per month) appears in Baton Rouge with the smallest population size among the 15 cities and the highest average monthly death rate (69 persons per month) is shown in New York with the second largest population size. Generally speaking, the death rate is linearly correlated with the population size of each city. As we expected, the monthly mean temperature and dew point temperature of each city are related to its geographic location. As for the air pollutants, none of them exceeds the WHO recommended criterion but CO shows large values in all cities compared with other pollutants. It is also observed that the levels of air pollutants vary from one city to another. In general, the cities with greater population have relatively higher air pollution levels.

Table 5.1 Descriptive Characteristics of the Fifteen Study Cities

City	*Population*	*Number of Deaths*	*Temperature* (°F)	*Dew Point* (°F)	PM_{10} (mg/m^3)	O_3 (*PPB*)	SO_2 (*PPB*)	NO_2 (*PPB*)	*CO* (*PPB*)
Baton Rouge	412852	2.0 (0.37)	68.3 (10.53)	58.7 (11.43)	27.3 (6.95)	23.6 (5.78)	5.8 (1.59)	15.0 (2.92)	466.3 (229.60)
Boston	689807	3.9 (0.65)	52.1 (15.23)	39.4 (15.74)	26.5 (5.76)	23.0 (8.54)	8.9 (4.53)	25.0 (2.77)	1347.6 (179.52)
Buffalo	950265	9.2 (1.15)	49.4 (17.12)	38.7 (15.25)	22.1 (7.34)	24.0 (8.44)	9.3 (2.40)	19.3 (2.62)	732.1 (125.77)
Chicago	5376741	36.1 (4.04)	50.8 (17.76)	39.9 (15.76)	34.0 (7.35)	20.2 (7.58)	5.8 (1.75)	25.0 (2.66)	969.5 (101.16)
Dallas/Fort Worth	4199873	16.2 (2.49)	68.4 (14.03)	51.7 (12.94)	26.2 (6.28)	27.2 (7.95)	3.5 (1.58)	14.9 (2.63)	676.6 (171.31)
El Paso	679622	2.3 (0.43)	68.3 (14.22)	35.7 (12.56)	36.7 (11.18)	25.7 (7.96)	9.1 (3.85)	19.3 (4.88)	1207.9 (417.34)
Houston	3400578	10.8 (1.59)	70.2 (11.10)	59.1 (10.91)	27.1 (7.25)	22.3 (6.01)	3.2 (0.79)	17.6 (3.54)	863.3 (233.33)
Jersey City	608975	3.4 (0.59)	56.8 (16.19)	42.2 (16.03)	31.4 (6.08)	20.8 (10.15)	10.6 (4.21)	28.6 (4.07)	2103.3 (359.13)
Los Angeles	9519338	54.7 (8.62)	63.4 (4.92)	52.7 (7.00)	42.8 (10.79)	24.5 (9.82)	2.5 (1.09)	36.2 (7.20)	1428.1 (640.13)
New York	8931737	68.9 (8.19)	55.6 (15.17)	42.4 (15.83)	30.2 (7.27)	20.3 (7.89)	13.5 (6.02)	32.3 (4.44)	2121.1 (213.06)
Philadelphia	1517550	11.9 (1.57)	57.1 (15.95)	43.8 (15.81)	41.5 (8.80)	22.4 (9.79)	10.0 (4.51)	32.4 (3.71)	1249.7 (255.19)
Pittsburgh	1281666	13.2 (1.79)	53.8 (16.45)	40.8 (15.05)	32.0 (8.73)	21.8 (8.56)	15.3 (4.08)	27.8 (2.72)	1184.6 (321.05)
Riverside	1545387	8.4 (1.69)	65.2 (7.77)	47.5 (7.21)	43.6 (14.40)	35.1 (13.32)	0.8 (1.01)	21.3 (5.08)	1065.9 (375.96)
San Bernardino	1709434	7.5 (1.32)	65.2 (7.77)	47.5 (7.21)	41.7 (13.32)	38.9 (14.09)	2.3 (0.96)	23.8 (5.29)	792.9 (268.66)
San Diego	2813833	15.6 (2.57)	63.6 (4.69)	54.3 (6.47)	34.0 (7.15)	31.7 (7.64)	2.8 (1.00)	22.9 (6.79)	1185.1 (465.99)

Note. Values are means (+ SE).

We include the time-series plots of the monthly data for several cities as examples in Appendix B. See Figures B.1-B.5. Although each city has its own characteristics, those figures illustrate some general features across the 15 cities. One common point is the clear seasonal variation in the patterns of temperature and humidity (or dew point temperature). Another point is that the mortality rate and the levels of the 5 air pollutants show certain degree of seasonable behaviors. This observation implies that weather conditions affects not only the number of death of cardiovascular and respiratory diseases but also the levels of air pollution. It reminds us to notice the possible existence of collinearity or nonlinear interaction in the multi-predictors when carrying out regression.

We have also examined correlations between the variables involved in our study. See Figures C.1-C.5 in Appendix C. These scatter matrix plots are for the same cities involved in Appendix B and typical to present some common properties. From the figures, we find that all predictors (temperature, humidity, PM_{10}, O_3, SO_2, NO_2 and CO) tend to have moderately strong correlations with mortality on average. Moreover, correlations among the predictors cannot be ignored. It implies that when we carry out regression on mortality rate with the predictors, the model may be difficult to be interpreted because of the collinearity in the predictors.

Based on those observations, we now construct the regression model on air pollution data. Consider the relative mortality rate as the response variable, that is, making a

logarithmic transformation of the average monthly mortality rate. As for the multi-predictors, we combine the weather-based variables (temperature (temp) and dew point temperature (humi)) and the air pollutants (PM_{10}, O_3, SO_2, NO_2 and CO) together. Then a general multivariate regression model is

$$\log(\text{death rate}) = g(\text{temp}, \text{humi}, PM_{10}, O_3, SO_2, NO_2, CO) + \varepsilon, \tag{5.1}$$

where ε is a $\mathbb{R}^1$-valued random variable. Our aim is to understand how these air pollutants and weather conditions affect public health via estimating $g(\cdot)$. For using the rMAVE method to reduce the high dimensionality in the predictors, all the variables in (5.1) need to be standardized. Furthermore, to avoid collinearity in the multi-predictors, we standardize the design matrix

$$\boldsymbol{X} = (\text{temp}, \text{humi}, PM_{10}, O_3, SO_2, NO_2, CO)_{n\times 7}$$

by its square root of covariance matrix $\boldsymbol{S}^{1/2} = (\boldsymbol{X}^T\boldsymbol{X})^{1/2}$. Let Y denote $\log(\text{death rate})$, $\tilde{\boldsymbol{X}} = (\widetilde{\text{temp}}, \widetilde{\text{humi}}, \widetilde{PM_{10}}, \widetilde{O_3}, \widetilde{SO_2}, \widetilde{NO_2}, \widetilde{CO})$ denote $\boldsymbol{X}\boldsymbol{S}^{-1/2}$. Therefore, after those transformations, (5.1) becomes

$$\log(\text{death rate}) = g(\widetilde{\text{temp}}, \widetilde{\text{humi}}, \widetilde{PM_{10}}, \widetilde{O_3}, \widetilde{SO_2}, \widetilde{NO_2}, \widetilde{CO}) + \varepsilon,$$

or more concisely

$$Y = g(\tilde{\boldsymbol{X}}) + \varepsilon,$$

where g is different from the one in (5.1).

Table 5.2 Estimated EDR dimension D for the 15 cities with h and CV-values.

City	*Dimension*	*Bandwidth*	*CV value*
Baton Rouge	3	0.034	0.52824
Boston	3	0.082	0.39871
Buffalo	2	0.037	0.33285
Chicago	2	0.028	0.31616
Dallas/Fort Worth	2	0.026	0.30465
El Paso	2	0.027	0.44786
Houston	3	0.044	0.32739
Jersey City	2	0.026	0.46318
Los Angeles	4	0.047	0.21953
New York	3	0.055	0.18693
Philadelphia	2	0.025	0.22150
Pittsburgh	2	0.025	0.27411
Riverside	3	0.035	0.34079
San Bernardino	3	0.045	0.32453
San Diego	3	0.040	0.22447

5.2 Dimension Reduction

Now let us search for the EDR space of the air pollution data in each city. Consider a semi-parametric model $Y = g(B_0^T \boldsymbol{X}) + \varepsilon$, or exactly the one after standardization,

$$Y = g(\tilde{B}_0^T \tilde{\boldsymbol{X}}) + \varepsilon, \tag{5.2}$$

and apply the rMAVE method to estimate $\tilde{B}_0$.

Table 5.2 shows the estimated dimension $\hat{D}$ of the EDR space for each city. It is found that $\hat{D}$ is most of the time equal to 2 or 3, with the exception of Los Angeles where $\hat{D}$ is 4. Comparing with the original 7 covariates, the number of predictors has been substantially reduced through linear combinations of the original covariates. It implies that the regression information contained in the original 7 covariates could be summarized into

a few directions of the EDR space. The reason that Los Angeles has a relatively higher dimension of the EDR space may be attributed to its largest population size among the 15 cities. Table 5.2 also lists the corresponding bandwidths and cross-validatory values for deciding the EDR dimensions of the 15 study cities. As we used the standardized observations, loosely speaking, we may interpret the CV-values as unexplained variation in Model (5.2) because CV can be view as an asymptotically biased estimator of the variance of the random term ε. The maximal CV-value is about 0.53 for Baton Rouge and the minimum is about 0.19 for New York City.

To further investigate the associations between the air pollution, weather conditions and the adverse health impacts, we should examine the corresponding direction estimates $\hat{B}_0 = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_{\hat{D}})_{p\times\hat{D}}$. In fact, the direct result of direction estimation from the rMAVE method is $\hat{\hat{B}}_0$, so we need to multiply $\hat{\hat{B}}_0$ by $\boldsymbol{S}^{-1/2}$ to obtain the meaningful EDR directions $\hat{B}_0$, namely, $\hat{B}_0 = \boldsymbol{S}^{-1/2}\hat{\hat{B}}_0$. Naturally, $\hat{B}_0$ need not satisfy the orthogonal condition that $\hat{B}_0^T\hat{B}_0 = I_{\hat{D}}$. Table 5.3 shows $\hat{B}_0$ for all the 15 cities. We notice that, in each city, every covariate partly contributes to all the EDR directions with small coefficients, none of which exceeds 0.5. This is related to the standardized transformation on data before applying the rMAVE method. To illustrate the meaning of an EDR direction estimation in Table 5.3, we select the third EDR direction for Baton Rouge as an

Table 5.3 Estimated EDR directions $\hat{B}_0^T = (\hat{\beta}_1, \ldots, \hat{\beta}_D)^T$ for the 15 cities.

City	*Dim*	*Temp*	*Humi*	PM_{10}	O_3	SO_2	NO_2	*CO*
Baton Rouge	1	-0.1218	0.0437	-0.0070	0.0399	-0.0003	0.0290	0.0031
	2	-0.0547	0.1539	-0.0443	-0.0059	-0.0348	**0.0863**	0.0098
	3	**0.1692**	**-0.1663**	-0.0298	0.0422	-0.0534	0.0430	-0.0423
Boston	1	0.0265	-0.0641	0.0025	0.0450	0.0523	-0.0029	0.0391
	2	**-0.4125**	**0.3744**	0.0722	-0.0298	0.0175	-0.0174	-0.0766
	3	0.1703	-0.1347	0.0341	-0.0292	-0.0330	-0.0963	**0.1423**
Buffola	1	0.0578	0.0294	0.0062	-0.0081	-0.0183	0.0075	0.0100
	2	**0.3416**	**-0.3125**	**-0.0571**	**0.0559**	0.0498	0.0020	-0.0119
Chicago	1	0.0128	0.0759	-0.0063	-0.0036	0.0152	0.0169	-0.0205
	2	**0.2039**	**-0.2121**	**0.0698**	-0.0299	0.0298	-0.0210	0.0222
Dallas/ Fort Worth	1	0.1027	-0.0164	-0.0129	0.0042	-0.0261	0.0126	0.0179
	2	**-0.0429**	**0.0870**	0.0061	-0.0435	**0.0606**	-0.0575	0.0228
El Paso	1	0.0908	0.0307	0.0092	-0.0133	0.0353	-0.0182	0.0334
	2	0.0363	0.0034	-0.0264	0.0135	0.0153	**0.0795**	-0.0741
Houston	1	0.1129	-0.0244	0.0118	0.0024	0.0069	0.0098	0.0356
	2	-0.2596	0.2191	-0.0351	**0.0977**	-0.0290	-0.0259	0.0089
	3	**-0.2850**	**0.2998**	0.0025	0.0272	**0.0834**	-0.0522	0.0486
Jersey City	1	**-0.0898**	**0.0396**	-0.0385	0.0622	0.0868	0.0074	-0.0248
	2	-0.0113	-0.1222	0.0970	-0.0291	**-0.1286**	-0.0058	0.0844
Los Angeles	1	-0.0407	-0.0214	0.0010	-0.0239	0.0287	0.0069	-0.0160
	2	0.0571	-0.0244	-0.0220	0.0456	-0.0514	0.0858	0.0766
	3	-0.0132	-0.0460	-0.0342	**0.2154**	-0.0046	-0.0986	**0.2321**
	4	**-0.1215**	**0.1997**	-0.0266	-0.0079	0.0899	-0.0229	0.0289
New York	1	0.0965	-0.0033	-0.0156	-0.0056	-0.0014	0.0070	0.0065
	2	**-0.2264**	**0.3508**	0.0099	-0.0596	0.0846	0.0148	-0.0953
	3	0.2781	-0.2263	-0.0162	**-0.1494**	-0.0603	0.0302	0.0055
Philadelphia	1	**-0.1518**	**0.0370**	-0.0038	**0.0717**	-0.0047	-0.0105	0.0380
	2	0.0194	-0.0264	0.0592	0.0478	0.0651	-0.0071	-0.0207
Pittsburgh	1	-0.1731	0.2573	-0.0373	0.0294	-0.0112	0.0007	0.0392
	2	**-0.4138**	**0.3714**	**-0.0445**	**0.0540**	-0.0109	0.0132	-0.0105
Riverside	1	-0.0065	0.0017	0.0215	0.0849	-0.0217	0.0022	0.0421
	2	**0.1574**	**-0.0996**	0.0479	-0.0991	0.0587	-0.0912	0.0090
	3	-0.0801	0.1037	**0.1030**	**-0.1120**	-0.0391	-0.0786	0.0447
San Bernardino	1	-0.0006	0.0294	0.0071	0.0597	0.0197	-0.0034	0.0269
	2	0.0637	0.0338	0.0555	-0.1034	0.0172	-0.0777	0.0542
	3	**-0.1832**	**0.0367**	-0.0223	**0.1160**	0.0294	0.0707	0.0200
San Diego	1	0.0039	0.0589	0.0025	0.0604	-0.0329	-0.0135	0.0637
	2	**0.0718**	**-0.1144**	-0.0245	0.0822	0.0570	0.0789	-0.0623
	3	0.0039	0.0916	-0.0090	0.0367	-0.0015	-0.1705	**0.3039**

Note. Bold-faced entries have relatively large absolute values.

example:

$$\begin{aligned} \boldsymbol{X}\hat{\boldsymbol{\beta}}_3 &= 0.1692\ \text{temp} - 0.1663\ \text{humi} - 0.0298\ \text{PM}_{10} + 0.0422\ \text{O}_3 \\ &\quad -0.0534\ \text{SO}_2 + 0.043\ \text{NO}_2 - 0.0423\ \text{CO}. \end{aligned}$$

In this EDR direction, temperature and dew point temperature are two dominant components because of their relatively greater coefficients. From the city-specific estimates in Table 5.3, we have observed that the fitted structures of the EDR space differ from city to city. However, we can still summarize two general but important features as follow:

1) The weather covariates are influential. Notice that many large coefficients are assigned to temperature and humidity in all the 15 cities. In particular, every city, with the exception of El Paso, contains an EDR direction having large coefficients for temperature and humidity at the same time but with different signs.
2) Different city has different influential pollutants. Notice that there is at least one pollutant presenting a relatively larger coefficient in each city. O_3 and PM_{10} seem to be the most influential because they show the highest frequencies of relatively larger absolute values in all 15 cities. And the least influential seems to be NO_2 with only two large values.

These statistical observations are consistent with the practical and medical observations. It is well-known that the extremes of temperature, both cold and hot, are not favorable to the disease suffers: viruses can survive longer and transmit in cold season

thus exacerbating other diseases; hot weather increases the risk of dehydration and other adverse effects. Humidity is also important to cardiovascular and respiratory diseases in that under wetter conditions it is easier for fungal to colonize, thus worsening the air quality and causing health problems. Especially, when both undue temperature and undue humidity conditions concur, e. g. either extremely hot and dry weather or extremely cold but high humid weather, their impacts on disease suffers could be even more severe.

Furthermore, our findings suggest that PM_{10}, O_3, SO_2, NO_2, and CO at current levels are poisonous to public heath, though their toxicities are different. The small coefficients indicate that their toxicities on human heath are not acute, but long-term exposure at current levels may be fatal. This evidence is consistent with the epidemiological understanding we introduced in Chapter 1. Moreover, it is an obvious phenomenon that air quality fluctuates by place and time, depending on many factors, such as type and composition of the gasoline used locally, type of emission sources, presence of local industrial sources and geographical location. Hence, the variation in city-specific estimations about air pollution is reasonable.

To summarize, the above results and discussions suggest that air pollution at current levels in populous cities in the United States have impacts on the overall mortality of elder population ($>$75 years). In particular, weather conditions, ozone and particular matter ($<10\mu$m) are more influential hazards to the death of cardiovascular and respiratory diseases.

5.3 Model Selection

In previous subsection, we have found the EDR directions and the influential factors in those directions for all the 15 study cities. However, for each city, it remains unknown how those linear combinations of the original covariates (i. e. those EDR directions) affect the relative mortality rates. Now we discuss this problem mainly by nonparametric fittings because of their flexibility on nonlinear smoothing. Especially, since Generalized Additive Models (GAMs) are prevalent in recent epidemiological studies on air pollution and heath effects, we focus on whether a GAM is superior to a general multivariate model for the air pollution data. The model selection criterion is the one based on the cross-validatory value to measure the predictive performance of a model, introduced in Chapter 3.

Consider the following models as our candidates to compare their forecasting ability for the air pollution data in the 15 cities:

1) Generalized Additive Model,

$$Y = g_1(X_1) + \ldots + g_7(X_7) + \varepsilon; \tag{5.3}$$

2) General multivariate regression model,

$$Y = g(X_1, \ldots, X_7) + \varepsilon = g(\boldsymbol{X}) + \varepsilon; \tag{5.4}$$

3) Generalized Additive Model of EDR components,

$$Y = g_1(\boldsymbol{\beta}_1^T \boldsymbol{X}) + \ldots + g_D(\boldsymbol{\beta}_D^T \boldsymbol{X}) + \varepsilon; \tag{5.5}$$

4) General EDR model,

$$Y = g(\boldsymbol{\beta}_1^T \boldsymbol{X}, \ldots, \boldsymbol{\beta}_D^T \boldsymbol{X}) + \varepsilon = g(B_0^T \boldsymbol{X}) + \varepsilon. \tag{5.6}$$

Note that in above equations (5.3)-(5.6), Y denotes $\log(\text{death rate})$, $\boldsymbol{X} = (X_1, \ldots, X_7)$ denotes $(\text{temp}, \text{humi}, \text{PM}_{10}, \text{O}_3, \text{SO}_2, \text{NO}_2, \text{CO})$, and ε denotes the random error. We include Models (5.3) and (5.4) also in our candidates to check the performance of EDR.

Now, for Models (5.3) and (5.4), directly use the air pollution data to fit and then compute the CV-values for each city under the two model respectively. While for Models (5.5) and (5.6), we first apply the rMAVE method to form the EDR data set for every city, then fit the two models using those restructured data sets, and finally calculate the corresponding CV-values. Then we have the following results shown in Table 5.4.

Our proposed criterion is to choose the model with relatively small CV-values. Firstly, we limit the comparison among the models before dimension reduction: GAM (5.3) and the general form (5.4). As shown in the left two columns in Table 5.4, each of the 15 cities has a smaller CV-value from the additive model than the CV-value from the general model ($CV^A < CV^G$), though the difference is subtle. This observation indicates that, for the original air pollution data, GAM (5.3) exhibits better performance than a

Table 5.4 Results of CV-value criterion for the 15 cities

City	***Before EDR***		***After EDR***	
	Additive	*General*	*Additive*	*General*
Baton Rouge	0.6700	0.6802	0.6472	**0.5251**
Boston	0.4839	0.5180	0.4294	**0.3939**
Buffalo	0.3800	0.4106	0.3448	**0.3382**
Chicago	0.3697	0.4029	0.3562	**0.3313**
Dallas/Fort Worth	0.3561	0.4402	0.3518	**0.3046**
El Paso	0.5549	0.6013	0.4945	**0.4419**
Houston	0.4487	0.4525	0.4183	**0.3141**
Jersey City	0.5903	0.6178	0.5416	**0.4600**
Los Angeles	0.2883	0.3050	0.2856	**0.2235**
New York	0.2467	0.2549	0.2496	**0.1878**
Philadelphia	0.3194	0.3533	0.2346	**0.2232**
Pittsburgh	0.3320	0.4050	0.2839	**0.2747**
Riverside	0.4454	0.4995	0.4333	**0.3803**
San Bernardino	0.4288	0.4435	0.4288	**0.3366**
San Diego	0.2621	0.3212	0.2763	**0.2603**

Note. Bold-faced entries are the smallest CV-values in rows.

general multiple model (5.4) in view of predictive ability. Therefore, it is valid to widely use GAMs to model the impacts of air pollution on human health in epidemiological studies, if we only have these two choices. The benefit of this common practice is that GAMs can improve our understanding of the link between health effects and air pollution, weather conditions in the sense that GAMs depict how each covariate nonlinearly affects the response variable, conditional on the other smooth functions in the model. However, the disadvantage of acceptance GAMs is that we believe the nonlinear interactions among the covariates are so small that could be ignored.

Now let us test the performance of EDR. After including Models (5.5) and (5.6) in our comparison, we reconsider the model selection problem for air pollution data. Table 5.4 illustrates that, among the four candidate models (5.3)-(5.6), the general multiple model after EDR searching through the rMAVE method (5.6) shows consistently smaller CV-values than those of the other three models for all cities. Another observations is that the additive model before dimension reduction (5.3) shows greater CV-values than those from the additive model after EDR (5.5). Hence, Model (5.6) is the preferred model across the four candidates, according to the CV-value based model selection criterion. These imply that EDR is necessary to quantify the effects of air pollution for the original data, and more importantly, that a GAM is NOT superior to general multivariate model once dimension reduction has been considered. We may interpret this as the "price" of using the reduced set with $2\sim4$ covariates instead of the original set with 7 covariates. We lose the intuitively interpreted feature of GAMs but we gain the better predictive ability in general multiple models. Furthermore, Model (5.6) raise our attention to the problem of nonlinear interactions among various pollutants and weather variates, although it can not explicitly figure out the interdependence.

To further illustrate that the GAM with the covariate set reduced by the rMAVE method (5.5) does not throw much light on understanding how the linearly combined covariates act together to influence the relative mortality rate, we include Figures 5.1-5.5 for several cities as representatives. Suppose that, for each city, we have fitted the

reduced air pollution data $(\hat{\boldsymbol{\beta}}_1^T \boldsymbol{X}, \ldots, \hat{\boldsymbol{\beta}}_D^T \boldsymbol{X})$ with GAM (5.5) and obtained the estimates $\hat{g}_1(\hat{\boldsymbol{\beta}}_1^T \boldsymbol{X}), \ldots, \hat{g}_D(\hat{\boldsymbol{\beta}}_D^T \boldsymbol{X})$. Then we define the estimated partial residuals for the j[th] EDR direction as

$$\hat{r}_j = Y - \sum_{i \neq j, i=1}^{D} \hat{g}_i(\hat{\boldsymbol{\beta}}_i^T \boldsymbol{X}), \quad j = 1, \ldots, D.$$

In those figures, we exhibit scatter plots of the estimated partial residuals for every EDR direction ($\hat{r}_j$ against $\hat{\boldsymbol{\beta}}_j^T \boldsymbol{X}$, j = 1, …, D), with superimposed lines being the corresponding smoothed additive terms ($\hat{g}_j(\hat{\boldsymbol{\beta}}_j^T \boldsymbol{X})$, j = 1, …, D). Now examining those figures for all cities, we find that those lines do not capture the patterns in the partial residuals in general. In some cases, the fitted lines are undersmoothed, such as those in Dalls/Fort Worth. In other case, the lines are oversmoothed for the clustered points, such as those in Los Angeles.

For comparison, we also include Figures 5.6-5.10 for the same cities in the end of this chapter, produced in the identical way as mentioned above, with the exception that do not apply the rMAVE method to estimate the EDR space. Specifically, for each city, we fit GAM (5.3) to the air pollution data $\boldsymbol{X}_{n\times 7} = (\text{temp}, \text{humi}, \text{PM}_{10}, \text{O}_3, \text{SO}_2, \text{NO}_2, \text{CO})$, and then plot the estimated partial residuals for every additive univariate function with the smoothed fitted lines superimposed. From these figures, we observe that the fits do not give much hint on understanding of the relationship between air pollution, weather conditions and relative mortality rates. However, when considering the impacts of using EDR, we observed that the partial residual plots of after EDR contain relatively

more information than those of without EDR, which indicates that EDR has effects on improvement of fitting the air pollution data.

In conclusion, our results suggests that the best model among the four candidate models (5.3)-(5.6), for this practical air pollution data set, is the one that incorporates EDR by the rMAVE method in a general multiple regression model (5.6), according to the proposed CV-value based model selection criterion that measures the predictive performance of one model.

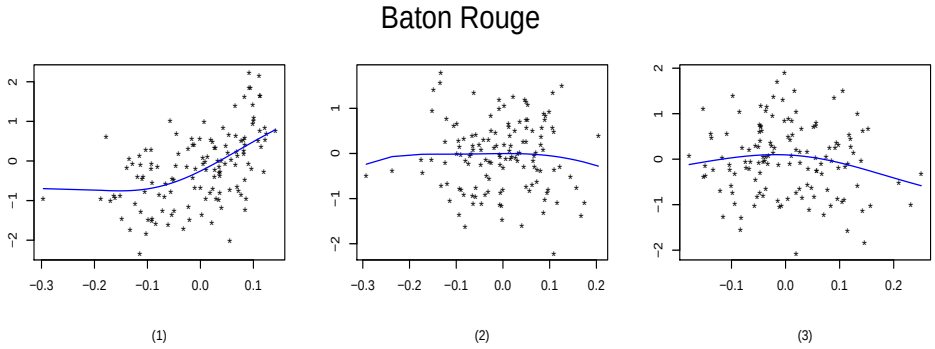

Figure 5.1 (1)-(3) are the scatter plots of the estimated partial residuals for the jth EDR direction $\hat{r}_j$ against the jth EDR direction $\hat{\boldsymbol{\beta}}_j^T \boldsymbol{X}$, j = 1, 2, or 3, respectively. The superimposed lines are the smoothed estimated additive terms $\hat{g}_j(\hat{\boldsymbol{\beta}}_j^T \boldsymbol{X})$, j = 1, 2, or 3, to help visualization.

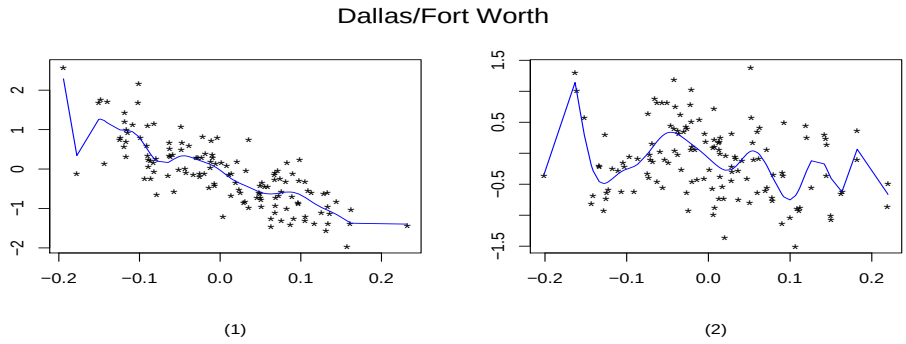

Figure 5.2 (1)-(2) are the scatter plots of the estimated partial residuals for the j[th] EDR direction $\hat{r}_j$ against the j[th] EDR direction $\hat{\boldsymbol{\beta}}_j^T \boldsymbol{X}$, j = 1 or 2, respectively. The superimposed lines are the smoothed estimated additive terms $\hat{g}_j(\hat{\boldsymbol{\beta}}_j^T \boldsymbol{X})$, j = 1 or 2, to help visualization.

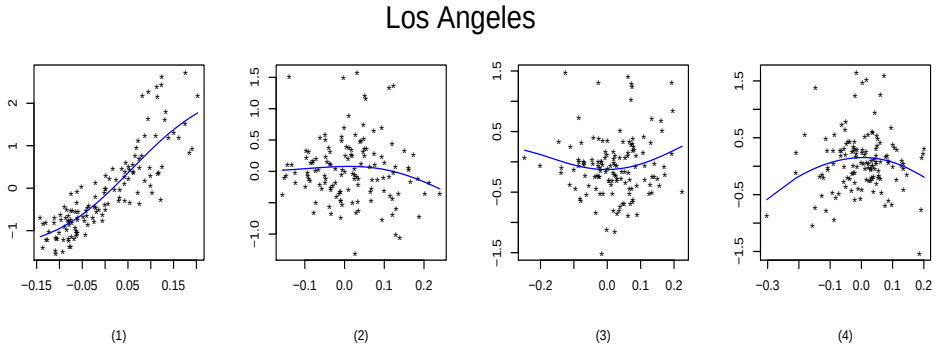

Figure 5.3 (1)-(4) are the scatter plots of the estimated partial residuals for the j[th] EDR direction $\hat{r}_j$ against the j[th] EDR direction $\hat{\boldsymbol{\beta}}_j^T \boldsymbol{X}$, j = 1, …, 4, respectively. The superimposed lines are the smoothed estimated additive terms $\hat{g}_j(\hat{\boldsymbol{\beta}}_j^T \boldsymbol{X})$, j = 1, …,4, to help visualization.

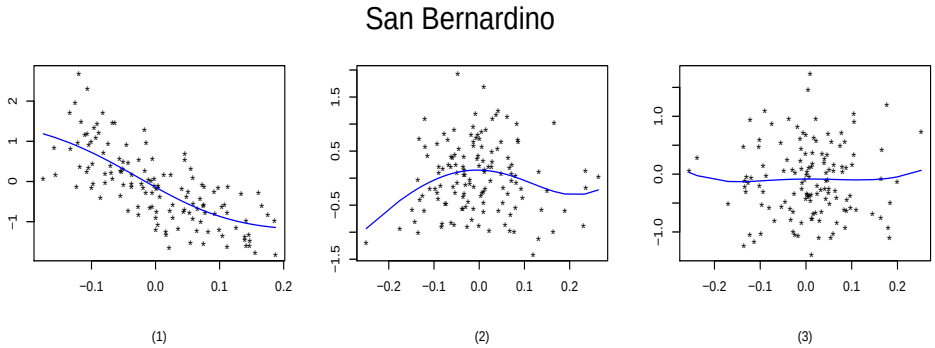

Figure 5.4 (1)-(3) are the scatter plots of the estimated partial residuals for the jth EDR direction $\hat{r}_j$ against the jth EDR direction $\hat{\boldsymbol{\beta}}_j^T \boldsymbol{X}$, j = 1, 2 or 3, respectively. The superimposed lines are the smoothed estimated additive terms $\hat{g}_j(\hat{\boldsymbol{\beta}}_j^T \boldsymbol{X})$, j = 1, 2 or 3, to help visualization.

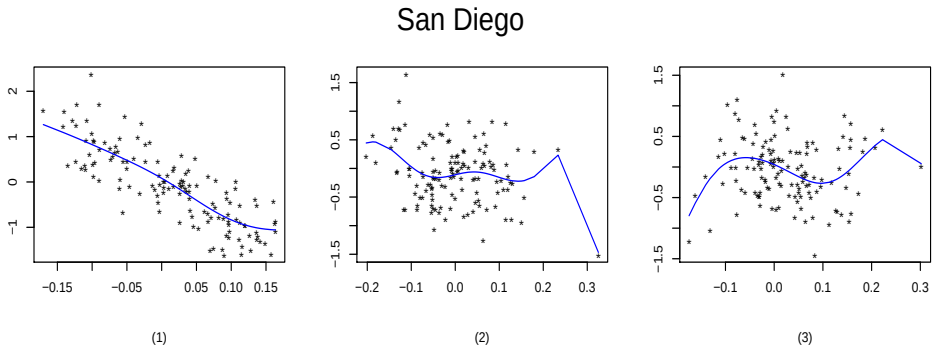

Figure 5.5 (1)-(3) are the scatter plots of the estimated partial residuals for the jth EDR direction $\hat{r}_j$ against the jth EDR direction $\hat{\boldsymbol{\beta}}_j^T \boldsymbol{X}$, j = 1, 2 or 3, respectively. The superimposed lines are the smoothed estimated additive terms $\hat{g}_j(\hat{\boldsymbol{\beta}}_j^T \boldsymbol{X})$, j = 1, 2 or 3, to help visualization.

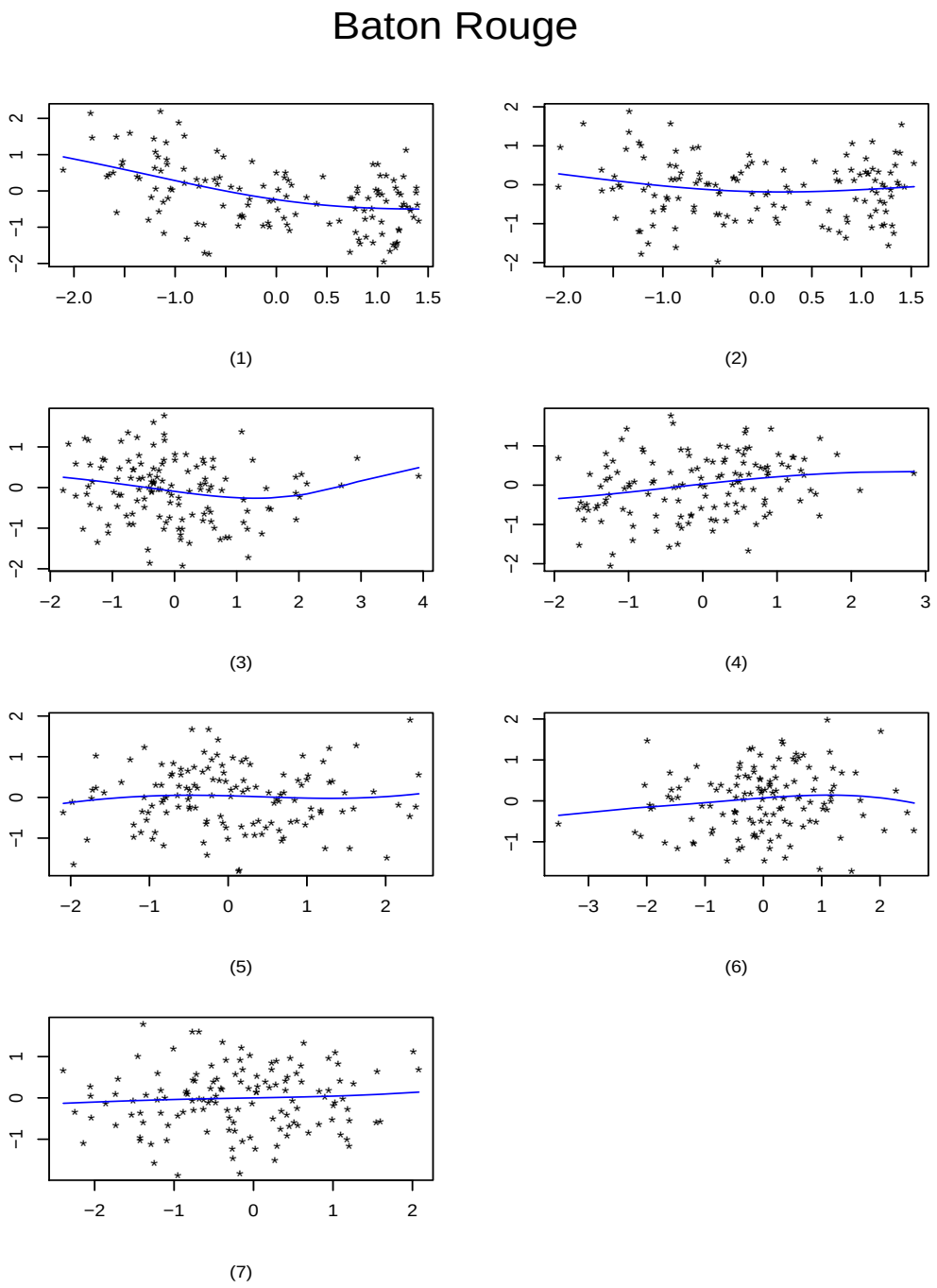

Figure 5.6 (1)-(7) are the scatter plots of the estimated partial residuals for temp, humi, PM_{10}, O_3, SO_2, NO_2 and CO respectively. The superimposed lines are the corresponding smoothed estimated additive terms to help visualization.

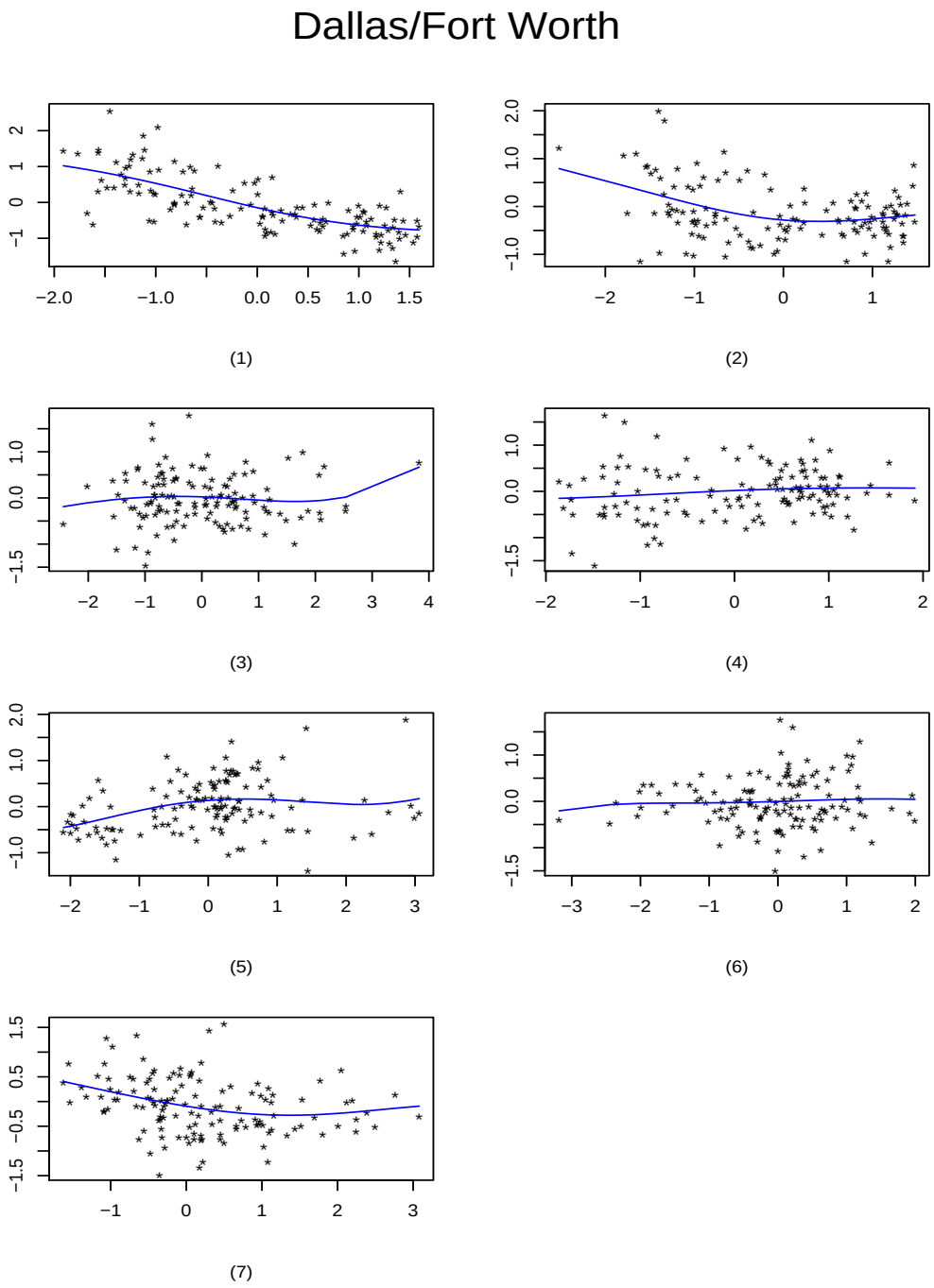

Figure 5.7 (1)-(7) are the scatter plots of the estimated partial residuals for temp, humi, PM_{10}, O_3, SO_2, NO_2 and CO respectively. The superimposed lines are the corresponding smoothed estimated additive terms to help visualization.

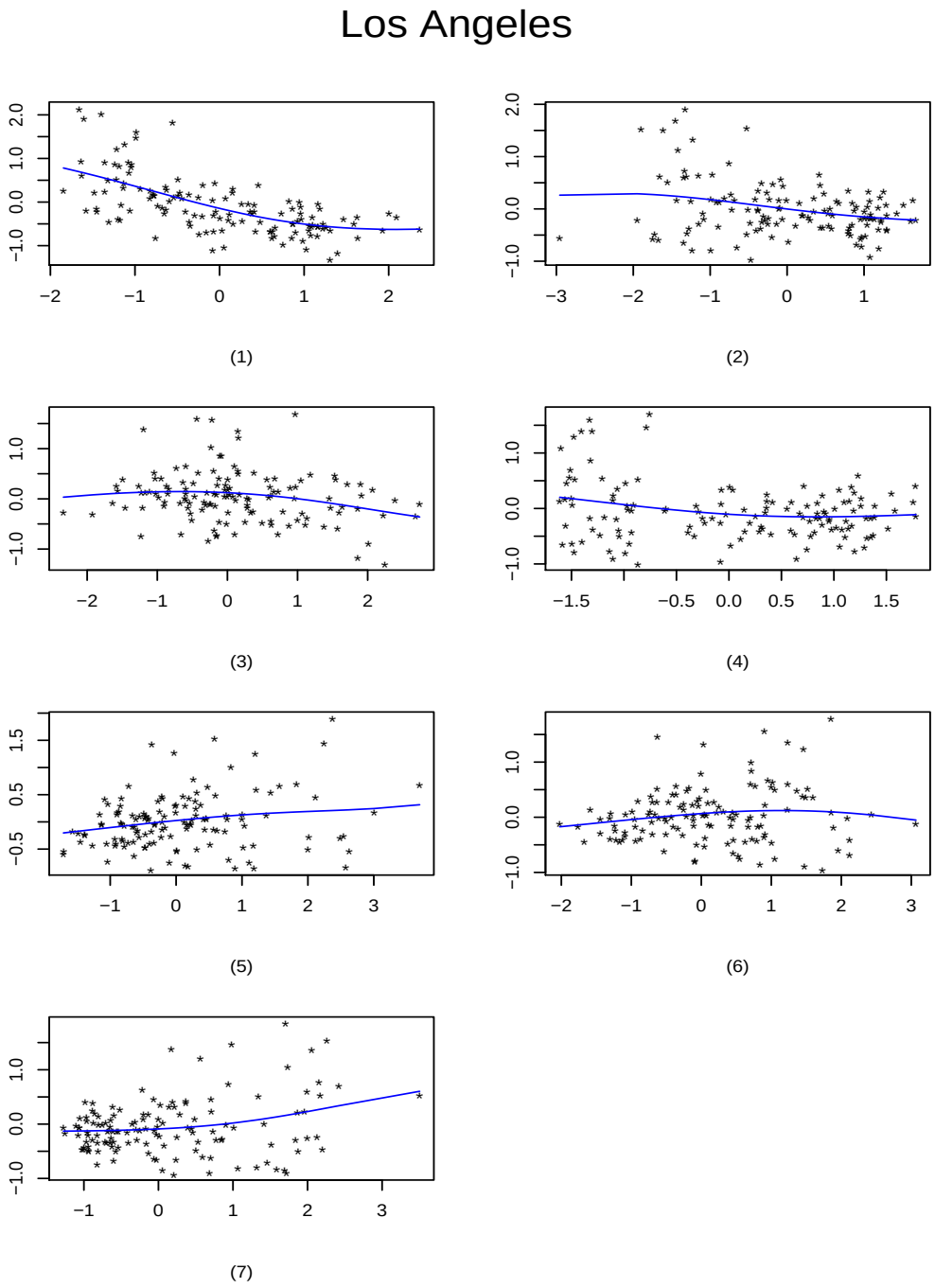

Figure 5.8 (1)-(7) are the scatter plots of the estimated partial residuals for temp, humi, PM_{10}, O_3, SO_2, NO_2 and CO respectively. The superimposed lines are the corresponding smoothed estimated additive terms to help visualization.

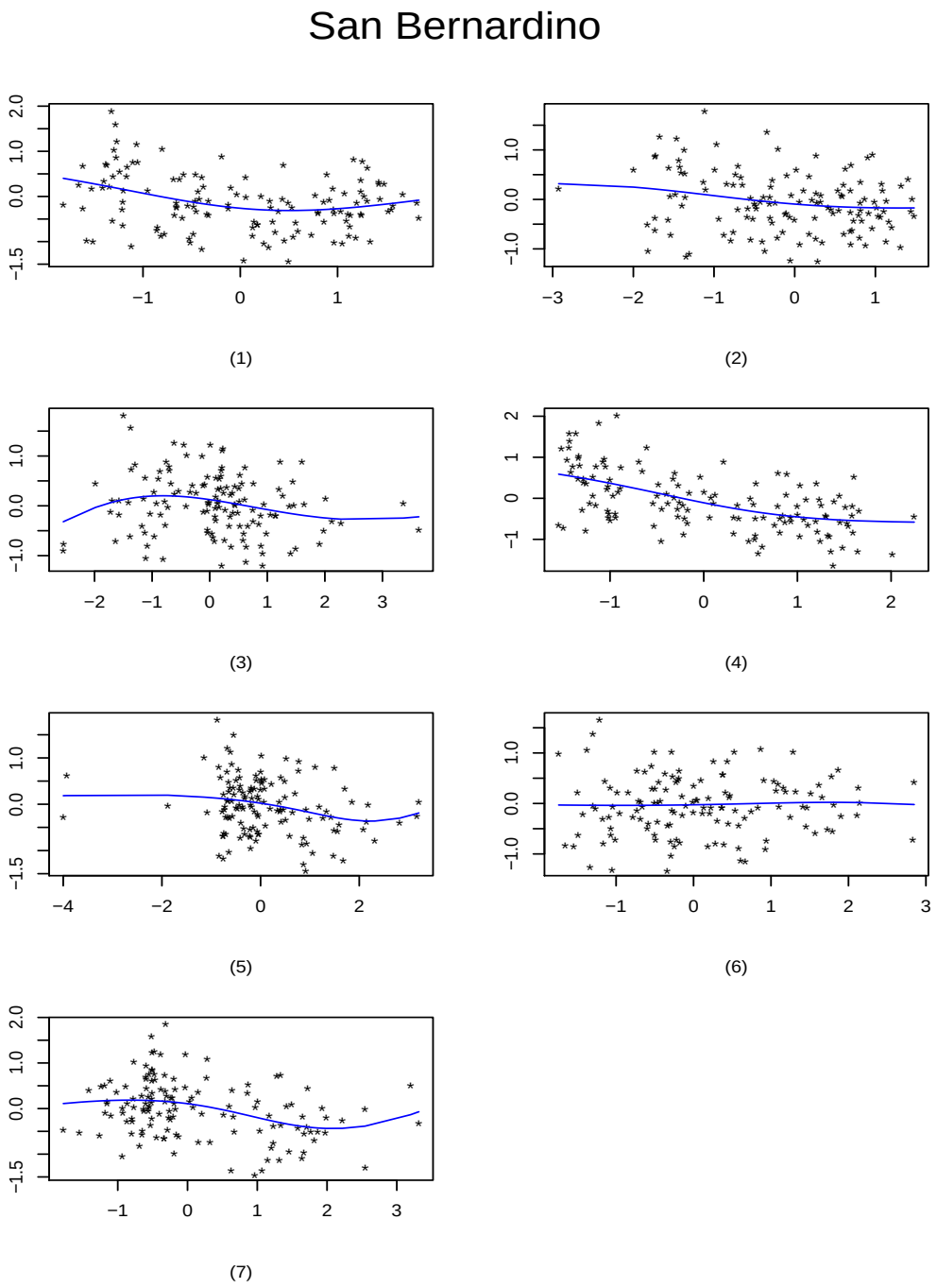

Figure 5.9 (1)-(7) are the scatter plots of the estimated partial residuals for temp, humi, PM_{10}, O_3, SO_2, NO_2 and CO respectively. The superimposed lines are the corresponding smoothed estimated additive terms to help visualization.

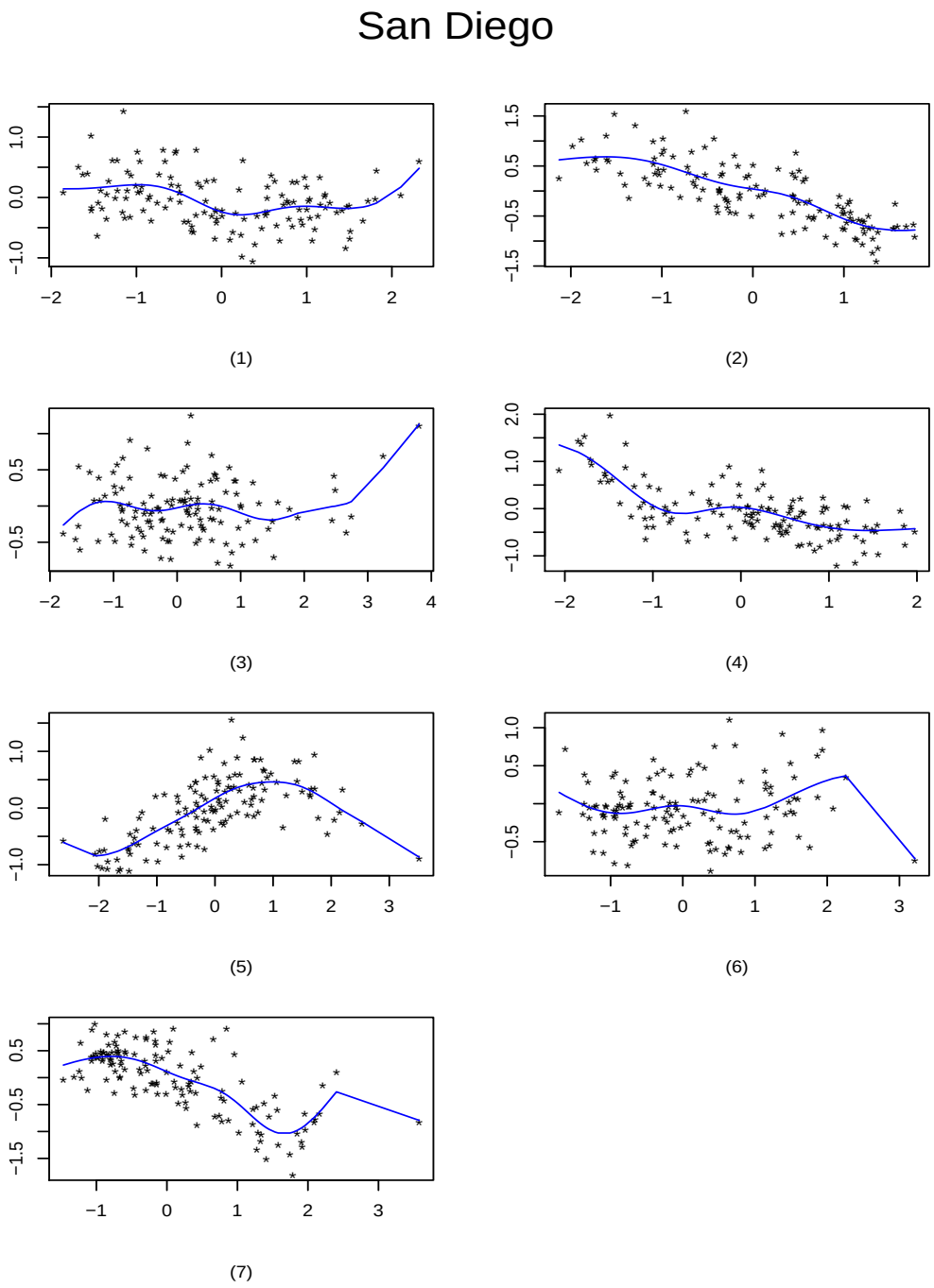

Figure 5.10 (1)-(7) are the scatter plots of the estimated partial residuals for temp, humi, PM_{10}, O_3, SO_2, NO_2 and CO respectively. The superimposed lines are the corresponding smoothed estimated additive terms to help visualization.

Chapter 6

Concluding Remarks

We have analyzed the effects of exposure to air pollution on public health across 15 populous cities in the United States. The method used in this analysis is mainly a two-stages nonparametric approach. The first step is to perform the rMAVE method proposed by Xia *et al*. (2002) to efficiently reduce the dimension in multiple regression problems. The second step refers to choose a proper model according to cross-validatory values which assess a model's forecasting ability. We focus on the Generalized Additive Models (GAMs) because of their prevalence in epidemiological studies on air pollution. As the alternative, we consider the general nonparametric multiple regression models because of their flexibility in fitting.

By applying the method to the practical air pollution data set extracted from the

NMMAPS database, our results confirm that air pollutants (PM_{10}, O_3, SO_2, NO_2 and CO) at current levels, acting with weather conditions (temperature and humidity) together, have adverse effects on human health. The results also indicate that the influential hazards to death are O_3, PM_{10}, and weather variates. As for model selection, our results suggest that dimension reduction through the rMAVE method is necessary to the original data, and that the general multiple model incorporating Efficient Dimension Reduction outperforms GAMs for all the observed city-specific data. Moreover, the model derived from our results emphasizes the existence of complex interactions among pollutants and weather conditions. Our conclusion is that the GAM is not a proper model in describing the effects of air pollution on public health although it is strongly recommended by epidemiologists.

Nevertheless, our analysis has some limitations. For exammple, from the rMAVE method we only have the estimates but can not assess the performance and significance of those estimates, since the asymptotic distribution has not been obtained. In spite of these limitations, our results still represent a starting point for refinement in the future analysis of the effects of air pollution on public health. It would seem appropriate then to investigate how to adjust the EDR space for the proper usage of GAMs to gain a better forecasting performance and a deeper understanding of the link between air pollution and health effects. For example, upon obtaining the EDR estimate $\hat{B}_0 = (\hat{\beta}_1, \ldots, \hat{\beta}_D)$, we may search for an orthogonal transformation on $\hat{B}_0$, namely $\Theta_0 = (\theta_1, \ldots, \theta_D) = \hat{B}_0\Gamma$

where Γ is subject to $\Gamma^T\Gamma = I_D$, such that the GAM after this orthogonal transformation

$$Y = g_1(\theta_1^T X_1) + \ldots + g_D(\theta_D^T X_D) + \varepsilon,$$

provides a better fit than the general EDR model (5.6).

Bibliography

Allen, D. M. (1974) The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**, 125–127.

Auestad, B. and Tjøstheim, D. (1990) Identification of nonlinear time series: first order characterisation and order determination. *Biometrika*, **77**, 669–688.

Cheng, B. and Tong, H. (1993) On residual sums of squares in non-parametric autoregression. *Stochastic Process with their Applications*, **48**, 154–174.

Cook, R. D. and Weisberg, S. (1999) *Applied Regression Including Computing and Graphics*. Wiley, New York.

Daniels, M. J., Dominici, F., Zeger, S. L. and Samet, J. M. (2004) The National Morbidity, Mortality, and Air Pollution Study. Part III: PM_{10} consentration-response curves and thresholds for the 20 largest US cites. *Research Report* 94, Health Effects Institute, Cambridge MA.

Denker, M. and Keller, G. (1983) On U-statistics and von Mises' statistics for weakly dependent processes. *Z. Wahrsch. Verw. Gebiete*, **64**, 505–522.

Dominici, F., McDermott, A. and Hastie, T. (2004) Improved Semi-parametric Time Series Models of Air Pollution and Mortality. *Journal of American Statistical Association*, **468**, 938–948.

Dominici, F., McDermott, A., Zeger, S. L. and Samet, J. M. (2002) On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology*, **156**, 193–203.

Dominici, F., Samet, J. M. and Zeger, S. L. (2000) Combining evidence on air pollution and daily mortality from the twenty largest US cities: a hierarchical modeling strategy (with discussion). *Journal of the Royal Statistical Society, Series A*, **163**, 263–302.

Hastie, T. J. and Tibshirani, R. J. (1986) Generalized additive models (with discussion). *Statistical Science*, **1**, 297–318.

Künzli, N., Medina, S., Kaiser, R., Quénel, P., Horak, F. Jr. and Studnicka, M. (2001) Assessment of deaths attributable to air pollution: should we use risk estimates based

on time series or on cohort studies? *American Journal of Epidemiology*, **153**, 1050–1055.

Lee, J. T., Kim, H., Hong, Y. C., Kwon, H. J., Schwartz, J. and Christiani, D. C. (2000) Air Pollution and Daily Mortality in Seven Major Cities of Korea, 1991-1997. *Environmental Research, Section A*, **84**, 247–254.

Lipfert, F. W. (1994) *Air pollution and community health*. New York, NY: Van Nostrand Reinhold.

McGeehin, M. A. and Mirabelli, M. (2001) The potential impacts of climate variability andchange on temperature-related morbidity and mortality in the United States. *Environmental Health Perspectives*, **109**, 185–189.

Pope III, C. A., Burnett, R. T., Thum, M. J., Calle, E. E., Krewski, D., Ito, K. and Thurston, G. D. (2002) Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Journal of American Medical Association*, **287**, 1132–1141.

Roussas, G. G. (1988) Nonparametric estimation in mixing sequences of random variables. *Journal of Statist. Plann. Inference*, **15**, 135–149.

Samet, J. M., Dominici, F., Curriero, F., Coursac, I., and Zeger, S. L. (2000) Particulate air pollution and motality: findings from 20 U.S. cities. *New England Journal of Medicine*, **343**, 1742–1757.

Schwartz, J., Spix, C., Touloumi, G., Bachárová, L., Barumamdzadeh, T., le Tetre, A., Piekarksi, T., Ponce, de Leon A., Pönkä, A., Rossi, G., Saez, M. and Schouten, J. P. (1996) Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *Journal of Epidemiological Community Health*, **50**, (Suppl 1), S3–S11.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, **36**, 111–147.

World Health Organization (2000) Quantification of the health effects of exposure to air pollution. *Report on a WHO working group*, Bilthoven, Netherlands, Nov. 2000.

World Health Organization (2002) Exposure assessment in studies on the chronic effects of long-term exposure to air pollution. *Report on a WHO working group*, Bonn, Germany, Feb. 2002.

World Health Organization (2003) Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide. *Report on a WHO working group*, Bonn, Germany, Jan. 2003.

Xia, Y. and Tong, H. (2005) Cumulative effects of air pollution on public health. *In press*.

Xia, Y., Tong, H., Li, W. K., and Zhu, L. (2002) An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society, Series B*, **64**, 363–410.

Appendix A

Conditions for Theorem 1

Let $\mathscr{F}_1^n(\boldsymbol{X})$ denote $\sigma(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)$, the σ-algebra generated by $(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)$.

(A.1) $\mathbb{E}(\varepsilon|\mathscr{F}_1^n(\boldsymbol{X}))=0$, almost surely.

(A.2) $\mathbb{E}(\varepsilon^2|\mathscr{F}_1^n(\boldsymbol{X}))=\sigma^2$, a strictly positive constant, almost surely.

(A.3) $K_D(\boldsymbol{u})=\prod_{i=1}^D k(u_i)$ for $\boldsymbol{u}=(u_1,\ldots,u_D)\in\mathbb{R}^D$.

(A.4) g is Hölder continuous, i. e. $\forall\ \boldsymbol{x}_1,\ \boldsymbol{x}_2\in\mathbb{R}^D,\ |g(\boldsymbol{x}_1)-g(\boldsymbol{x}_2)|\leq c_1\parallel\boldsymbol{x}_1-\boldsymbol{x}_2\parallel^{\mu}$, where $0<\mu\leq 1$ and $\parallel\ \ \parallel$ denotes the Euclidean norm in $\mathbb{R}^D$.

(A.5) ω is a weight function which has a compact support S and

$$0<\int_{\mathbb{R}^d}\omega(\boldsymbol{x})d\boldsymbol{x}<\infty,\quad 0\leq\omega(\boldsymbol{x})\leq 1.$$

(A.6) Let f denote the probability density function of $\boldsymbol{X}=(X_1,\ldots,X_D)$, which is strictly positive on S, and $\forall\ \boldsymbol{x}_1,\ \boldsymbol{x}_2\in\mathbb{R}^D,\ |f(\boldsymbol{x}_1)-f(\boldsymbol{x}_2)|\leq c_2\parallel\boldsymbol{x}_1-\boldsymbol{x}_2\parallel$.

(A.7) k has compact support, and $\forall\, x_1,\ x_2 \in \mathbb{R}^1,\ |k(x_1)-k(x_2)| \le c_3 \parallel x_1 - x_2 \parallel$.

(A.8) For every $t,s,\tau,t',s'\tau' \in \mathbb{N}$, the joint probability density function of $(\boldsymbol{X}_t, \boldsymbol{X}_s, \boldsymbol{X}_\tau, \boldsymbol{X}_{t'}, \boldsymbol{X}_{s'}, \boldsymbol{X}_{\tau'})$ is bounded.

(A.9) Let $\frac{1}{p}+\frac{1}{q}=1$. For some $p>2$ and $\delta>0$ such that $\delta < \frac{2}{q}-1$, $\mathbb{E}|\varepsilon|^{2p(1+\delta)} < \infty$ and $\mathbb{E}|g(\boldsymbol{X}_1)|^{2p(1+\delta)} < \infty$.

(A.10) For δ in condition (i) and some $\varepsilon > 0$, $\beta_j^{\delta/(1+\delta)} = o(j^{-2+\varepsilon})$, where

$$\beta_j = \sup_{i\in\mathbb{N}} \left(\mathbb{E}\left[\sup_{A\in\mathscr{F}_{i+j}^n(\boldsymbol{X})} \{|Pr(A|\mathscr{F}_1^n(\boldsymbol{X})) - Pr(A)|\} \right] \right).$$

(A.11) Let $j=j(n)$ be a positive integer and $i=i(n)$ be the largest positive integer such that $2ij \le n$,

$$\limsup_{n\to\infty} \left(1 + 6e^{1/2}\beta_j^{1/(1+i)}\right)^i < \infty.$$

(A.12) For $i=i(n)$ in condition (k) and the bandwidth h,

$$\limsup_{n\to\infty} \{i(n)h^D\} < \infty.$$

(A.13) $nh^{2D} \to \infty$ as $n \to \infty$.

(A.14) For μ in assumption (d), $nh^{2D+2\mu} \to 0$ as $n \to \infty$.

(A.15) For q, δ and ε in condition (i) and (j) $n^\varepsilon h^{-2D+\theta} \to 0$ as $n \to \infty$, where $\theta = 4D/(q+q\delta)$.

We briefly describe here some explanation of these conditions in order, which were given in Cheng and Tong (1993). Conditions (A.1)-(A.4) are self-explanatory. Condition(A.5) is the introduction of a weight function ω, the purpose of which is to overcome

the "infinite integration problem" in asymptotic expansion encountered by Auestad and Tjøstheim (1990). Conditions (A.6), (A.7), (A.9), (A.13) and (A.14) are standard conditions in nonparametric inference. Condition (A.8) is a mild condition, which will be useful when we use a mixing inequality. Condition (A.10) is a very mild condition, which is weaker than geometric absolute regularity. Conditions (A.11) and (A.12) were given by Roussas (1988). They may be replaced by other assumptions on the mixing coefficient β, if other methods are used to show the almost sure convergence of $\hat{f}_n$ and $\hat{g}_n$. Condition (A.15) is necessary for proposition 2 of Denker and Keller (1983). Note that conditions (A.10) and (A.15) do not contradict each other.

Appendix B

Time-Series Plots

In this appendix, we include the time-series plots for the monthly averages of death rate, temperature, dew point temperature (i. e. humidity), PM_{10}, O_3, SO_2, NO_2 and CO from January 1987 to December 1998, in Baton Rouge, Dallas/Fort Worth, Los Angeles, San Bernardino, and San Diego respectively. We select them as representatives of all the 15 study cites.

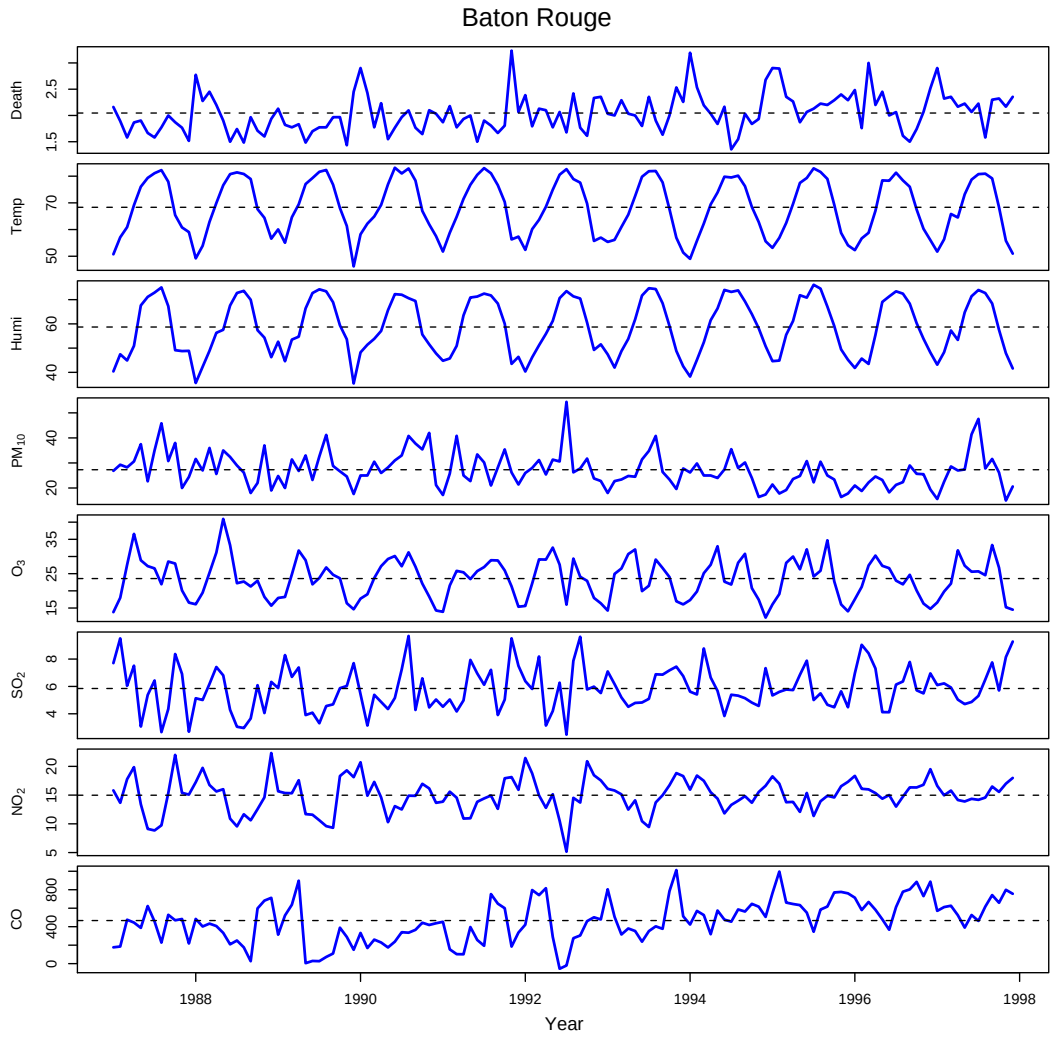

Figure B.1 Time-series plots for Baton Rouge.

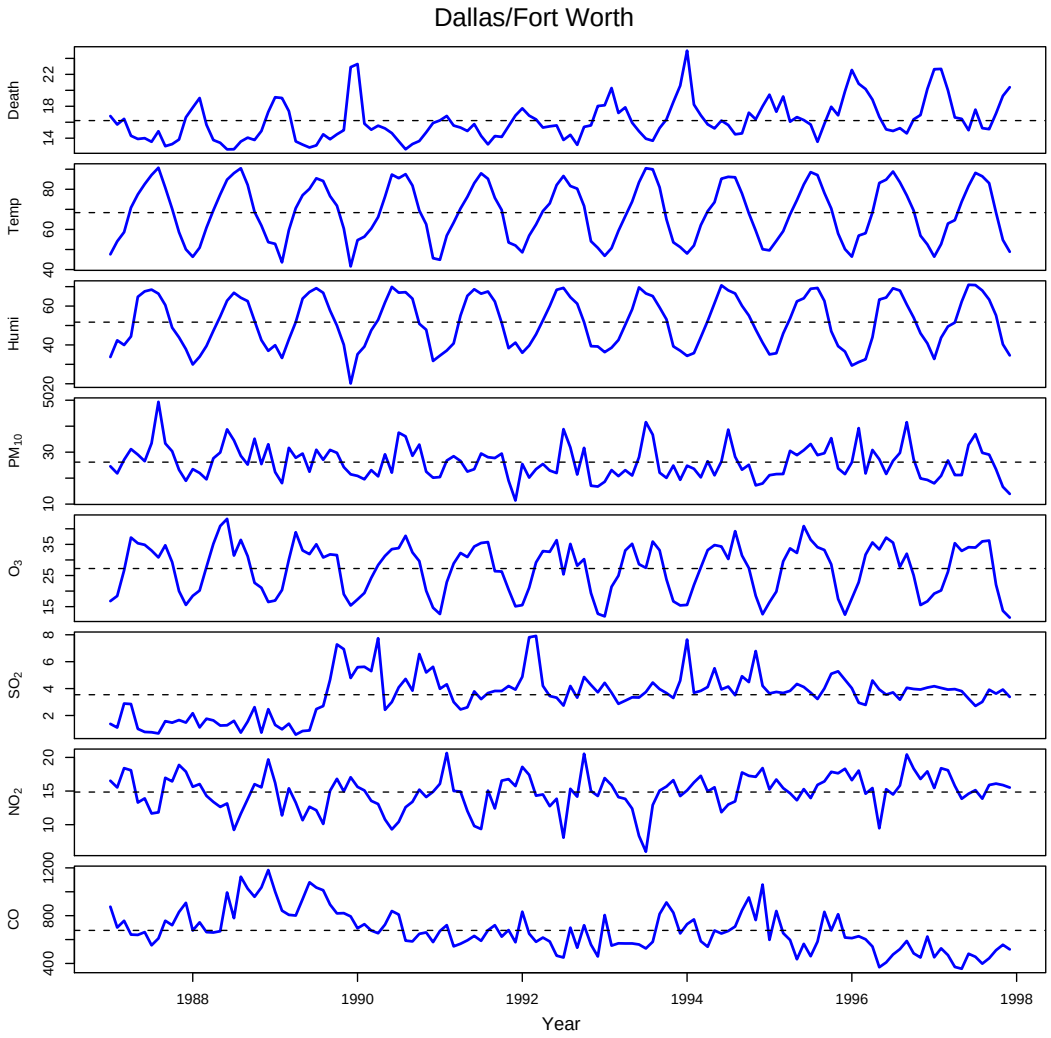

Figure B.2 Time-series plots for Dallas/Fort Worth.

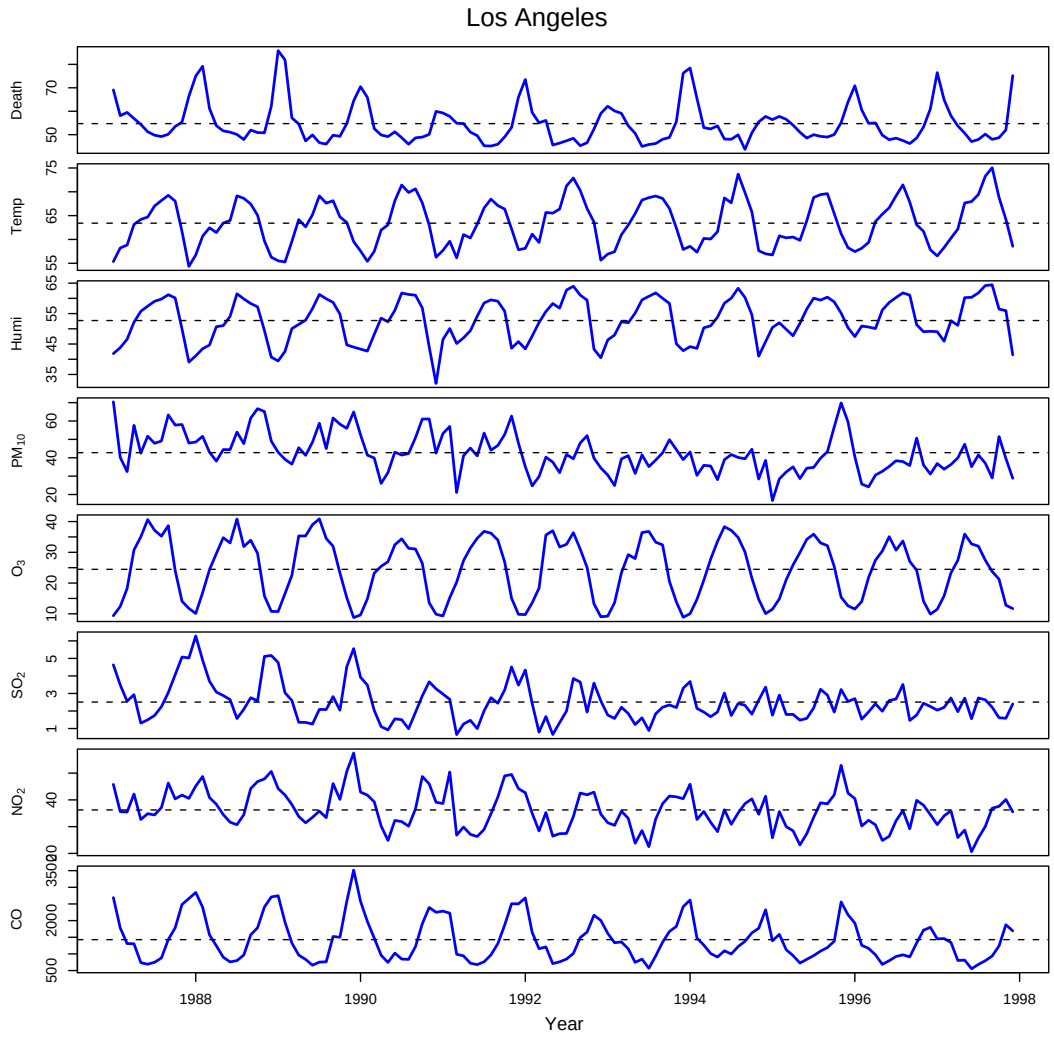

Figure B.3 Time-series plots for Los Angeles.

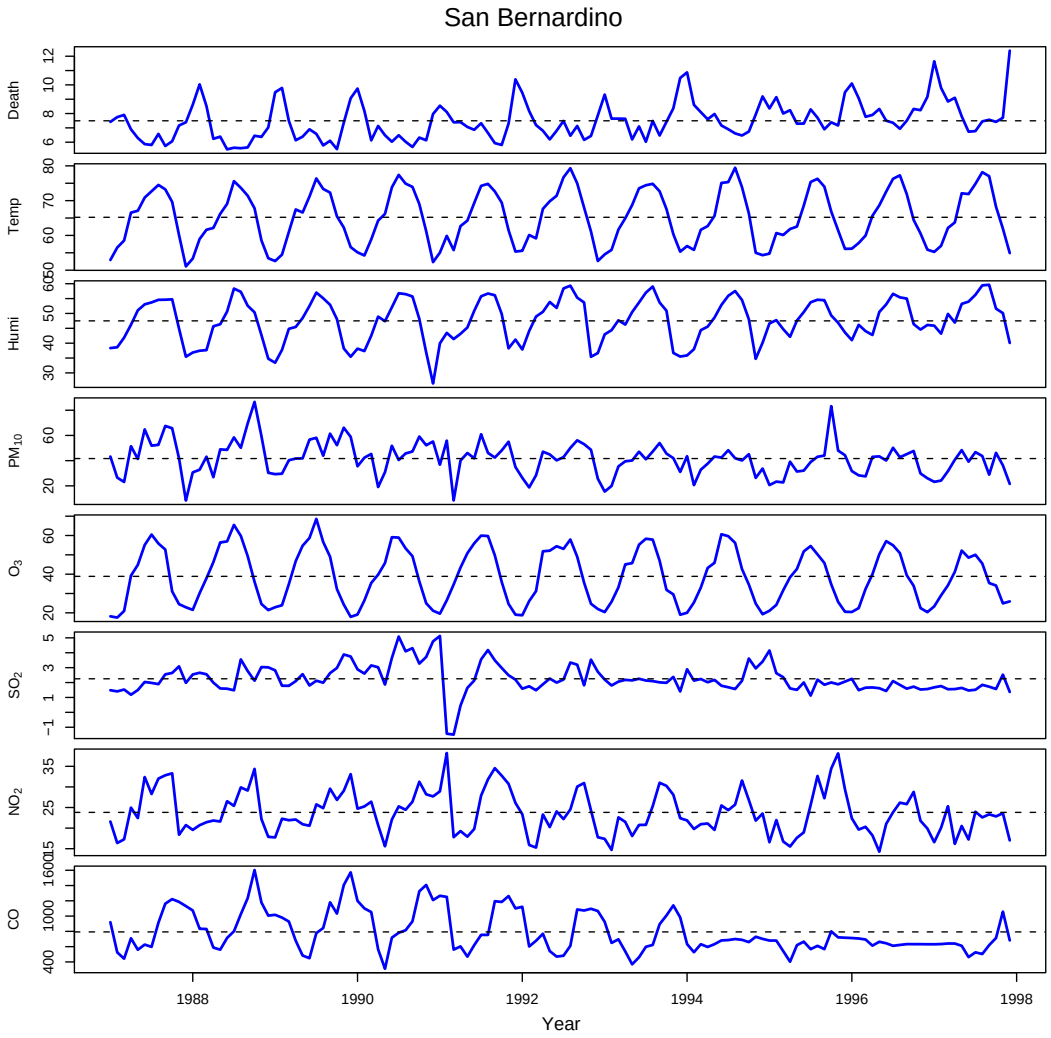

Figure B.4 Time-series plots for San Bernardino.

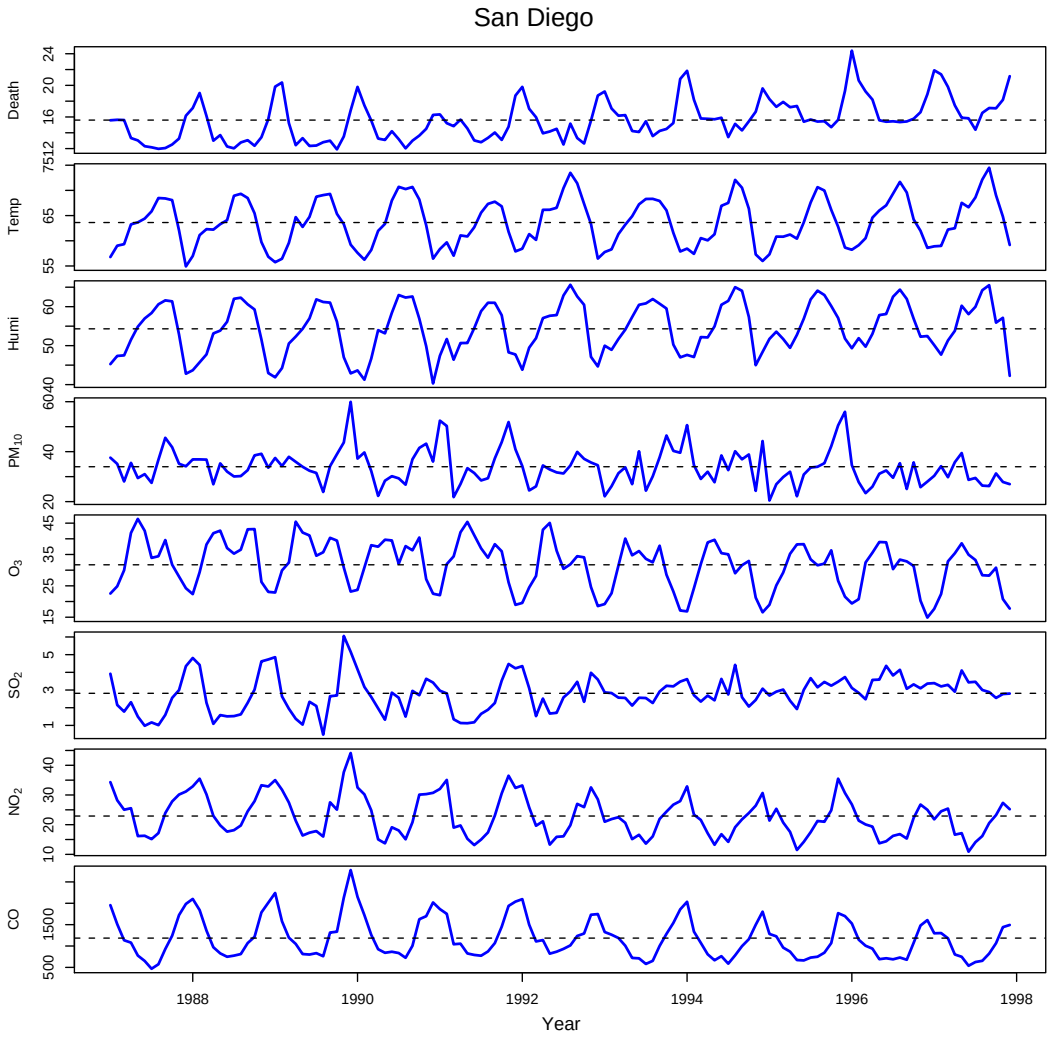

Figure B.5 Time-series plots for San Diego.

Appendix C

Scatter Plot Matrix with Correlations

In this appendix, we include the scatter plot matrix with correlations among the monthly averages of death count, temperature, dew point temperature (i. e. humidity), PM_{10}, O_3, SO_2, NO_2 and CO from January 1987 to December 1998, in Baton Rouge, Dallas/Fort Worth, Los Angeles, San Bernardino, and San Diego respectively. We select them as representatives of all the 15 study cities.

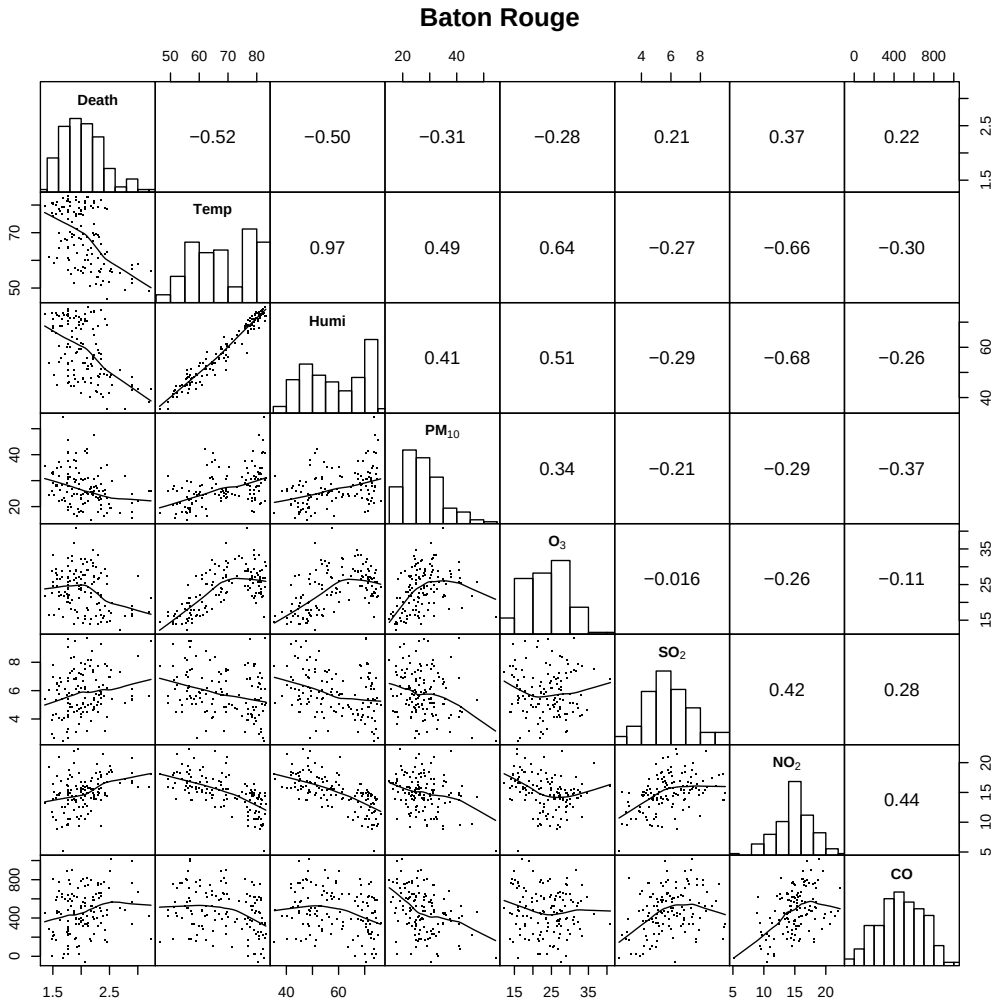

Figure C.1 Scatter plot matrix with correlations for Baton Rouge.

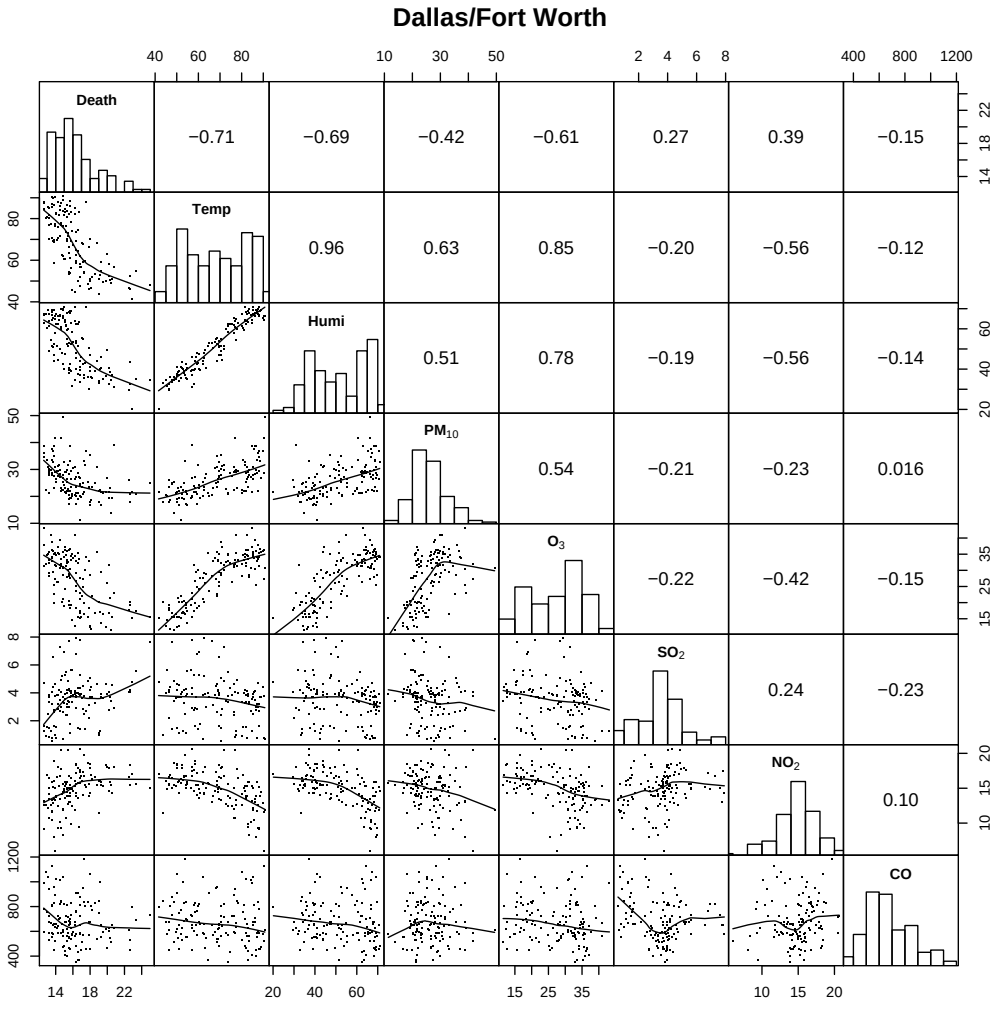

Figure C.2 Scatter plot matrix with correlations for Dallas/Fort Worth.

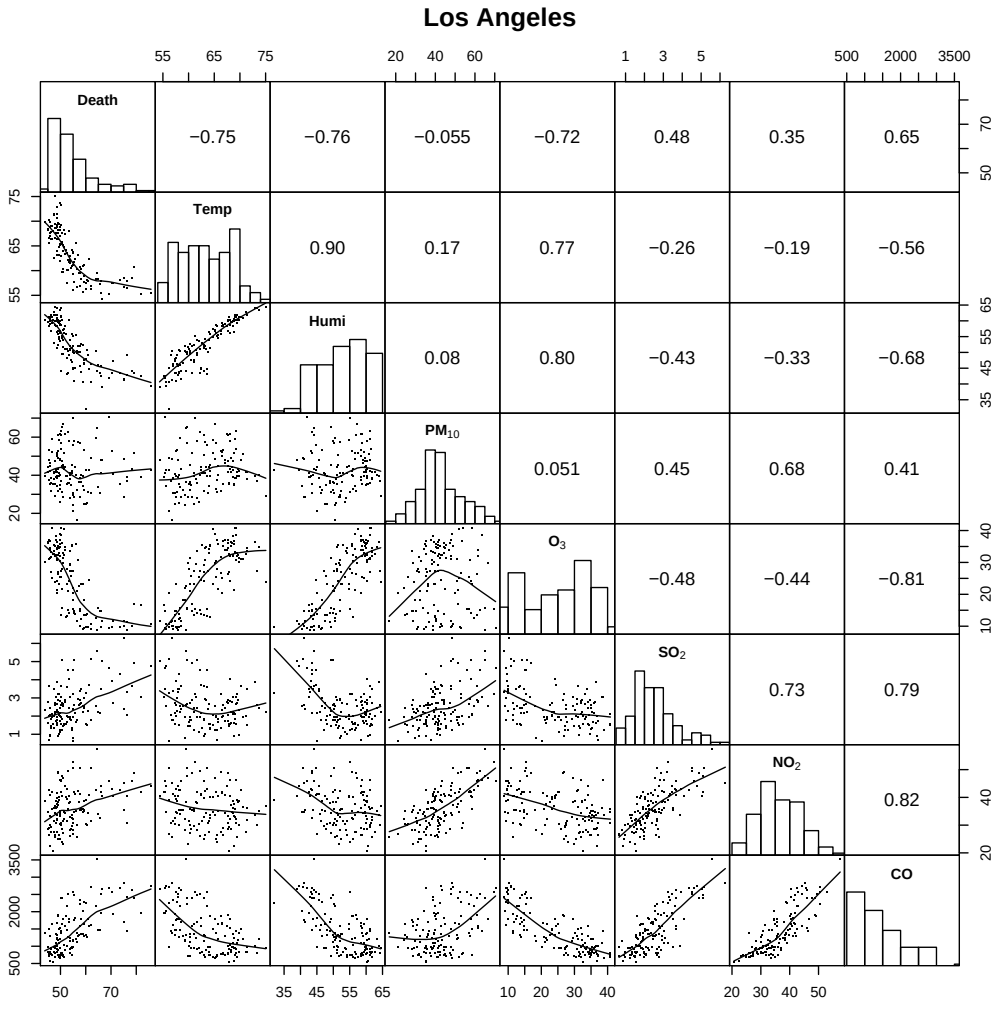

Figure C.3 Scatter plot matrix with correlations for Los Angeles.

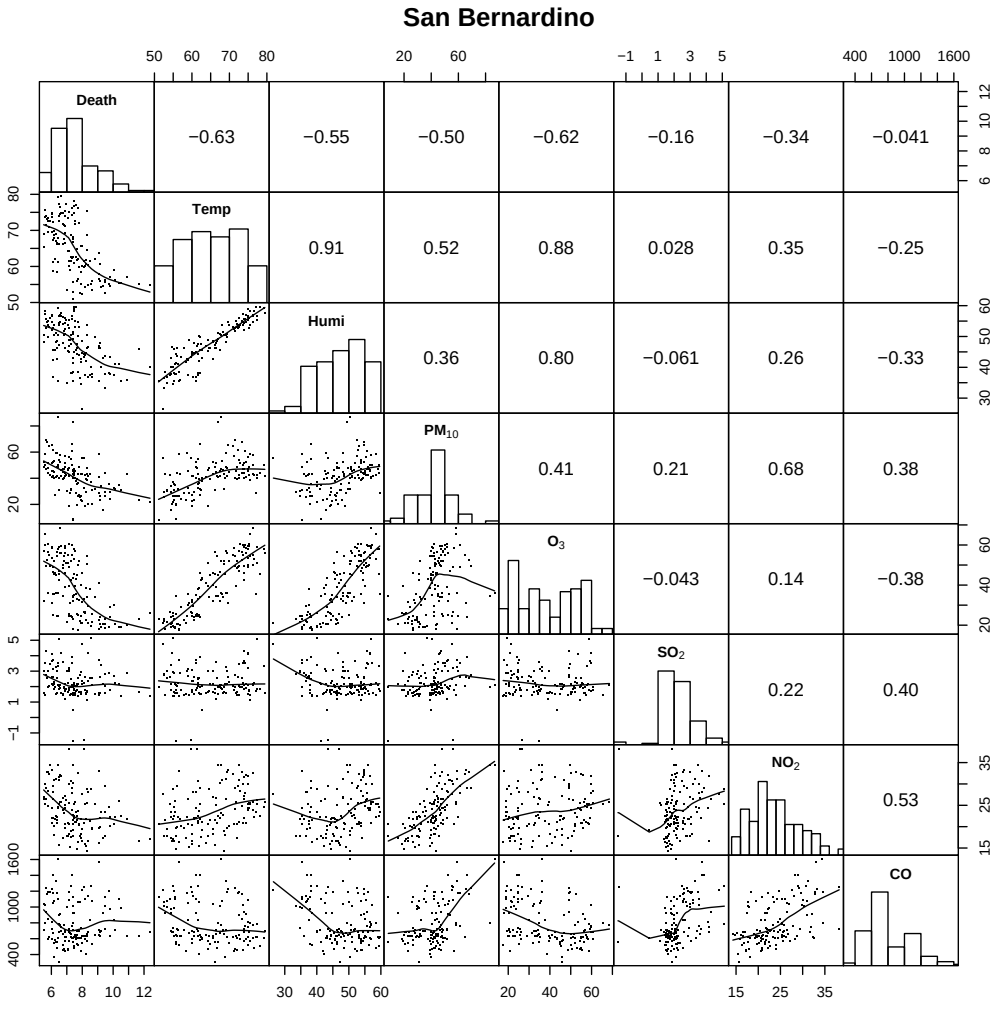

Figure C.4 Scatter plot matrix with correlations for San Bernardino.

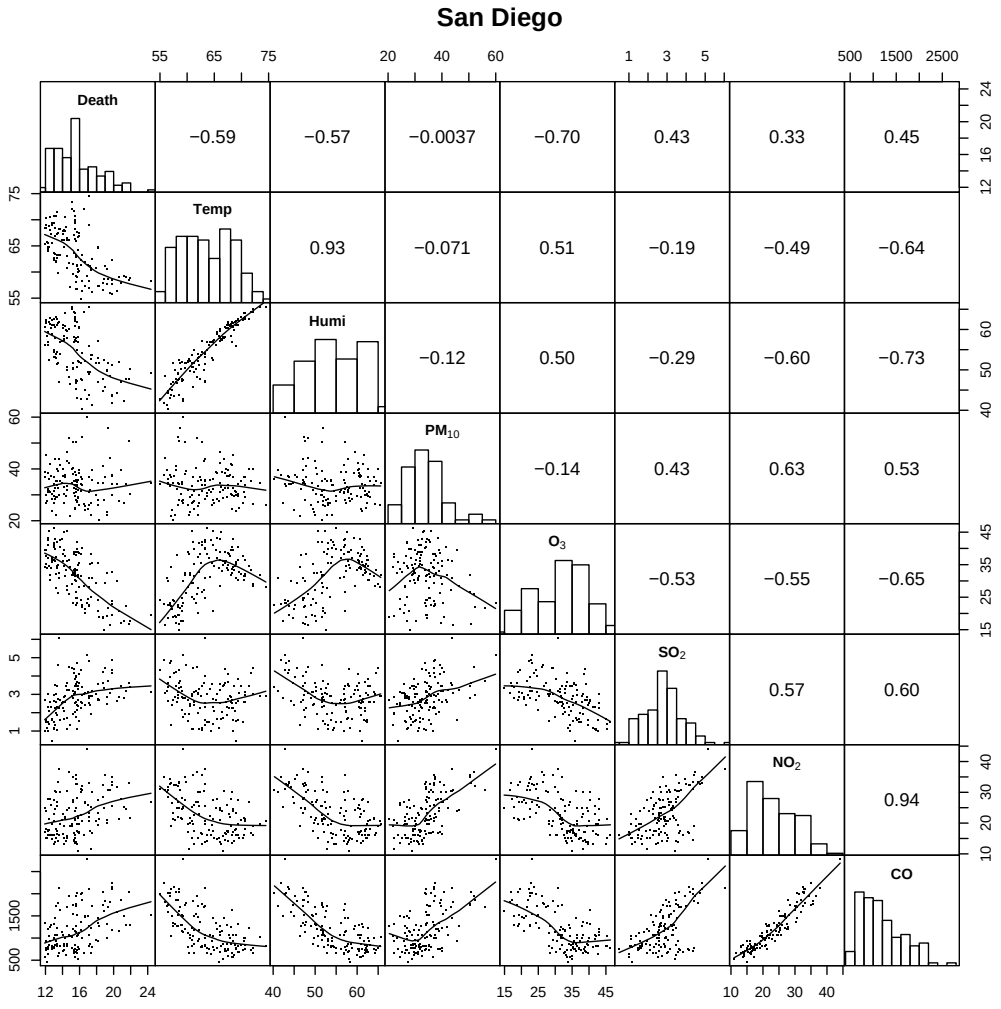

Figure C.5 Scatter plot matrix with correlations for San Diego.

NONPARAMETRIC MODELING OF THE EFFECTS OF AIR POLLUTION ON PUBLIC HEALTH PENG QIAO 2005