Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXVII

Functional Data Analysis for Environmental Pollutants and Health

Direttore della Scuola: Prof. Monica Chiogna

Supervisore: Prof. Francesco Pauli

Dottorando: Maeregu Arisido

January 29, 2015

To my father,

Woldeyes Arisido

Abstract

The adverse health effect of exposure to high pollutant concentration has been the focus of many recent studies. This is particularly true for ground level ozone which is considered in the present thesis. The effect has been estimated at different geographic locations, and it has been shown that it may be spatially heterogeneous. Within such widely accepted studies, two major issues arise which are the focus of this thesis: how to best measure daily individual exposure to a pollutant and how the health effect of the exposure is affected by geographic location both in strength and shape. The first issue is related to the fact that the concentration of ozone varies widely during the day, producing a distinctive daily pattern. Traditionally, the daily pattern of the pollutant is collapsed to a single summary figure which is then taken to represent daily individual exposure. In this thesis, we propose a more accurate approaches to measure pollutant exposure which address the limitations in the use of the standard exposure measure. The methods are based on principle of functional data analysis, which treats the daily pattern of concentration as a function to account for temporal variation of the pollutant. The predictive efficiency of our approach is superior to that of models based on the standard exposure measures. We propose a functional hierarchical approach to model data which are coming from multiple geographic locations, and estimate pollutant exposure effect allowing daily variation and spatial heterogeneity of the effect at once. The approach is general and can also be considered as the analogue of the multilevel models to the case in which the predictor is functional and the response is scalar.

Sommario

Numerosi studi recenti hanno mostrato l'effetto dannoso che l'esposizione a elevate concentrazioni di inquinanti ha sulla salute umana. In particolare, questo avviene per l'ozono, del quale ci occupiamo nel presente lavoro. Stime ottenute in diversi siti mostrano che l'effetto è geograficamente eterogeneo. Nel contesto degli studi menzionati emergono due aspetti di particolare importanza, e su cui è incentrato il presente lavoro: come misurare al meglio l'esposizione individuale e come e in che misura l'effetto vari geograficamente, sia quanto a intensità che a forma. La prima questione è legata al fatto che la concentrazione di ozono mostra ampie variazioni nel corso di una giornata. Di tale andamento giornaliero non si tiene conto nella maggior parte degli studi epidemiologici, e si assume che possa essere efficacemente riassunto da una statistica unidimensionale. Nel presente lavoro proponiamo degli approcci che si basano sull'impiego di misure della concentrazione che tengono conto dell'andamento temporale della stessa. Tali approcci sono basati sulla metodologia dell'analisi dei dati funzionali, che consiste nel trattare il dato sulla concentrazione giornaliera come una funzione, tenendo così conto delle sue variazioni durante la giornata. In termini previsivi, si è verificato che tale approccio porta a un miglioramento rispetto ai modelli basati su una statistica giornaliera. Questo approccio è poi esteso al caso di dati multisito, per i quali si propone un modello funzionale gerarchico, che consentono di stimare l'effetto dell'esposizione all'inquinante tenendo conto da un lato della variazione giornaliera della concentrazione dello stesso e dell'eterogeneità nello spazio di tale effetto. Questo approccio può essere visto come l'analogo di un modello multilivello per il caso in cui il predittore è funzionale e la variabile risposta scalare.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Francesco Pauli, his guidance and support enabled to accomplish the work presented in this thesis. His invaluable suggestions and encouragement made the entire research period so much enjoyable. I am grateful to University of Padova for financial support. I am also grateful to professors at the department of statistics who gave me enormous knowledge and skills in the PhD course program. particularly, I would like to thank Monica Chiogna for crucial suggestion in identifying research questions and reading of the papers. I would also wish to thank cycle 27 PhD students: Marco, Vera, Gloria, Gianluca, Leonardo, Valentina and Paola. You have been a wonderful colleagues, I learnt so much from you over the whole three years.

My special thanks first go to my mother, she has always been with me in good times or bad, showing unlimited love and support in all aspects. I am grateful to my wife, Arsema Desalegn, her companionship has been an inspiration and strength in all my endeavour. I would like to express my thanks to my unborn daughter, she makes me so proud. Finally, I wish to thank my brother and sisters for their encouragement and support over these years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The adverse health effect of exposure to pollutants has become a global issue since early twentieth century when a series of severe air pollution episodes occurred in different areas of the world. Some of the episodes that caused acute respiratory issues were the 1930 Meuse Valley fog in Belgium (Stern, 1973), the air pollution crisis in some cities of the USA in 1930s (Dewey, 2000) and the Poza Rica episode, Mexico in 1950 (Stern, 1973). More severe health episode including deaths were caused by the London 'Great Smog' in 1952 (Neidell, 2009) and the Donora killer smog in 1948. These events led to formulation of strategies to reduce the environmental pollution levels. The USA and many countries in Europe adopted national air pollution legislations. For example, the UK had passed the Clean Air Act in 1956 following the severe London smog episode, and the United States federal government enacted a similar legislation in 1970. These Clean Air Acts have been reformulated after some years to drastically improve the air quality (US Environmental Protection Agency, 2007). Nevertheless, pollution persists at high levels, and studies continued to detect effect on human health from exposure to pollutants.

Environmental studies of exposure to pollutants and health use data which consist of the concentration measure of pollutants, health outcome data and various confounding covariates for a particular study region. The health outcome data contains daily counts of mortality or morbidity (hospital admissions) for the population residing within the study area. The commonly studied pollutants for their effect on health are Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), ground level Ozone (O3), particulate matter (PM) and Sulphur Dioxide (SO_2), with ground level ozone being the main pollutant studied in this thesis. The association between exposure to ozone and health is widely studied. A search on PubMed with the search key ”ozone epidemiology” leads to 1322 papers, the oldest being from 1974. An analogous search on Google Scholar leads to 35500 results, while a search on PubMed with search key ”ozone mortality” leads to 732 and ”ozone morbidity” to 72900 results. Research moves toward various directions, among these the connections with climate change (De Sario et al., 2013), the study of the effect of ozone on specific pathologies such as asthma (Sousa et al., 2013), allergies, heart diseases (Shah et al., 2013), birth weight (Stieb et al., 2012), lung function decrements (Hazucha et al., 1989; Mudway and Kelly, 2000). A thorough review would clearly be out of scope here, we limit our focus on important aspects of the studies.

One aspect of these studies is the time scale over which adverse effect is most apparent. Some studies consider the short-term effect of exposure to pollutants estimated over a few days or weeks rather than a long-term exposure effect estimated by following cohorts over years to decades. Examples of air pollution effect from long-term exposure studies involving cohorts include Dockery et al. (1993), Violato et al. (2009) and Pope III and Dockery (2006). The majority of the studies including this thesis examine the short-term effect of pollutants on health. For a thorough review of short-term effect of exposure to pollutants see (Samoli et al., 2001) and (Dominici et al., 2003).

The short-term exposure studies use measurements of pollutants concentration obtained from a network of monitoring stations located throughout the study region. The concentrations levels of the pollutants are measured frequently throughout the day. Although more frequent measures can also be noticed, usually the concentrations are measured on hourly basis. To represent a daily exposure to a pollutant, a single summary figure derived from the hourly records is used, such as the daily average or maximum. Common statistical approaches used to analyse the data are the generalized linear models (GLM, McCullagh and Nelder (1989)) or the generalized additive models (GAM, Hastie and Tibshirani (1990)). These methods are used to estimate effects associated to exposure to pollutants by regressing day-varying health outcome against day-varying exposure measure, typically the daily summary figures, accounting for other confounding covariates. Covariates studied for their confounding effect are weather condition, particularly temperature and less frequently humidity, days of the week and calendar year. Some studies consider the confounding effect of concentrations of other pollutants (see for example, Chiogna and Pauli (2011)).

In this thesis, we shall propose approaches to measure exposure to pollutants to address the limitations of the standard exposure measures and improve effect estimation. The standard methods estimate the health effect of a pollutant by collapsing the hourly measurements into single daily summary figures. We employ a functional data analysis (FDA) technique to turn all hourly discrete records of a day into one smooth function with little information loss. Thus, the daily exposure to a pollutant is represented by a single curve which takes into account the daily variation of the concentration. Such measure will be used as a predictor in a functional regression model, within such model, the health effect is given by a functional coefficient.

The health effect of exposure has been studied at different geographic locations (Gryparis et al., 2004; Zhang et al., 2006), and it has been shown that the estimated effect is spatially heterogeneous (Bell et al., 2004). To address the heterogeneity issue, large multi-city studies have been implemented. In this regard, the European approach (APHEA project, for example Katsouyanni et al. (1996)) and the National Morbidity, Mortality and Air Pollution Study (NMMAPS, for example Samet et al. (2000)) in the USA are well known. The first approach uses the standard Poisson regression methods and the latter adopts a hierarchical model to combine evidence from single city analyses. Other researchers advocate meta-analysis techniques to reduce bias from estimates of single (city) studies (see for example Ji et al. (2011)). However, like other single city studies, the multi-city and the meta-analysis studies collapse the daily pattern of pollutant concentration to represent the daily exposure to the pollutant by a single figure summaries for each day and city.

We propose a functional hierarchical modeling approach to estimate pollutant exposure effect allowing for the daily variation of the concentration and spatial heterogeneity of the effect at once. Using this approach, we estimate an overall functional regression coefficient as well as location-specific coefficients in the Bayesian paradigm using Markov Chain Monte Carlo techniques. We shall also exploit the idea of functional principal component analysis (FPCA) to derive principal scores from hourly measurements of ozone, which are believed to capture the most important portion of the daily concentration curve. These principal scores will form an additional exposure measures for researchers to be used as alternative to our functional exposure measure and other approaches in literature. The reminder of this thesis is organized in 6 Chapters and the reminder of this Section discusses each Chapter in detail.

Chapter 2 reviews the statistical methods which are used to study the health effect of exposure to pollutants. The first part of the Chapter briefly discusses Generalized linear models (GLM) and Generalized additive models (GAM). We present a review of more advanced methods, Functional data analysis (FDA) and Bayesian methods. Discussion of functional data analysis comprises estimating functions from discrete observed measurements, function alignment methods, functional principal component analysis (FPCA) and functional regression models. In the Bayesian method Section, we review prior distributions, Markov Chain Monte Carlo (MCMC) simulation and Bayesian hierarchical models.

Chapter 3 motivates the main problem in more detail, and outlines common issues which are encountered in estimating the association between exposure to pollutants and health. The Chapter includes discussion of the nature of pollutant data and how the concentration measures enter to the model. Particularly, we explore how studies use the available pollutant measurements to represent exposure to the pollutant. This particular aspect of the pollutant and health is the focus of the work presented in this thesis. The Chapter includes a brief discussion on overdispersion, model selection, autocorrelation and lag. The use of confounding covariates is also discussed.

Chapter 4 discusses the use of functional data analysis technique to measure daily exposure to ozone and presents functional regression models to investigate the dependence of health outcome on a functional measure of ozone. The Chapter is intended to study data coming from a single geographic location. For application, we consider pollutant and health outcome data from the city of Milan, Italy. A functional exposure measure of the pollutant is estimated from the discrete hourly measurements using functional data analysis tools. We employ function alignment (Ramsay and Silverman, 2005) which aligns the common features of functional observations to identify the portion of the daily ozone concentration curve potentially linked to health. Thus, the functional linear regression model is fitted considering both the aligned and non-aligned ozone curves as functional covariate and hospital admission counts as response. Within this Chapter, the assumption of linearity in the dependence of health on ozone exposure measured as function is relaxed, and the functional generalized additive model is considered to estimate a non-linear flexible ozone effect. We close this Chapter by demonstrating that the functional regression models have superior predictive performance over the standard models using out-of-sample predictive study. The work presented in this Chapter is currently under revision with the Journal of Environmetrics with the title *Flexible Functional Modelling of Short-term Effect of Ozone: Application to City of Milan, Italy.* The same work has been published in the proceedings at the 21st International Conference on Computational Statistics (COMPSTAT2014) in Geneva, Switzerland, 2014, with the title *Functional data modeling to measure exposure to ozone.* A reduced version of the same work has also been presented at 1st International Workshop on Large scale population-based surveys on respiratory health in Italy and Europ, Verona, Italy with the title *Modeling exposure to ozone and hospital admission.*

Chapter 5 discusses an extension of methods in Chapter 4, used when data come from multiple geographic locations. We propose a functional hierarchical modeling approach to estimate pollutant exposure effect allowing for the daily variation of the concentration and the spatial heterogeneity of the effect at once. The approach is developed using the functional regression model discussed in Chapter 4 and the Bayesian hierarchical model paradigm. An application is considered to data from 15 USA cities for the summer periods (June-July-August) of years 1987-2000. We consider two possible specification, the first specification is an overall model fitted by pooling all the city data together and can not account spatial heterogeneity. Thus, we estimate one marginal ozone effect as a function of daily time from the pooled data. The second specification is a hierarchical model in which we obtain an overall functional regression coefficient as well as location-specific coefficients. The approach is general and can also be considered as the analogue of the classical multilevel (hierarchical) models to the case in which the predictor is functional and the response is scalar. The work presented in this Chapter is currently under working paper for publication with the title *Functional hierarchical model for pollutants and health.* The same work has also been presented at joint Graybill/ENVR conference on Modern Statistical Methods for Ecology, Colorado State University, USA with the title *Functional hierarchical model for pollutants and health.*

Chapter 6 provides an alternative and simpler approach, which still meets the aim of providing a representative exposure measure allowing for daily variation of the pollutant concentration but in a more parsimonious way. This approach uses a fixed number of principal scores derived from the hourly concentrations using functional principal component analysis technique. The principal scores capture the important portion of the variability of the concentration curve to be used as a potential exposure measure. To model the health effect of the pollutant exposure measured by principal scores, we adopt two model specifications, in the same spirit as Chapter 5, first the effect of principal scores on health is studied assuming homogeneous exposure effect across the different geographic locations. The second approach considers the Bayesian hierarchical model to allow for spatially heterogeneous effect, which assumes the association between the principal scores and health can possibly vary across the cities. This framework allows to estimate city-specific effect as well as the overall scores effect.

Chapter 7 presents the conclusion of this thesis by reviewing the research questions and discusses the main results. The Chapter synthesises the findings of the thesis in a wider context. The limitations of the thesis and future work are discussed.

1.2 Contributions of the Thesis

The work presented in this thesis focuses on two major problems of exposure to environmental pollutants and health. These are how to best measure daily exposure to a pollutant and how the association between exposure to the pollutant and health is affected by geographic locations (in strength and shape). There have been rare methodological analyses to address these problems. Studies traditionally collapse the daily pattern of ozone concentration to a single daily summary figures which fail to account for the temporal variations. Some studies have forwarded their approach to deal with the issue, Chiogna and Pauli (2011) attempted to address the issue by defining a number of alternative measures to account for different features of the daily pattern. Staniswalis et al. (2009) adopted historical functional model to examine the effect of particulate matter on daily mortality. Our approach (Arisido, 2014) used the functional regression approach (Ramsay, 2006) to effectively account for the daily fluctuations of the pollutant. The superiority of measuring exposure in the form of function over the other approaches is demonstrated using out-of-sample predictive performance.

The other main issue for environmental studies of pollutants and health is that the estimated health effect is heterogeneous across different geographic locations or studies (Bell et al., 2004), a circumstance that may be due to many factors, for instance differences in the industrialization levels of the cities or their weather conditions. To resolve the issue, meta-analysis (Dumouchel, 1995; Clayton et al., 1993) and Bayesian hierarchical models (Dominici et al., 2000) have been advocated to obtain a pooled and an unbiased location-specific estimates by sharing 'strength' across the different locations. However, these multi-study or -location methods depend on the daily summary measures of ozone exposure, thus ignoring the daily variation of the pollutant. To address the issue, we shall illustrate a functional hierarchical regression model using Bayesian paradigm. In general, the contributions of this thesis can be illustrated as follows:

- Often pollutants are monitored hourly, but in the standard models, the exposure is typically reduced to an aggregate measure such as the daily maximum or 24 hour average. We instead propose hourly ozone measurements of one day as a function, and then using the function as the exposure measure.
- The health effect of exposure to the pollutant is estimated as a function of daily time to examine the influence of the exposure continuously throughout the day in contrast to one single scalar estimate provided by the standard methods.
- As far as the health effect is concerned, different parts of the daily concentration curve may be more or less relevant, we identify these degrees of relevance using functional data analysis.
- We propose a functional version of the Bayesian hierarchical model to estimate health effect of exposure accounting for temporal variations of the pollutant and spatial heterogeneity of the effect. The approach will serve as the analogue of the standard hierarchical/multilevel model when observations are obtained in the form of functions and is used as a method of combining functional information from different locations to make inference in the overall effect of exposure.
- We present an alternative approach in the form of principal scores to measure the daily exposure to the pollutant. The principal scores are computed from hourly measurements of pollutant concentration allowing for daily variation, and capture the most important portion of the daily concentration curve to be used as potential measure of daily exposure to explain health. This approach will serve as an additional exposure measure for researchers to be used as alternative to our functional exposure measure and other approaches in literature.

Chapter 2

Statistical Methods Review

Most of the earliest studies estimated the association between exposure to pollutants and health using Poisson log-linear models. The changes in environmental and weather conditions, which have a direct influence on pollutants, required to use more general model to accommodate various confounders such as seasonal and weather variables. Generalized additive models (GAM) have been widely used to flexibly control for confounders. Despite these advances to model daily exposure to pollutants and health, there are crucial issues, as highlighted in the introduction, which remain to be addressed. These issues motivate to implement more flexible complex modelling techniques. In this Chapter, we review both the widely used standard methodology and the more complex methods which are used in this thesis.

The Chapter is organized as follows. Section 2.1 reviews the standard statistical methods used to model the association between health outcomes and exposure to pollutants. Section 2.2 discusses the functional data analysis technique, which includes representing a function using the spline basis approach, function alignment, functional principal component analysis and functional regression methods. Section 2.3 describes the Bayesian method.

2.1 Classical Statistical Methods

This section discusses the standard inferential tools used to study the association between exposure to pollutants and health. The most commonly used methods are the generalized linear models (GLM) and the generalized additive models (GAM).

2.1.1 Generalized Linear Models

The classical linear model is used to study the relationship between T independent response data $\boldsymbol{y} = (y_1, \ldots, y_T)'$ and a matrix of covariates $\boldsymbol{X}$ with the assumption that $\boldsymbol{y}$ is normal variate. Despite this restrictive distributional assumption, the linear model has widespread applications and the statistical theory for inference on its parameters is well developed. The generalized linear models (GLM) are an extensions of the classical linear regression model to non-normal response variables $\boldsymbol{y}$. All generalized linear models are based on a family of distribution called exponential family. A random observation Y taken from a distribution that is a member of the exponential family, has probability function in the form

$$f(y|\theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right], \tag{2.1}$$

where θ is parameter of the exponential family and ϕ is the dispersion parameter used to represent the scale parameter of the distribution. In some cases ϕ may be known, for example the Poisson distribution, in that case θ is commonly known as canonical parameter. When ϕ is unknown, the family is more properly called the exponential dispersion family. For instance, both the Gamma and Normal probability distributions have their own scale parameter and ϕ is unknown. The expression $b(\theta)$ is a known function which is useful to derive the mean and the variance of the given exponential family.

$$\mathbb{E}(y) = \mu = b'(\theta) \quad \text{and} \quad \mathbb{V}ar(y) = \phi b''(\theta),$$

where $b'(.)$ and $b''(.)$ indicate the first and second derivatives of b respectively. The mean is fully specified by the parameter θ only, while the variance is a function of θ and the dispersion parameter ϕ. The variance function $b''(\theta)$ specifies the link between the mean and the variance. Here, a Poisson distributed random variable has no dispersion parameter, as a result ϕ is assumed to be 1. However, care must be taken before assuming that there is no extra variation in the data to be accounted by ϕ, that leads to overdispersion issue. We shall discuss overdispersion in Section 3.3.1. Assume that Y_t can come from any exponential family distribution, a generalized linear model can be specified using the link function $g(.)$ to describe how the mean response, $\mathbb{E}(Y_t) = \mu_t$, is linked to the predictors. Then, a generic form of generalized linear model is

$$\begin{aligned} Y_t &\sim f(y_t|\mu_t, \phi), \qquad t = 1, \ldots, T \\ g(\mu_t) &= \boldsymbol{X}_t'\boldsymbol{\beta}, \end{aligned} \tag{2.2}$$

where $\boldsymbol{X}_t' = (X_{t1}, \ldots, X_{tp})$ is a $T \times p$ design matrix of predictors and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ denote coefficients specifying the effect of the associated predictor on the response. Parameter estimation and inference of a generalized linear model is based on the theory of maximum likelihood estimation (Pfanzagl, 1994). For more details on GLM, see McCullagh and Nelder (1989), Myers et al. (2012) and Dobson (2001). In epidemiological studies of pollutants and health, the response is daily mortality or morbidity counts. Since these daily counts are assumed to have Poisson distribution, a log link is the natural choice to specify a generalized linear model which describes the association between exposure to a pollutant and daily mortality or morbidity accounting for confounders. In this context, expression (2.2) can be modified as

$$\begin{aligned} Y_t &\sim \text{Poisson}(\mu_t), \qquad \text{for } t = 1, \ldots, T \\ \log(\mu_t) &= X_t\beta + \text{confounders}, \end{aligned} \tag{2.3}$$

where the predictor X is a measure of daily exposure to a pollutant, the parameter β represents the effect of exposure to a pollutant as measured by X on the response. In this modeling approach, the pollutant measure and confounding covariates have a linear relationship with a transformation of the mean of the response.

2.1.2 Generalized Additive Models

The generalize linear model specified in (2.2) is fully parametric and some situations require more flexibility, specifically, to allow for a non-linear effects. The generalized additive models (GAM, Hastie and Tibshirani (1990)) are a semiparametric extension of generalized linear models. The basic idea is to replace one or more of the linear predictors in (2.2) by a spline function, yet other predictors may still be included as linear components. The general form of a generalized additive model can be presented as

$$g(\mu_t) = \boldsymbol{X}_t'\boldsymbol{\beta} + f_1(z_{1t}) + f_2(z_{2t}) + \cdots + f_J(z_{Jt}), \tag{2.4}$$

the first part of the right side of the equation is the same as (2.2), and the smooth functions f_j of the additional covariates z_j are included to allow flexibility on the dependence of the response on the covariates z_j. To estimate such a model, each function f_j should be specified as a linear combination of known basis functions $\phi_{j1}(z), \ldots, \phi_{jK}(z)$. We then have

$$f_j(z_{jt}) = \sum_{k=1}^{K} c_{jk}\phi_{jk}(z_{jt}), \tag{2.5}$$

where $(c_{j1}, \ldots, c_{jK})$ are unknown coefficients and they will be estimated from the model fitting techniques. Here, suitable choice of basis functions has to be made for appropriate representation of f; the chosen basis should aid in determining a smooth curve $f(z_j)$ that approximates the effect of z_j (for more details about basis functions, see Section 2.2.1). A model matrix

$\boldsymbol{Z}_j$ can be created for each function, so that the functions can be represented as

$$\boldsymbol{f}_j = \boldsymbol{Z}_j \mathbf{c}_j, \tag{2.6}$$

where, $\boldsymbol{Z}_j = (\phi_{j1}(z_{jt}), \ldots, \phi_{jK}(z_{jt}))$ and $\boldsymbol{c}_j = (c_{j1}, \ldots, c_{jK})'$. Considering both the linear and the non-linear components, the model matrix can be specified in the form $[\boldsymbol{X} : \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_J]$ and model parameters $\boldsymbol{\theta} = [\boldsymbol{\beta} : \boldsymbol{c}_1, \ldots, \boldsymbol{c}_J]$. Parameters in generalized additive models are often estimated by penalized likelihood maximization, where the penalties are used to suppress the roughness of the f_j terms. Assume that we have a penalty matrix $\boldsymbol{S}_j$ to penalize the functions f_j. Then, a penalized log-likelihood has the form

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - \frac{1}{2} \sum_j \lambda_j \mathbf{c}_j' \boldsymbol{S}_j \mathbf{c}_j, \tag{2.7}$$

where λ_j are smoothing parameters which are used to control the trade-off between goodness of fit of the model and smoothness. The estimation of model parameters are conditional on the unknown λ_j, therefore λ_j must be estimated first using, for example, generalized cross-validation (GCV, Golub et al. (1979)). More discussion on GCV is provided in Section 3.3.2. Assuming that λ_j are known, the penalized likelihood maximization estimates parameters using iterative procedures. Different iterative algorithms have been proposed: local scoring (Hastie and Tibshirani, 1990), backfitting (Buja et al., 1989) or the penalized version of iteratively re-weighted least squares (P-IRLS, Wood (2006)). For more details of generalized additive models, see Wood (2006), Li (1986) and Ruppert et al. (2003). Now, we shall discuss the method to model exposure to environmental pollutants and health.

The generalized additive models are the most widely used modelling approach in pollutants and health studies. The flexibility of the models allow to introduce confounding covariates as smooth spline functions. Assuming X is a measure of daily exposure to a pollutant and $f(z)$ is smooth function of a confounding covariate, we can specify (2.2) in the form of generalized additive model

$$\begin{aligned} Y_t &\sim \text{Poisson}(\mu_t), \quad t = 1, \dots, T \\ \log(\mu_t) &= X_t \beta + f(z_t). \end{aligned} \tag{2.8}$$

The parameter of interest is β which describes the association between day-to-day variability of the pollutant and health. An advantage with this specification is that the researcher has complete control to decide the degrees of freedom to estimate $f(z)$. Although such choice is an issue in its own right (Peng and Dominici, 2008), sensible choice can be made so that the estimate is not over- or under-smooth. Inferences on β may correspond to quantifying its uncertainty using confidence interval. For the estimate of curve f_j, inference is typically made using point-wise confidence bands. A further general discussions on the use of generalized additive models for estimating the association between exposure to pollutants and health are available in Dominici et al. (2002) and Zanobetti et al. (2000).

2.2 Functional Data Analysis Method

The term "Functional data" was introduced by Silverman and Ramsay (2005) to denote when the collected data are available in the form of curves. The expression "Functional Data Analysis" (FDA) is used to indicate the methodology for dealing with functional data (Ferraty, 2011). The functional data analysis is a new area of Statistics and extends established methodologies and theories from the field of image analysis, generalized linear models, multivariate data analysis, nonparametric statistics and many others. This Section presents a review of functional data analysis with the main focus being on estimating a smooth function from the discrete observations, function alignment, functional principal component analysis and the functional regression models.

Within the field of functional data analysis, there are two schools of thought based on how they conceptualize functional data (Shang, 2014). On one side, some researchers consider functional data analysis as a smoothed version of multivariate data analysis, and functional data analysis represent the multivariate data analytical tools in the language of functional analysis. On the other side, researchers underline that functional data analysis is the development of the statistical application of spline functions, particularly in the scope of nonparametric function estimation. Although there is a difference between the two stances, the common feature is that a single observation in a functional data analysis is a whole function defined on bounded common interval, as opposed to focusing on the discrete number of observed values at particular points in the interval. When the discrete observed values appear collectively as a function, they reflect a certain smoothness property which allows functional data interpretation. Further, the discrete values in a function may display high correlation, in which case the standard multivariate data analysis fails. The other feature of functional data analysis is that a function may be estimated from the fixed number of discrete observed measurements, but functional data are intrinsically infinite dimensional.

These ideas will be made more clear at different stages of this Section. In particular, in what follows we will discuss how functional data can be represented and how their representation could be estimated from the discrete observed data. We consider a situation where a random quantity is observed at several different times $(h_1, \dots, h_J)$, and the discrete observations can be represented as $\tilde{X}(h_1), \dots, \tilde{X}(h_J)$. Then, a continuous form of the expression is given by $\{X(h) : h \in (h_1, h_J)\}$. Thus, a functional data is the observation of T functionals $X_1(h), \dots, X_T(h)$. These functional data and their discrete form can be viewed as a traditional data matrix as shown in Table 2.1.

$X_1(h_1)$	$X_1(h_2)$	$\dots$	$X_1(h_J)$
$X_2(h_1)$	$X_2(h_2)$	$\dots$	$X_2(h_J)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$X_T(h_1)$	$X_T(h_2)$	$\dots$	$X_T(h_J)$

Table 2.1: Functional data and its discretized observations presented in the form of classical data matrix.

There are J measurements at $h_1, \dots, h_J$ for each curve. The basic assumption is that there are underlying continuous functions that generated those discrete observations and the interest is mainly in such functions. Further, we assume that the underlying function $X(h)$ is *smooth*, so that a pair of adjacent observations $X(h_j)$ and $X(h_{j+1})$ are linked together to some extent and unlikely to be wildly different from each other. This principle is the base of treating the data as functional rather than just multivariate.

In order to perform any type of functional data analysis, the first step is to estimate the underlying function using the discrete observations. The use of basis functions and smoothing techniques are the main components to produce a flexible representation of the discrete observations by functional. Generally, the basis function methods represent a function $X(h)$ by a linear combination of K know basis functions $\boldsymbol{\phi}(t) = (\phi_1(h), \dots, \phi_K(h))$ as

$$X(h) = \sum_{k=1}^{K} c_k \phi_k(h), \tag{2.9}$$

where $\mathbf{c} = (c_1 \dots, c_K)'$ are the coefficients of the basis to be estimated from the data using smoothing methods. A sufficiently good approximation is achieved when K is large. However, the approximation depends not only on K, the type of basis $\phi_k(h)$ and the method used to estimate c_k are particularly important.

2.2.1 Basis Functions

The basis function representation is used to approximate a function by taking a linear combination of sufficiently large number K of basis functions. The most frequently used bases methods are polynomial bases, Fourier series and the spline bases. Different basis have different properties, and which is the most appropriate one depends on the characteristics of the function to be approximated. For instance, Polynomial bases are convenient when the interest focuses on properties of $X(h)$ in the vicinity of a single specified point as opposed to over its whole domain (Wood, 2006). A Fourier series is used if the observed data are periodic and uniformly smooth. In this Section, we shall discuss the use of the popular spline bases method.

B-Spline Basis Functions

Splines are pieces of polynomials with order m that are tied together smoothly. A spline function on a given interval is constructed by dividing the interval into subintervals, whose limits are called knots. Thus, to define a spline, the order m of the polynomials must be decided as well as the location and number of knots. The order m of a polynomial is the number of constants required to define it and is one more than its highest power (degree). Over each interval, a spline is a polynomial of specified order m. The locations of the knots must be chosen according to the nature of the data. Typically, the knots would either be evenly spaced through the range of observed $X(h)$ values, or placed at quantiles of the distribution of unique $X(h)$ values (see for example Ruppert et al. (2003) and Wood (2006)). The higher the order of the spline, the more flexible the shape can be between the knots and the better will be the approximation. Increasing the number of knots allows more flexibility as well. There are different spline basis used to construct spline function, in this Section, we will illustrate the widely used B-spline basis, which provides great flexibility and computational efficiency. The basis functions are local, that means each basis function is only non-zero over the intervals between $m+2$ adjacent knots. Assume that a basis composed of K functions is to be used, and define $K+m$ knots, $h_1 < h_2 < \cdots < h_{K+m}$, the $m-1$ order B-spline basis functions are defined recursively in the form

$$\phi_k^{m-1}(h) = \frac{h - h_k}{h_{k+m-1} - h_k}\phi_k^{m-2}(h) + \frac{h_{k+m} - h}{h_{k+m} - h_{k+1}}\phi_{k+1}^{m-2}(h), \tag{2.10}$$

and

$$\phi_k(h) = \begin{cases} 1 & h_k \le h < h_{k+1} \\ 0 & \text{otherwise,} \end{cases}$$

see De Boor (1976) for further details. Figure 2.1 illustrates an example of B-splines basis with varying order, knots and number of basis defined over the range [0,23], say the range contains the daily hours. The B-splines shown in panel a is an order one with single knot, and this is a step function, panel b depicts three piecewise linear B-splines with a single knot, Panel d shows an eight B-splines of order 4 defined by four equally spaced knots. Looking at panel d, the basis functions in the middle are non-zero in a maximum of four adjacent subintervals. The basis functions at both ends are also positive in at most four adjacent subintervals. This property of B-spline can be generalized as: an order m B-spline basis function is positive in at most m adjacent intervals, and this property is the main reason for B-spline being mathematically efficient and flexible.

2.2.2 Estimating Function from Discrete Measurements

Let $y_t = (y_{t1}, \ldots, y_{tJ})$ be the discrete observations for which the replicate t is observed at each j for $j = 1, \ldots, J$ and its functional version be represented by $X_t(h) = \sum_{k=1}^{K} c_{tk}\phi_k(h)$, where $\phi_1(h), \ldots, \phi_K(h)$ denote the B-spline basis. The classical least square provides a simple linear smoother by minimizing the sum squared error (SSE)

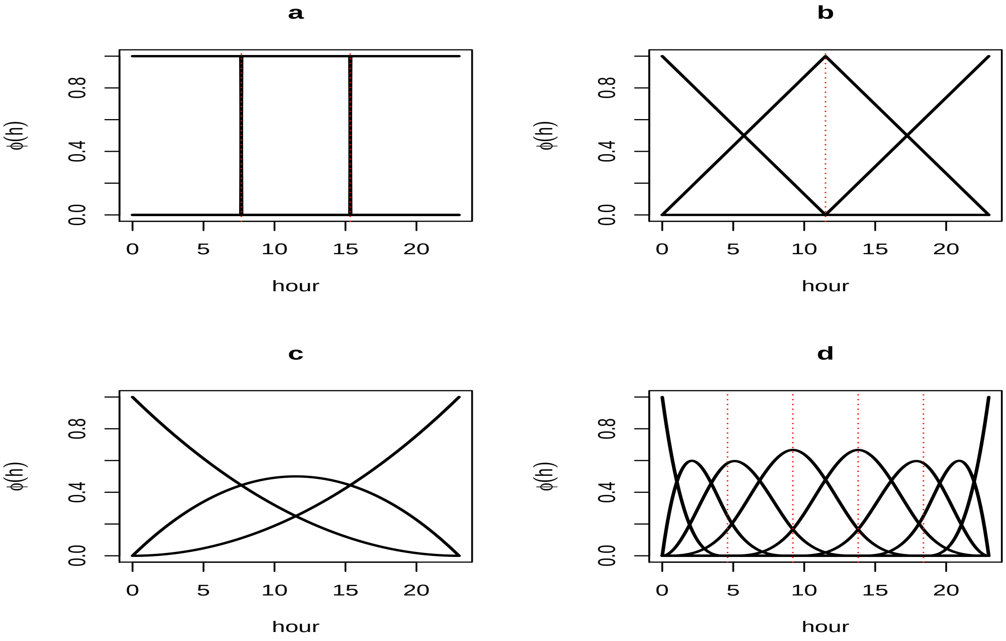

Figure 2.1: B-spline basis with 3 basis functions of order 1 with 2 knots (a), 3 basis functions of order 2 with 1 knots (b), 3 basis functions of order 3 without knot (c) and 8 basis functions of order 4 with 4 knots (d). The vertical dotted lines are the positions of the knots.

$$\mathrm{SSE}(X_t|y_t) = \sum_{j=1}^{J}\left[y_{tj} - \sum_{k=1}^{K} c_{tk}\phi_k(h_j)\right]^2. \tag{2.11}$$

The least square procedure has different modifications to obtain a good approximation. The weighted least square and the localized least square or the kernel smoothing are among those. For more details, see Eubank (1999), Green and Silverman (1993) and Fan and Gijbels (1996). Here, we opt to present the roughness penalty approach, which is more powerful and elegant in the context of smoothing a function. In the roughness penalty approach (O'Sullivan, 1986), a penalty is imposed on the sum squared of the error to control the smoothness of the estimated curve as flexibly as possible. Thus, the penalized sum of squared error (PSSE) compromises the smoothness against the goodness of the fit to the data

$$\text{PSSE}_\lambda(X_t|y_t) = \sum_{j=1}^{J} [y_{tj} - X_t(h_j)]^2 + \lambda \int_1^J \left[D^2 X_t(h_j)\right]^2 \text{dh}, \tag{2.12}$$

where D^2 indicates the second derivative, and the integrated squared second derivative measures the roughness of the curve. The smoothness parameter λ controls the smoothness of the curve, and can be selected by generalized cross-validation criterion (GCV, Golub et al. (1979)). As λ becomes larger, the weight given to the integrated squared second derivative becomes larger, consequently the criterion $\text{PSSE}_\lambda(X_t|y_t)$ places more emphasis on the smoothness of $X_t(h)$ and less on the goodness of fit. For small λ, the function becomes more variable. To briefly show how $\text{PSSE}_\lambda(X_t|y_t)$ is computed, let us use the matrix form $X_t(h) = \sum_{k=1}^{K} c_{tk}\phi_k(h) = \boldsymbol{c}_t'\boldsymbol{\phi}(\boldsymbol{h})$, $\mathbf{c}_t$ is the K-vector of coefficients associated to replicate t and $\boldsymbol{\phi}(h)$ is the K-vector of B-spline basis functions. From a simple minimization problem of (2.11) without penalization, $\mathbf{c}_t$ has the solution $\hat{\boldsymbol{c}}_t = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'y_t$. Here, $\boldsymbol{\Phi}$ is the J by K matrix containing the values of K basis functions at the J sampling points, and y_t is the vector of discrete data associated to replicate t. The idea is to find the analogous estimate of $\mathbf{c}_t$ corresponding to the roughness penalty approach specified in (2.12). The integrated squared of second derivative is expressed in matrix form as follows

$$\begin{aligned}
\int \left[D^2 X_t(h_j)\right]^2 \text{dh} &= \int \left[\boldsymbol{c}_t' D^2 \boldsymbol{\phi}(h)\right]^2 \text{dh} \\
&= \int \boldsymbol{c}_t' D^2 \boldsymbol{\phi}(h) D^2 \boldsymbol{\phi}'(h) \boldsymbol{c}_t \text{ dh} \\
&= \boldsymbol{c}_t' \left[\int D^2 \boldsymbol{\phi}(h) D^2 \boldsymbol{\phi}'(h) \text{dh}\right] \boldsymbol{c}_t \\
&= \boldsymbol{c}_t' \boldsymbol{Q} \boldsymbol{c}_t,
\end{aligned}$$

where $\boldsymbol{Q}$ is the roughness penalty matrix, given by

$$\boldsymbol{Q} = \int D^2\boldsymbol{\phi}(h)D^2\boldsymbol{\phi}'(h)\mathrm{dh}, \tag{2.13}$$

we then have a minimization problem in the matrix form

$$\mathrm{PSSE}_{\lambda}(y_t|\boldsymbol{c}_t) = (y_t - \boldsymbol{\Phi}\boldsymbol{c}_t)'\,(y_t - \boldsymbol{\Phi}\boldsymbol{c}_t) + \lambda\boldsymbol{c}_t'\boldsymbol{Q}\boldsymbol{c}_t.$$

Taking the derivative with respect to $\mathbf{c}_t$ and manipulating the resulting expression for the estimated coefficient, we obtain

$$\hat{\boldsymbol{c}}_t = \left(\boldsymbol{\Phi}'\boldsymbol{\Phi} + \lambda\boldsymbol{Q}\right)^{-1}\boldsymbol{\Phi}'y_t.$$

Depending on the type of basis functions, $\boldsymbol{Q}$ may be computed analytically or numerically. For example the FDA package by Graves et al. (2009) contains programming code for the B-spline basis functions. However, in many applications numerical approximation to the integrals are implemented. Even in the B-spline case, the details are fairly complicated (see Ramsay (2006)), and researchers usually opt for numerical approximations.

2.2.3 Aligning Functional Observations

Functional observations often share common features such as the peak, the minimum or the valley. These observable features vary according to the size and the position of the features. Thus, functional data display two type of variations, the first one is amplitude or vertical variation, which is a variation in the vertical size of a particular feature in a sample of curves. For instance, the peak of one curve may be greater or less than the other. The second is phase or horizontal variation, which is a variation in the location of a sample of curves features along the horizontal axis. For instance, the position of peaks in a sample of curves my not be obtained at the same time point. Figure 2.2 presents a synthetic example to illustrate the issues of the amplitude-phase variations. The existence of these two type of variations in functional observations poses several issues. For instance the cross-sectional mean function may not be used to represent the average of functions, since the mean function is dissimilar to each of the curves in the presence of phase variation (Silverman and Ramsay, 2005). This issue can also be observed in Figure 2.2 (a) where the mean function (shown as dashed curve) does not resemble any curve.

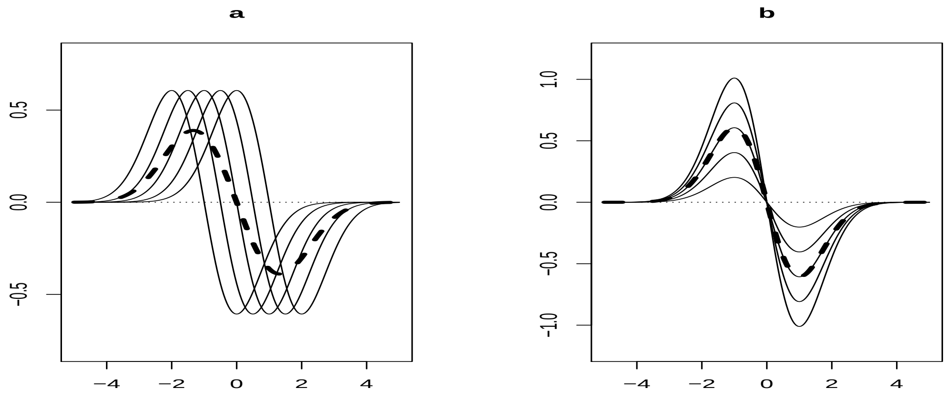

Figure 2.2: An illustration of phase and amplitude variations, which exist in functional observations, taken from Graves et al. (2009). Panel a shows five curves varying only in phase, panel b shows five curves varying only in amplitude. The dashed line in each panel indicates the mean of the five curves.

The issue of phase-amplitude variations is not only restricted to the cross-sectional functional mean, Kneip and Ramsay (2008) suggests other issues related to computing variances, correlations and principal components analysis. Thus, curve aligning or registration is basically a means of avoiding the unnecessary phase variation by transforming the curves via transformation of their arguments. We shall implement curve alignment in order to align the daily pollutant curve to remove unwanted phase variation. Further, it allows us to identify the portion of the daily curve which is assumed to be potentially harmful to health. These applications will be discussed in Chapter 4, in this Section, we shall provide general discussion as background information.

Different curve alignment methods have been proposed. The commonly used methods are: shift, landmark and continuous curve alignment. Shift alignment is the simplest method which is used to align the time scale. The curves are shifted by an amount δ_t to align a feature at a common time point. The aligned curves are then $X_t^*(h) = X_t(h + \delta_t)$. The parameter δ_t is estimated using the version of least square criterion to identify a shift δ_t for curve t. Further details on this method are available in Silverman and Ramsay (2005). The continuous method is used when there are no clearly identifiable features. In this case, the method uses the entire curves rather than their values at specified points. The method involves aligning the curves to a target curve or other function. This could be the average curve of the functions, one of the curves or some other function of interest. The curves are made as similar to a target function as possible. For more details, see Graves et al. (2009). The most popular method is Landmark alignment method, used when a curve has a clear feature or landmark that we can associate with a specific argument value h. We shall discuss this method in more detail, since it is the method we will use in the next Chapter to align the daily ozone curves.

Landmark alignment method removes phase variation by monotonically transforming the argument for each curve so that points specifying the locations of the features are aligned across curves. Features or landmarks of a curve may be maxima, minima and crossings of fixed thresholds, which may be defined at the level of one or more derivatives. Notationally, we can specify the location of a feature f for which $f = 1, \ldots, F$ as h_{tf}, which can also be the argument values for each curve associated to the feature. Assume that W_t is a transformation of curve X_t, then landmark alignment is given in the form

$$X_t^*(h) = X_t[W_t(h)],$$

where $X_t^*(h)$ is the aligned curve and W_t is a strictly monotonic function called warping function which defines the alignment. Through this definition, each aligned curve has identical argument values at the target location h_{0f}. In order to use the warping functions to align curves at the same location, W_t must satisfy the following properties over the interval [1,J]: $W_t(h_1) = h_1$, $W_t(h_J) = h_J$ and $W_t(h_{0f}) = h_{tf}$ for $f = 1, \ldots, F$. These properties ensure that the beginning and ending locations are already aligned and that all the transformation occurs between those locations. We refer to Chapter 4 for a further illustration with examples.

2.2.4 Functional Principal Component Analysis

Functional Principal Component analysis (FPCA) is one of the main standard inferential tools for the analysis of functional data. The core objectives of FPCA are capturing the principal mode of variations on one side and dimension reduction on the other side. In order to understand FPCA, consider a set of variables denoted by a vector X. Variation on X is often summarized by either the covariance or the correlation matrix in multivariate context. Then, the vector X is decomposed into components using the spectral decomposition of symmetric (covariance/correlation) matrix (see Härdle and Simar (2007) for multivariate statistics). The majority of multivariate PCA theories carry over to the FPCA. Particularly, Mercer's theorem (Indritz, 1963) provides an analogous spectral decomposition for functional data. Assume that a functional observation $X_t(h)$, where h is observed in the interval $[1, J]$, is a square integrable random function with mean $\mu(h)$

$$\mu(h) = \mathbb{E}[(X_t(h)], \quad \text{for } \ t = 1, \ldots, T$$

and the covariance operator $K(s, h)$

$$K(s,h) = \mathbb{C}ov[X_t(s), X_t(h)] = \mathbb{E}\{[X_t(s) - \mu(s)][X_t(h) - \mu(h)]\}.$$

We assume that there is an orthogonal expansion in terms of eigenfunctions $\phi_l(h)$ for $l = 1, \ldots$ and the associated non-increasing eigenvalues $\lambda_1 \geq \lambda_2 \ldots$ to decompose $K(s,h)$ in the form

$$K(s,h) = \sum_{l=1}^{\infty} \lambda_l \phi_l(s) \phi_l(h), \quad s,h \in [1, J]. \tag{2.14}$$

An important advance on the theory of FPCA is the Karhunen-Loeve expansion (Karhunen (1947); Loeve (1965)) which allows to express a random curve $X_t(h)$ as

$$X_t(h) = \sum_{l=1}^{\infty} \xi_{tl} \phi_l(h), \quad h \in [1, J], \tag{2.15}$$

where the coefficients

$$\xi_{tl} = \int_1^J X_t(h) \phi_l(h) dh,$$

are uncorrelated random variables with zero mean and variance λ_l. These random variables are called principal component scores. Hence, X_t is decomposed into orthogonal components with uncorrelated principal scores as coefficients. The eigenvalues λ_l are a measure of the variation in X_t in the ϕ_l direction. The aim is to retain only the first L eigenvalues and eigenfunctions in Karhunen-Loeve expansion to capture the important modes of variations. Hence, the FPCA can achieve dimension reduction. To motivate this further, we choose the first eigenfunction $\phi_1(h)$ to capture types of variation that are very strongly represented by the data by maximizing the variance (λ_1) of the principal score ξ_{tl} subject to the constraint $||\phi_1^2(h)|| = \int \phi_1^2(h) dh = 1$. The second eigenfunction can be obtained similarly, find $\phi_2(h)$ and compute ξ_2. The value ξ_2 has maximum variance λ_2, subject to the constraint $||\phi_2^2(h)|| = \int \phi_2^2(h) dh = 1$ and additional requirement $\int \phi_1(h) \phi_2(h) dh = 0$.

The idea here is that we seek the dominant mode of variation again, but we require the second eigenfunction $\phi_2(h)$ to be orthogonal to the first $\phi_1(h)$, so that they are presenting different information. With successive steps the amount of variation explained will decline on each step.

In practice, the principal scores (ξ_{tl}) are popularly used to describe the important components of variations. They can be estimated by plugging the estimate of eigenfunction $\phi_l(h)$ in the formula $\xi_{tl} = \int X_t(h)\phi_l(h)dh$ and evaluating over the grid of points. This is the standard method in FPCA, we shall show in Chapter 5 that the ξ_{tl} can be estimated as part of model parameters in the Bayesian model setting. To choose the number of eigenfunctions, the share of estimated explained variance is often used. For illustration, we shall consider ozone data coming from the city of Milan, Italy. More elaboration and application on the data will be given in Chapter 4. Now, let $X_t(h)$ denote ozone curve recorded over the daily hour $h \in [0, 24]$ in the day t for $t = 1, \ldots, T$. We want to retain only a few principal components to highlight the dominant mode of variability across the daily hours. We can rely on the traditional scree plot to identify the number of principal components. Figure 2.3 shows that the first six principal components can be deemed sufficient to describe the mode of variability observed in the original functional data. The cumulative variance plot in the right panel shows the proportion of variance against the associated component which suggests the same conclusion as the scree plot.

The popular advantage of FPCA is its application in functional regression models. Postponing details of functional regression models until Section 2.2.5, FPCA can be an important ingredient for the functional regression models, since the response variable of a model can be specified as a function of functional principal component scores of the predictor process. The idea of using principal components of the predictors in place of the predictors was first raised by Jolliffe (2002). The motivation is that in addition to dimen-

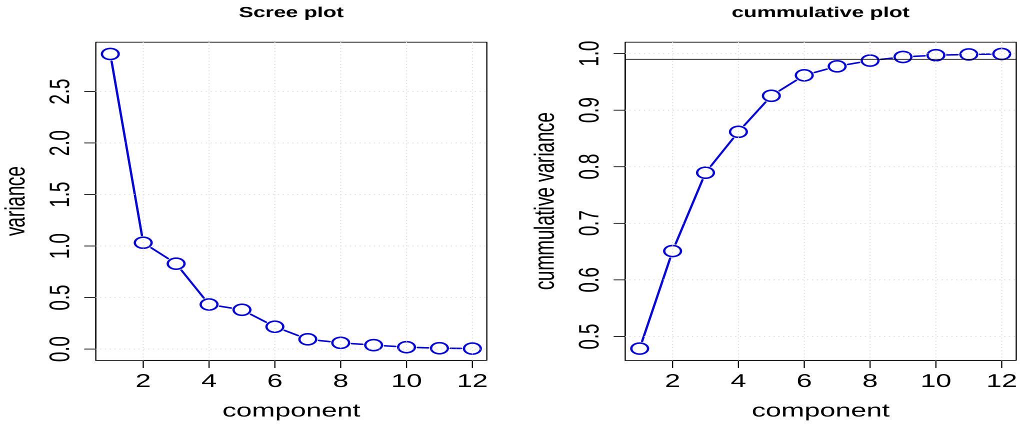

Figure 2.3: The Scree and cumulative variance plots for the functional principal component analysis of the Milan ozone data.

sion reduction, the retained principal components are uncorrelated, which allows to resolve multicollinearity related issues caused by the presence of multiple predictors that are potentially correlated. Shang (2014) points out further practical advantages in the use of PCA for regression model. In the functional data analysis context, Ramsay and Dalzell (1991) forwarded the idea of using functional principal scores instead of the functional variable. Since then FPCA has been used for several functional regression problems. James (2002) presented a discussion to show how FPCA can be used to gain insight into the relationship between the response and functional predictors and applied it to standard missing data problems. James and Silverman (2005) considered the decomposition of the predictor function into a sum over its functional principal components for functional adaptive model estimation. Müller and Yao (2008) describe the use of functional principal components in an additive non-linear structure rather than linear way to explain the response variable. In a further advance, Di et al. (2009) and Crainiceanu et al. (2009) introduced functional principal component scores that can be used for multilevel regression models. The approach is also implemented in the Bayesian framework (Crainiceanu and Goldsmith, 2010).

2.2.5 Functional Regression Models

Functional regression models allow to describe the variability of a response using covariates. Both the response and the predictor may be functional, so there are three different scenarios: First, when the response variable y is functional and the predictor X is scalar. Second, when both the response and the predictor are functional, and the last scenario is when the response is scalar and the predictor is functional. The first situation involves predicting functional response $y(h)$ using multivariate covariates. Particularly, a T-vector of smooth functions $y(h)$ is related to known T by P design matrix of covariates ($\mathbf{X}$) by a linear combination of P parameter functions $\boldsymbol{\beta}(h)$ which are to be estimated using the data. The general form of the model is $y_t(h) = \boldsymbol{\beta}(h)\mathbf{X}_t + \epsilon_t(h)$ where $\epsilon_t(h)$ is an independent realization of stochastic process with mean zero and same bivariate covariance function $K(h_1, h_2)$. The P-vector parameter $\boldsymbol{\beta}(h)$ can be estimated using the least square version of the model. For more detail, see Faraway (1997) and Silverman and Ramsay (2005).

The second scenario constitutes a regression problem in which a function $X_t(s)$, for $s \in [0, S]$, is used as predictor to explain the variation in the response function $y_t(h)$, for $h \in [0, H]$, through a linear model

$$y_t(h) = \int_0^S X_t(s)\beta(s,h)ds + \epsilon_t(h), \tag{2.16}$$

where $\epsilon_t(h)$ is a residual function and $\beta(s,h)$ is the bivariate regression coefficient function. Assuming that both s and h are on the same interval ($S = H$), three different cases can be identified: the first is $y_t(h)$ can be affected by $X_t(s)$ for a future time $s > h$. This situation is usually applicable if the process is periodic (Malfait and Ramsay, 2003). The second case is $y_t(h)$ can be affected by $X_t(s)$ at the same time $s = h$, in this case the model is called concurrent or point-wise, in the sense that X_t only influences $y_t(h)$ through its value $X_t(h)$ in contrast to the influence of X_t can involve a range of argument values $X_t(s)$. The third case addresses when the behaviour of y at time h depends only on the behaviour of X at times $s \leq h$. This case is the most common implementation of model (2.16), sometimes called the 'historical functional linear model' named by Malfait and Ramsay (2003). A well known application is that for an indicator of a patient's recovery $y(h)$ which may depend linearly on the time course of treatment $X(s)$ and the association only involves times $s \leq h$. Other applications and more details are available in Bosq (2000), Cardot et al. (1999) and Ramsay and Silverman (2002). Since our focus lies on modeling a scalar response and functional predictor in the subsequent Chapters, we shall not provide any more discussion on other scenarios.

Let $X_t(h)$ for $h \in [1, J]$ denote the functional predictor and $\mathbf{y} = (y_1, \ldots, y_T)'$ represent the scalar response. The functional linear regression model which assumes a linear relationship between the functional predictor and the response is given by

$$y_t = \int_1^J X_t(h)\beta(h)dh + \epsilon_t, \quad t = 1, \ldots, T, \tag{2.17}$$

where $\beta(h)$ is the coefficient function, ϵ_t is the error term. In practice, $X_t(h)$ is observed at a finite set of discrete time points. We may imagine simply replacing the integral with a summation over the observed times and see this as a finely discretized version of the functional model being considered. However, this approach has two main issues: first it may lead to fit an extremely high dimensional vector of coefficients, resulting in large or infinite variance terms (Silverman and Ramsay, 2005). Second, the procedure fails to make use of the intrinsic relationship between values of X_t observed at close proximity (James, 2002). Instead, we use a basis expansion approach to allow for the underlying smooth pattern. Then, both the covariate function $X(h)$ and the functional coefficient $\beta(h)$ can be specified as

$$X_t(h) = \sum_{l=1}^{L} c_{tl}\phi_l(h) = \mathbf{c}_t'\boldsymbol{\phi}(h) \text{ and } \beta(h) = \sum_{k=1}^{K} b_k\theta_k(h) = \mathbf{b}'\boldsymbol{\theta}(h),$$

where ϕ_l for $l = 1, \dots, L$ and θ_k for $k = 1, \dots, K$ are basis functions for $X_t(h)$ and $\beta(h)$ respectively, $\mathbf{c}_t$ is L - vector of coefficients (for more discussions on estimating $X_t(h)$, see Section 2.2.2). However, the real parameter of interest is the K-vector parameter $\mathbf{b}$ which describes the dependence of the response on the functional covariate. Now, we can re-write (2.17) as

$$\hat{y}_t = \int_1^J X_t(h)\beta(h)dh = \int_1^J [\mathbf{c}_t\boldsymbol{\phi}(h)]\left[\boldsymbol{\theta}(h)'\mathbf{b}\right]dh = \mathbf{c}_t\mathbf{J}_{\phi\theta}\mathbf{b}, \tag{2.18}$$

where the L by K matrix $\mathbf{J}_{\phi\theta}$ is defined as

$$\mathbf{J}_{\phi\theta} = \int_1^J \boldsymbol{\phi}(h)\boldsymbol{\theta}(h)'dh. \tag{2.19}$$

The objective is to estimate $\mathbf{b}$ in order to recover the estimate of the functional coefficient $\beta(h)$. Recalling from Section 2.2.2 that the roughness penalty approach allows the trade off between smoothness of the function and the fit of the data, we can define the penalized residual sum of squares (PSSE) by imposing a penalty term on $\beta(h)$ as follows

$$PSSE_\lambda(\beta) = \sum_{t=1}^{T}\left[y_t - \int_1^J X_t(h)\beta(h)dh\right]^2 + \lambda\int_1^J \left[D^2\beta(h)\right]^2 dh. \tag{2.20}$$

Using definition (2.19) for $\mathbf{J}_{\phi\theta}$ and redefining $\mathbf{Q}$ from (2.13) as

$$\boldsymbol{Q} = \int_1^J D^2\boldsymbol{\theta}(h)D^2\boldsymbol{\theta}'(h)\mathrm{dh},$$

the penalized residual sum of squares can be expressed in the form

$$PSSE_\lambda(\beta) = ||\mathbf{y} - \mathbf{c}\mathbf{J}_{\phi\theta}\mathbf{b}||^2 + \lambda\mathbf{b}'\boldsymbol{Q}\mathbf{b}.$$

If we represent the coefficient matrix $\mathbf{c}\mathbf{J}_{\phi\theta}$ by Z, the expression further simplifies to

$$PSSE_{\lambda}(\beta) = ||\mathbf{y} - Z\mathbf{b}||^2 + \lambda \mathbf{b}'\boldsymbol{Q}\mathbf{b}. \tag{2.21}$$

Thus, the minimizing value $\hat{\mathbf{b}}$ satisfies

$$\left(Z'Z + \lambda\boldsymbol{Q}\right)\hat{\mathbf{b}} = Z'\mathbf{y}. \tag{2.22}$$

The smoothing parameter λ can be chosen using the cross-validation paradigm. Once an estimate of $\beta(h)$ is found, inferences can be made on the estimated coefficient. Particularly, the variance of the estimated function can be estimated in order to compute the point-wise confidence intervals. Using specification (2.22), the variance of estimated $\mathbf{b}$ is

$$\mathbb{V}ar(\hat{\mathbf{b}}) = \mathbb{V}ar\left[\left(Z'Z + \lambda\boldsymbol{Q}\right)^{-1} Z'\mathbf{y}\right], \tag{2.23}$$

things are straight here, since the variance-covariance matrix computed from the residuals is a scalar estimate σ_{ϵ}^2 and the other factors are just constants. Thus, the variance of $\hat{\mathbf{b}}$ is

$$\mathbb{V}ar(\hat{\mathbf{b}}) = \sigma_{\epsilon}^2\left(Z'Z + \lambda\boldsymbol{Q}\right)^{-1} Z'Z\left(Z'Z + \lambda\boldsymbol{Q}\right)^{-1}.$$

We have discussed the simplest form of linear functional regression problem consisting of a scalar outcome and a single functional predictor. Extensions of the model can be considered, for instance, when there are multiple functional predictors and a mix functional and scalar predictors. In general, functional linear models are a recent advance and they have crucial limitations. All the above discussions are based on the assumption that the response is Gaussian, hence the models are restrictive, and a full framework to account several distributions is required. For example, in the next Chapters we shall consider a functional pollutant curve to predict the daily mortality counts and hospital admissions, needing to fit a Poisson regression model. Some researchers have been actively working in this context.

The functional generalized linear model (FGLM), the extension of the generalized linear model (GLM) of McCullagh and Nelder (1989), has been proposed to model various distributional outcomes and functional predictors. Müller and Stadtmüller (2005) put forward a method using a link function $g(.)$ to associate the expected value of the scalar response $\mathbb{E}(y_t)$ and a linear predictor for which the linear predictor is obtained by a scalar product of the functional predictor $X_t(h)$ and a smooth parameter function $\beta(h)$. A related work is presented by James (2002) by extending GLM to handle functional predictors, and interpretable results are presented by decomposing $X_t(h)$ using functional principal component analysis. In Chapter 4, we shall adopt a functional generalized linear model in order to explain the Poisson distributed hospital admission counts using the profile of daily ozone curve without decomposing $X_t(h)$.

Another limitation of the standard functional linear model is the linearity assumption imposed on the dependence of the scalar response on the functional predictor. The idea of GAM (see, Section 2.1.2) has been used for functional data to allow a more flexible association of scalar response and functional predictor. James and Silverman (2005) presented estimation method to extend various models such as GLM and GAM when the predictor is a functional observation. The method depends on a functional principal components decomposition of the predictor function. Müller and Yao (2008) proposed a functional additive model to replace the linear association with an additive structure, implemented through a projection on the eigenbasis of the covariance operator of the functional components in the model. These two are recent advances to establish more general and flexible functional regression methods. However, these approaches rely on the functional principal component decomposition. Mclean et al. (2014) argues that a model that is additive in the principal component scores is not additive in $X_t(h)$ itself, and proposed a functional additive model which regresses on the functional predictors directly. In Chapter 4, we shall follow this latest paradigm to examine the association between pollutant curves and hospital admission data relaxing the linearity assumption.

2.3 Bayesian Statistical Method

The methods presented in this thesis so far have been related with the frequentist statistical paradigm. The alternative statistical paradigm is the Bayesian one. The Bayesian data analysis method estimates a parameter using two sources of information: the observed data and information about the parameter that is known before the data is observed. The information from the observed data is represented by the likelihood $f(\boldsymbol{y}|\theta)$, and the prior knowledge about the parameter is given in the form of a probability distribution $f(\theta)$, which is called prior probability distribution. The interest in Bayesian method is to obtain an updated distribution for θ by combining the prior knowledge about the parameter and the data. This can be achieved using Bayes' theorem

$$f(\theta|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\theta)f(\theta)}{f(\boldsymbol{y})} = \frac{f(\boldsymbol{y}|\theta)f(\theta)}{\displaystyle\int f(\boldsymbol{y}|\theta)f(\theta)d\theta}, \tag{2.24}$$

where $f(\theta|\boldsymbol{y})$, the updated distribution, is called the posterior probability distribution of θ. The Bayes' theorem states that the posterior distribution is proportional to the product of the observed data and the prior distribution, since the denominator which depends only on the data, is simply a normalizing constant. In the Bayesian statistics, the normalizing constant is known as marginal likelihood, since it is the weighted average of $f(\boldsymbol{y}|\theta)$ with a weight function being the prior distribution $f(\theta)$. Hence, in a Bayesian model, both the prior distribution and the likelihood must be fully specified. The likelihood comes from the usual model specification, but the information about the parameter in the form of prior distribution may not be available. Thus, choosing plausible prior distribution is needed before embarking on Bayesian modeling. Here, we present a review of the main classes of prior distributions.

2.3.1 Prior Distributions

The choice of prior distributions depends on a variety of situations. Traditionally, choice of prior has been influenced by computational issues. For instance, computationally tractable posterior is obtained by adopting conjugate prior. This type of prior has the nice property of implying a posterior of the same distributional family as the prior (Ntzoufras, 2011). This means that the posterior distribution follows a known parametric form, making computations simpler. However, in many occasions it may not be possible to achieve a conjugate prior distribution. Furthermore, the emergence of powerful and efficient computational algorithms means conjugacy may no longer be the main driver in choosing a prior. When there is available information or personal opinion about the parameter before the data is seen, this can be incorporated in the model as prior information. Prior can also be chosen to incorporate information about the model or study design (Berger, 1985). On the other hand, we may be prior ignorant and know nothing about the parameter before fitting the model. Prior distributions can therefore belong to one of two classes: informative or noninformative.

Informative prior distributions reflect the case when substantial prior information is available that can be turned into a probability distribution. The information can be in the form of historical data, previous studies, expert knowledge or following formal rules which express prior skepticism and optimism. Noninformative prior distributions, also called weakly informative, diffuse or vague, are selected to reflect our ignorance about the parameter, and the posterior is mainly driven by the data. The motivation is that we may not wish to make use of prior knowledge for some reasons. For instance, lack of any prior knowledge or we wish to make inference on the basis of the data only and we want to let the data to speak for themselves without introducing external information. It must be recognized that when a noninformative prior is preferred, we sometimes adopt an improper prior, that is, $f(\theta)$ is not a probability distribution, since it does not integrate to 1, this is generally accepted as long as it leads to a proper posterior. Otherwise, the Bayesian analysis is not valid, since the posterior is not a probability distribution.

2.3.2 Bayesian Hierarchical Models

Hierarchical models are increasingly important to model different types of complex data, since they allow to capture aspects of the data that can be interest to study. The Bayesian hierarchical model typically express the model in multiple levels to estimate the parameters that are connected in some way. The simplest form of Bayesian hierarchical model is to model the data $\mathbf{y}$ conditional on a parameter θ and this parameter is in turn described by a probability distribution with the underlying parameter α. The parameter α is called hyperparameter. These concepts of Bayesian hierarchical models are popularly used when the data are grouped. In this type of application the parameter θ will be group specific to describe a specific group and α can be treated as overall parameter. For example, we may have t observations within each of G groups. Assume that the data are distributed within groups according to some distribution with group varying parameters (θ_g). We assume that the group varying parameters (θ_g) come from a common distribution with shared parameter α. A posterior distribution for all unknown parameters is

$$f(\theta_g, \alpha|y) \propto f(y|\theta_g, \alpha)f(\theta_g|\alpha)f(\alpha).$$

Here, θ_g have the prior distributions $f(\theta_g|\alpha)$, they are conditional on another parameter α (hyperparameter) which has its own prior $f(\alpha)$ (called hyperprior). The product of the last two terms yield a joint distribution of $f(\theta_g, \alpha)$, this is the marginal joint distribution of the two parameters, which is in turn multiplied by the distribution of the data. By the Bayes' theorem, this gives the joint posterior distribution of the parameters. In principle, we can specify a model with any number of levels by introducing more hyperprior distributions. However, Bayesian hierarchical models that have greater than three levels involve interpretation issue (Gill, 2007). A further discussion of Bayesian hierarchical model is available in Chapter 5, where we incorporate the ideas of this section in the functional hierarchical modeling procedure.

2.3.3 Markov Chain Monte Carlo (MCMC)

In the Bayesian analysis, when the prior and the likelihood are conjugate, the posterior distribution of a parameter can be found analytically. In more general situations, often the parameter space is high dimensional and computing the integrals becomes extremely prohibitive. Thus, to learn about a particular parameter θ_q from p dimensional parameter space $\theta_1, \dots, \theta_p$, we wish to summarize the marginal posterior of θ_q given the data, which involves integrating out all the parameters except θ_q

$$f(\theta_q|\boldsymbol{y}) = \int f(\theta_1, \dots, \theta_{q-1}, \theta_{q+1}, \dots \theta_p|y) d\theta_1 \dots d\theta_{q-1} d\theta_{q+1} \dots d\theta_p,$$

the posterior distribution needs to be estimated using sampling algorithms. The Markov Chain Monte Carlo (MCMC) method is widely used sampling algorithm to simulate samples from the posterior. The Markov Chain Monte Carlo do not require integration, it uses simulation procedures to simulate repeatedly from the joint posterior of all parameters. We shall present here a brief review of Markov Chain Monte Carlo, for a more detailed discussion, see Gill (2007) and Gelman et al. (2003). The foundation of Markov Chain Monte Carlo is based on two methods: the Monte Carlo simulation and Markov Chain sampling. The Monte Carlo simulation is a method used to calculate numerically integrals of complex function. The idea in Monte Carlo simulation is to simulate random values which are assumed to be from the correct distribution, then use these generated empirical values to approximate the unknown integral quantity. The key assumption in the use of Monte Carlo simulation is that with large number of independent simulated values, the Monte Carlo approximation will converge to the true values. This is guaranteed by the strong law of large numbers. On the other hand, the theory of Markov Chains for a stochastic process is that future states are independent of past states given the present state. Thus, the Markov property ensures that the simulated sample values form a chain are slightly dependent on the previous one. The chain wanders around the parameter space, remembering only where it has been in the last period. The transition kernel or transition function governs the probability of moving the chain to some other state based on the current state (Albert, 2007). The Markov Chain Monte Carlo therefore simulates draws that are slightly dependent. The transition kernel is defined so that the ergodic distribution of the chain is the distribution we want to simulate from, that is the posterior. We then take the draws and compute quantities of interest for the posterior distribution. In Bayesian statistics, there are generally two Markov Chain Monte Carlo algorithms that can be used to draw samples: the Gibbs sampler and the Metropolis-Hastings (MH).

The Gibbs sampler (Geman and Geman (1984)) draws successive samples from the full conditional probability distribution of each parameter in turn, conditional on the current values of the other parameters and the data. The conditional distribution of θ_q is

$$f(\theta_q|\mathbf{y}, \theta_1, \dots \theta_{q-1}, \theta_{q+1}, \dots \theta_p).$$

However, to determine the conditional distribution of θ_q, the joint distribution must be known. The Hammersley-Clifford theorem (Robert and Casella, 2004) is used as the foundation to obtain the joint distribution from knowledge of the conditional distribution. Thus, the full conditional distribution for each parameter is the distribution of the parameter conditional on the data and all the other parameters. The Gibbs sampler sampling procedure can be summarized as follows

1 . Choose a vector of initial values $\boldsymbol{\theta}^{(0)}$

2 . Draw a value $\theta_1^{(1)}$ from the full conditional $f(\theta_1|\theta_2^{(0)},\ldots,\theta_p^{(0)},\mathbf{y})$

3 . Draw a value $\theta_2^{(1)}$ from the full conditional $f(\theta_2|\theta_1^{(1)},\theta_3^{(0)},\ldots,\theta_p^{(0)},\mathbf{y})$. Here,the updated $\theta_1^{(1)}$ from step 2 is used. Step 3 is repeated for the remaining parameters.

4 . Draw $\theta_1^{(2)}$ from the full conditional $f(\theta_1|\theta_2^{(1)},\ldots,\theta_p^{(1)},\mathbf{y})$

5 . Draw a value $\theta_2^{(2)}$ from the full conditional $f(\theta_2|\theta_1^{(2)},\theta_3^{(1)},\ldots,\theta_p^{(1)},\mathbf{y})$. Repeat step 5 for the remaining parameters by continually using the most updated values.

6 . Repeat the process until we get M draws for each parameter.

The idea is then to regard the final part of the chain as an identically independently distributed sample from the posterior distribution. The Gibbs sampler fails when the full conditional distributions can not be obtained. The Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) provides a method to sample from the posterior distribution without the need to determine the full conditionals. The approach is based on randomly proposing a new value θ^* for the parameter. If this proposed value is accepted according to a specified acceptance criterion, then the next value in the chain becomes the proposed value $\theta^{(m+1)} = \theta^*$. If the proposal is rejected, then the previous value is retained $\theta^{(m+1)} = \theta^{(m)}$, and another proposal is made and the chain progresses by assessing this new proposal.

The most common way to create proposal values is to add a random variable to the current value: $\theta^* = \theta^{(m)} + Q$. If we wish to make θ^* closer to $\theta^{(m)}$, we could choose Q from a standard normal distribution with a relatively low variance. If we wish all proposals within one unit of the current value to be equally likely, then we can use the uniform distribution $Q \sim U[-1,1]$. In either case, the probability distribution of Q is known as proposal density. The acceptance criterion is given in the form

$$\theta^{(m+1)} = \begin{cases} \theta^*, & \text{if } U < a \\ \theta^{(m)}, & \text{otherwise} \end{cases}$$

where U is a randomly drawn from a uniform distribution between 0 and 1, and a is acceptance probability, which can be given as

$$a = \min\left\{ \frac{f(\theta^*|\mathbf{y})}{f(\theta^{(m)}|\mathbf{y})} . \frac{Q(\theta^{(m)}|\theta^*)}{Q(\theta^*|\theta^{(m)})}, 1 \right\},$$

under the original Metropolis algorithm constraints, the two conditionals can be symmetric, that is $Q(\theta^{(m)}|\theta^*) = Q(\theta^*|\theta^{(m)})$, in that case a can be simplified to

$$a = \min\left\{ \frac{f(\theta^*|\mathbf{y})}{f(\theta^{(m)}|\mathbf{y})}, 1 \right\}.$$

Whatever Markov Chain Monte Carlo algorithm has been used, exploratory analysis is often conducted to monitor if the draws are approximately from the posterior distribution. This includes assessing the convergence of the algorithm to the target distribution so that the chain is actually drawn from the posterior. Part of the convergence assessment is to investigate if the chain has sufficiently explored the entire posterior distribution. Usually, we need to ensure that the chain is long enough, since the chain is not immediately taken from the posterior distribution, rather we have to wait until the sampling distribution has converged to the posterior. The initial part of the chain (called burn-in) is therefore not representative for the posterior, and can be avoided when computing the posterior summary measures. Thus, we need to determine the size of burn-in as part of convergence assessment. In general, the convergence can be assessed using graphical and formal diagnostic techniques. The graphical techniques include plots of Trace, Autocorrelation and Running mean. The details of the graphical techniques are available in Albert (2007). The formal diagnostics include the Geweke (Geweke et al., 1991) and Brooks-Gelman-Rubin diagnostics (Gelman and Rubin, 1992).

2.3.4 Inference

Bayesian inference about a parameter θ can be made using the Markov Chain Monte Carlo sample once the convergence assessment methods ensured that the Markov chain is drawn approximately from the true posterior. To make an inference, we need to compute posterior quantities of the Markov chain to summarize the posterior distribution. The posterior summaries of θ are a point estimate such as the mean, median or mode. Often, the mean or the median is widely used, the choice between the two is dependent on the nature of the posterior density. The posterior mean is the common approach in most cases. However, the posterior median is preferred if the distribution is skewed. By definition, the posterior mean is computed from the posterior distribution $f(\theta|\mathbf{y})$ as

$$\mathbb{E}(\theta|\mathbf{y}) = \int \theta f(\theta|\boldsymbol{y}) d\theta.$$

However, in many situations the integrals can not be computed analytically, and we have a Markov chain which is drawn from the posterior using appropriate Markov Chain Monte Carlo algorithm. Let assume that our Markov chain have T sample values after burn-in which are drawn from the posterior, then the mean is

$$\mathbb{E}(\theta|\mathbf{y}) \approx \bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \theta^t.$$

The posterior variance is used to specify the uncertainty in the parameter. The variance is also crucial to make Bayesian inferences in terms of credible intervals. The credible interval is the analogue of the concept of confidence interval used in classical statistics. The posterior variance can be computed as

$$\mathbb{V}ar(\theta|\mathbf{y}) \approx \sigma_{\theta}^{2} = \frac{1}{t-1}\sum_{t=1}^{T}(\theta^{t} - \bar{\theta})^{2}.$$

Chapter 3

The Problem and Modeling Issues

In this Chapter, we shall motivate the main problems in more detail, and outline common issues which are encountered in estimating the association between exposure to pollutants and health. We start by presenting a general description of data commonly used in pollutants and health studies, Section 3.2 focuses on the main issues which are the core of the work presented in this thesis. Section 3.3 illustrates some issues which commonly arise in estimating the health effect of exposure to pollutants and gives directions on how these will be addressed in the subsequent Chapters.

3.1 Data Description

Environmental pollutant exposure studies use data which typically consists of health data, the measurements of pollutants concentration and various confounding variables including a measure of weather conditions. The health data includes daily counts of hospital admission (Morbidity) or death (Mortality). Such health data are usually collected by medical facilities and should be classified for list of causes using internationally accepted standards. The classification is used to identify the causes of the admission or the mortality. Particularly, to study the health effects of exposure to pollutants, the total non-accidental causes are commonly used as health outcome variables. However, there are studies which use cause-specific outcomes such as respiratory (Mudway and Kelly, 2000) or cardiovascular issues (Shah et al., 2013). In Chapter 4, we use non-accidental hospital admission from city of Milan, Italy as health outcome variable and in Chapters 5 and 6 the outcome variable will be mortality from 15 different cities of the USA. A common issue in those type of health data is the presence of a seasonal trend, which needs to be adequately addressed before fitting a model. Section 3.3.3 assesses seasonal trend in the outcome variables which would be modeled in the subsequent Chapters.

The commonly studied pollutants for their impacts on health are Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), ground level Ozone (O3), Particulate Matter (PM) and Sulphur Dioxide (SO_2). The concentrations of a pollutant is typically measured using the metric of micrograms per meter cube ($\mu g/m^3$). Among these pollutants, ozone and particulate matter are frequently studied in different regions of the world. Ozone is the main pollutant studied in the remainder of this thesis, and Section 3.2 gives greater detailed discussion about ozone. Particulate matter is a complex mixture of small particles and liquid droplets including acids and other organic chemicals. The concentrations of particulate matter are measured in two different metrics: ($PM_{2.5}$) and (PM_{10}) to denote particles that are less than $2,5\mu g/m^3$ and $10\mu g/m^3$ in diameter respectively. Particulate matter is usually included in the studies of exposure to ozone and health as confounder, since any observed relationship between ozone and health may reflect Particulate matter effect. Thus, the models in Chapter 4 accommodate the effect of particulate matter for the dependence of health on ozone. The Chapter did not include the other pollutants, since too many pollutants associated with health outcomes are frequently correlated to each other (Levy et al., 2005). Further, the health effect of the other pollutants is generally ignored

(Bell et al., 2005). Chapters 5 and 6 consider only the health effect of ozone, since the concentrations of particulate matter including the other pollutants are missing in the majority of the study days.

Commonly included covariates as confounding are related to weather conditions. Such measures include temperature, dew point temperature and humidity. These measures are used to remove the influence of weather from the estimated pollutant effect on health. Temperature is the most commonly included covariate in pollutants and health studies, and its effect on health is widely studied (Armstrong, 2006). The daily measure of temperature is often given in the form of daily maximum or average. Figure 3.1(a) reports daily maximum temperature against hospital admission for the city of Milan in the years 1996-2002. The Figure shows high mortality rate at more extreme temperature (low and high). This type of 'U' shaped temperature effect is commonly estimated (see for example, Armstrong (2006)), other researchers have obtained 'N' shaped estimate (see for example, Pauli and Rizzi (2006)). To capture this inherent pattern, temperature is include as smooth function (see Section 2.1.2) into the regression model rather than as a linear effect.

In the remainder of the thesis, we include a smooth function of daily maximum temperature in the models. In addition, categorical covariates day of the week and calendar year will be included in our analysis. Inclusion of day of the week allows different baseline health outcome within each day of the week, and calendar year is typically used to protect the association between pollutant and health from confounding by longer-term trends due to changes in health status, influenza epidemics and seasonality. Their inclusion in the studies of pollutants and health have been widely advocated (Staniswalis et al., 2009; Dominici et al., 2000). Figure 3.1(b) displays the distribution of hospital admission data for the city of Milan across day of the week.

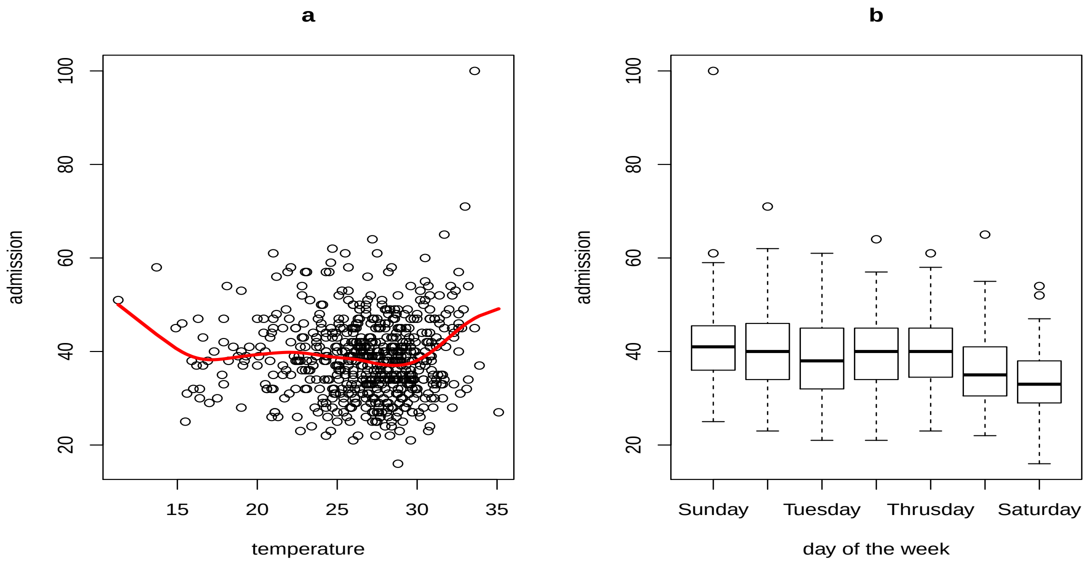

Figure 3.1: The relationship between the daily maximum temperature and the number of hospital admission (a), where the shape of the relationship is indicated by the bold red line. The distribution of hospital admission across the day of week (b).

3.2 The Problem

Ground level ozone (O3) is a potent environmental pollutant and can cause a variety of health problems including asthma and other lung diseases. Ozone is a secondary pollutant, that is the primary pollutants such as hydrocarbons and nitrogen oxides which are direct products of combustion, undergo a chemical reaction in the atmosphere in the presence of sunlight to form ozone. The process of ozone formation is therefore a continuous process, which follows different daily fluctuations exhibiting strong daily patterns (Gao, 2007). The concentrations of ozone are usually measured by a network of fixed number of monitoring sites in a particular study region. Each monitoring site typically measures hourly throughout the day, which leads to 24 measurements of pollutant concentrations for a day. Some researchers study the between site variability within the study regions. However, the majority of studies focus on estimating the effect in the region by aggregating the hourly measurements of each site. Thus, for a given study region or city there is a $T \times 24$ matrix of measurements, which relate to the T days of the study. The effect of ozone is likely to be detected in the summer time (June - July - August) rather than the whole year. During the summer time, the concentrations of ozone could potentially reach unhealthy levels, since the presence of a relatively strong sunlight drives the formation of ozone. All our analyses will therefore be limited to the summer periods. For example the hourly distributions of ozone for city of Washington DC is displayed in Figure 3.2. The Figure shows a typical summer time hourly ozone concentration. There is a clear pattern for which the concentrations reach maximum in the afternoon hours, and measurements remain low in the morning and night hours.

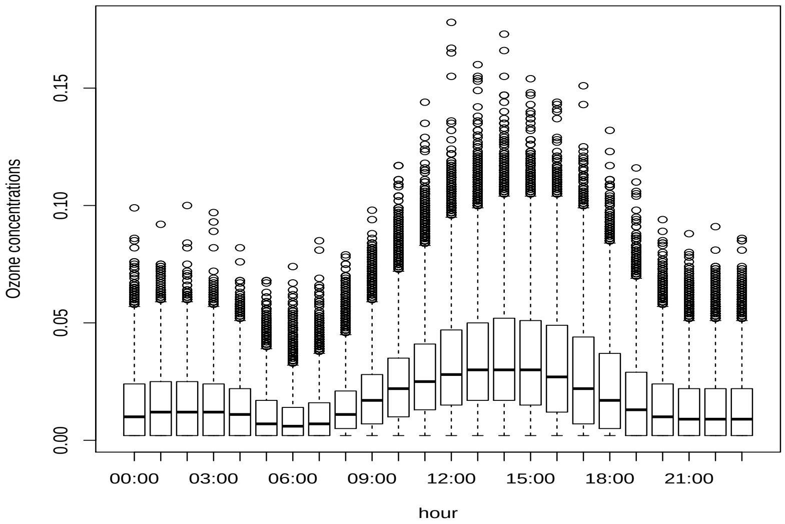

Figure 3.2: The distribution of hourly ozone concentrations for city of Washington DC.

Despite the advances in ozone monitoring and recording, there continue to be gaps in analyzing the monitored data (Seinfeld, 1991). One main issue in this regard is the method to measure human exposure to ozone. Systematic modeling that takes into account the daily patterns of ozone has been particularly rare. Despite the availability of the hourly measurements from monitoring networks, studies collapse the hourly measurements into single daily summaries. The most frequently used daily summary figures are the average and maximum of the 24-hour measurements. Other point summaries have also been used such as 8-hour maximum (Goldberg et al., 2001; Marr and Harley, 2002). These summary figures lead to the use of the classical statistical methods such as the generalized linear models (GLM) and the generalized additive models (GAM) which are discussed in Section 2.1. However, these daily summary figures are rough synthesis of the hourly measurements of pollutant concentrations and they totally disregard the temporal variability observed in the daily concentrations. Further, the daily summaries may not be representative of the actual personal exposure to individuals, since they are likely to ignore the portion of the time spent outdoor.

There have been rare methodological analyses that address the problems in using daily scalar summaries to represent exposure. Gao (2007) proposed the full use of all hourly measurement of a day to study differences of ozone concentrations across days of the week, but ozone and health outcome relationship was not part of the study. Staniswalis et al. (2009) adopted historical functional model to examine the effects of Particulate Matter ($PM_{2.5}$) on daily mortality. They showed that the highest association between particulate matter mass concentration and daily mortality was found to occur in the morning when particulate matter concentrations peak. Chiogna and Pauli (2011) addressed the issue by defining a number of alternative measures of ozone which help to account for different features of the daily patterns and then employing variable selection methods to determine which features are more relevant. They concluded that the common daily summaries may not be the best choice compared to other measures such as the area of the concentration curve above a certain threshold. However, the technique still rely on some form of data reduction: the hourly measurements being collapsed to two or three summaries. Our proposal (Arisido, 2014) uses the functional regression approach which is discussed in Section 2.2. The method effectively accounts for the daily fluctuations of the pollutant. We will discuss in the next Chapter that our approach has superior predictive accuracy than the standard methods in terms of predictive performance measure.

The other main limitation of studies examining the association between health outcomes and exposure to pollutant concentrations is modeling spatial heterogeneity when data are coming from multiple geographic locations. Most of these studies estimated a statistically significant effect of exposure to pollutants. Such effect has been estimated at various geographic regions and it is spatially heterogeneous (Gryparis et al., 2004), a circumstance that may be due to many factors, for instance differences in the industrialization levels of the locations or their weather conditions. To resolve the issue, meta-analysis (Dumouchel, 1995; Clayton et al., 1993) and Bayesian hierarchical models (Dominici et al., 2000) have been considered as more appropriate methods to assess health effect of exposure to pollutants. These methods have been advocated to obtain pooled estimate and unbiased location-specific estimates by sharing information across the different locations. For instance, Ji et al. (2011) conducted a meta-analysis of short-term ozone exposure and respiratory hospitalizations to evaluate variation across studies. Dominici et al. (2000) applied the hierarchical methodology to pool the estimates of the pollutant effect from the largest 20 USA cities, and Richardson and Best (2003) briefly re-considered this same application. However, these multi-study or -location methods for combining the estimated effects from exposure to pollutants depend on the daily summary measures of pollutant concentration to represent daily exposure, thus ignoring the daily variation of the pollutant.

We wish to extend the functional regression model to the functional hierarchical approach to model data from different geographic regions. The standard hierarchical regression models are well established to deal with a hierarchical or multilevel data. Although functional regression models are increasingly popular to model observations which are represented by functions, there has been rare progress in developing methods that can be used when the observed functions are organized in hierarchical fashion. In fact, the linear functional regression models which are discussed in Section 2.2.5 are not fully developed yet. For example, the excellent monograph for functional data analysis presented by Silverman and Ramsay (2005) which provides a wide range of functional data analysis approaches, allows to fit a functional regression model only to Gaussian outcomes. Some relevant extensions have been made in both frequentist and Bayesian point of views. Di et al. (2009) introduced a multilevel model in the context of functional principal component analysis, which is designated to extract the intra- and inter-subject components of multilevel functional data. A similar idea was implemented in Crainiceanu et al. (2009) to model multi-visit patient data in cohort study and the approach was extended to Bayesian paradigm by Crainiceanu and Goldsmith (2010). We shall propose the functional version of the Bayesian hierarchical model framework in order to estimate pollutant effects accounting for geographic variation. Further, the method can allow us to combine functional information across different cities. This approach is presented in Chapter 5.

3.3 Issues in Pollutants and Health Studies

In the studies of pollutants and health, researchers have been committed to estimating the health effect of exposure to pollutants, and any critique on the estimated effect have been directed to the statistical methods used. However, there are other issues which can influence the estimated effect. In this Section, we briefly discuss the main issues: overdispersion, model selection, autocorrelation and lag.

3.3.1 Overdispersion

A basic issue underlying the use of models for Poisson distributed data is the presence of overdispersion. Under the Poisson assumption, the variance is completely determined by the mean, $\mathbb{V}ar(y_t) = \mu_t$. In practice, the variance of a count response may be greater than its mean. The existence of more variability than assumed by the mean-variance relationship of the model is called overdispersion (McCullagh and Nelder (1989) and Wedderburn (1974)). When the data are overdispersed, the variability is underestimated, which leads to under coverage of confidence intervals. Different proposals have been put forward to tackle the issue. Several researchers relax the mean variance equality assumption by introducing an overdispersion parameter ϕ, such that $\mathbb{V}ar(y_t) = \phi\mu_t$. This approach allows the counts y_t to have variances that might exceed their means μ_t. Gelman and Hill (2006) provided a method to estimate the overdispersion parameter ϕ using residuals. Ruppert et al. (2003) addressed the issue of overdispersion in the framework of random effects models. For more discussions of overdispersion in pollutants and health studies, see Peng and Dominici (2008). To investigate overdispersion in our datasets, we shall use the standardized residuals computed as

$$\frac{y_t - \hat{y}_t}{\sqrt{\hat{y}_t}}, \tag{3.1}$$

where $\hat{y}_t$ denote the fitted values. If we observe that these residuals have a dispersion larger than 1, this may indicate existence of extra variation than assumed by the Poisson model. The distributions of the standardized residuals in our models are fairly normal with mean zero and standard deviation 1 and the size of the residuals is between -2 and 2. The estimate of ϕ in our model computed from the standardized residuals is 1.05, implying that there is no serious overdispersion. For this reason, the models in the reminder of this thesis use the Poisson assumption.

3.3.2 Model Selection

The objective in this section is to briefly discuss the methods used for model selection in the studies of pollutants and health when there is a set of candidate models. The choice of criterion to select the best model among the candidates depends on the statistical method used to model health effect of exposure to pollutants. For instance, if the effect of exposure is estimated using generalized linear models, then the deviance is mostly used. By definition, the deviance is twice the difference between the log-likelihood for the saturated model and the log-likelihood for the present model, meaning that it allows to compare the goodness of the fit of the present model with the fit of the saturated model. Thus, the deviance can be computed for any specific generalized linear model. For Gaussian linear regression models, the deviance is the classical residual sum of squares. For Poisson outcome data, the deviance can be computed as

$$D(\boldsymbol{y};\hat{\boldsymbol{y}}) = 2\sum_{t=1}^{T}\left[y_t \log(\frac{y_t}{\hat{y}_t}) - (y_t - \hat{y}_t)\right]. \tag{3.2}$$

For a more detailed specification and discussions of deviance, see Ruppert et al. (2003) and McCullagh and Nelder (1989). If the health effect of exposure is estimated using generalized additive models (GAM), two criteria are suggested (Wood, 2006) for model selection and choosing the degree of smoothness of non-linear components: the unbiased risk estimate (UBRE) or the generalized cross-validation (GCV). Estimating the smoothing parameter λ depends on the specific distribution of the response $\boldsymbol{y}$. When the dispersion parameter ϕ is known, for example, Poisson distributed data, we can base model selection on UBRE, which is given in the form

$$\frac{D}{T} + 2\frac{P}{T},$$

where D is the deviance, T is the number of observations and P is the total degrees of freedom. The lower the UBRE score, the better the model is in explaining the response by the given predictors. When the dispersion parameter ϕ is unknown, the generalized cross-validation (GCV) is used for model selection. The GCV score is given as

$$\frac{TD}{(T-P)^2} \tag{3.3}$$

Further discussions on GCV is available in Golub et al. (1979), Hastie and Tibshirani (1990) and Wahba (1990).

3.3.3 Autocorrelation

The health outcome data in pollutants and health studies are often mortality or morbidity time series data. Exploratory investigation of such data can be performed to identify whether there are patterns or seasonal trends. Particularly, neighbouring values the data are more similar than those far apart (Brockwell and Davis, 2002). This feature can be examined in any time series data using autocorrelation function. The autocorrelation function (ACF) indicates the strength of correlation between successive values. The function is given usually in the form of

$$\mathrm{acf}(k) = \frac{1}{T}\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t+k} - \bar{y})/c(0) \tag{3.4}$$

where

$$c(0) = \frac{1}{T}\sum_{t=1}^{N}(y_t - \bar{y})^2,$$

the value k indicates the lag of the variable. The autocorrelation function can be plotted against the lag k. In this thesis, we model time series data from Milan, Italy and 15 USA cities. The former dataset contains morbidity to measure health risks and the latter contain daily mortality counts. The presence of significant correlation in these datasets can be an issue. The autocorrelation function reported in Figure 3.3 shows that the morbidity counts in Milan and the daily mortality in New York contain correlation between successive counts.

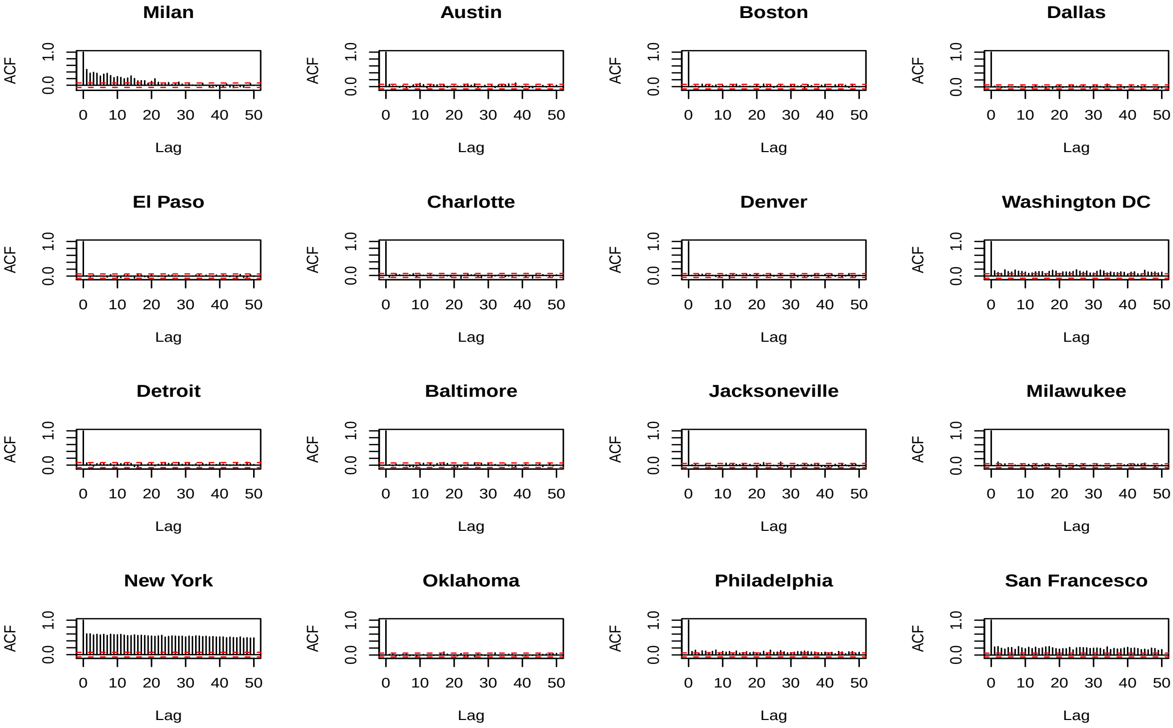

Figure 3.3: Autocorrelation function (ACF) computed from daily hospital admission for the city of Milan in the summer periods of years 1987-2000 and 15 USA cities daily mortality of the summer periods of years 1987-2000. The New York and Milan ACF plots indicate the presence of correlation.

One way to resolve the presence of autocorrelations in modeling environmental exposure and health is to include adjustment predictors (Peng and Dominici, 2008), for example seasonality or meteorological measures. In the reminder of this thesis, we shall include the seasonality adjustment predictors such as calendar year to control the confounding effects of season in modelling the association between health outcomes and the pollutant ozone.

3.3.4 Lag

The adverse effect of exposure to a pollutant may not occur immediately. For this reason, measure of pollutant exposure is lagged by a number of days. Some studies still report the health impacts of exposure to the pollutant measured at the same day (Moolgavkar et al., 1995). When the interest is to look for an association in the following days, it is advisable to report the findings for a number of lags. However, if the number of lags is relatively high and the exposure measure in different lag is serially correlated, this makes the model susceptible to multicollinearity related issues. Zanobetti et al. (2000) suggested to use distributed lag model (DLM) to restrict the estimated parameters to being a low degree polynomial in the lags. The distributed lag model can be reordered more flexibly using nonparametric smoothing through spline functions (Corradi, 1977). There is no single lag that has been consistently used, Zeger et al. (2000) suggested that anywhere between zero and five days is appropriate. In the following chapters, we shall use up to three days lag, that is, we assume that the pollutant effect persist up to three days after exposure. We shall also investigate whether it is likely to persists more than three days.

Chapter 4

Functional Data Analysis for Pollutants and Health

In Chapter 3, we have illustrated issues in the use of daily summary measures, which are obtained by collapsing the daily pattern of pollutant concentration, to represent daily exposure. In this Chapter, we tackle these issues, and propose a more representative daily exposure using functional data analysis. We adopt the corresponding functional regression models to estimate health effect of exposure accounting for the temporal variation of the pollutant. The remainder of this Chapter is organized as follows, Section 4.1 discusses the motivation of the work in more detail. Section 4.2 describes the Milan pollution data set. Section 4.3 outlines the functional data analysis method to estimate a function from the discrete hourly measurements of ozone. Section 4.4 describes aligning the ozone curves using features from the curves. Section 4.5 discusses functional regression approach to predict health outcome using the functional exposure measure. Section 4.5.1 presents the results of the model. Section 4.5.2 compares the predictive performance of our approach and other candidate models which are commonly used and finally, Section 4.6 provides concluding discussions.

4.1 Motivation

One of the main challenges involved in environmental studies of human exposure to ground level ozone has been the approach used for measuring daily exposure to the pollutant to estimate the association between short-term effect of ozone and health. Studies usually collapse the hourly monitored ozone concentration to a single daily summary measures and estimate health effect of exposure by regressing day-varying heath outcome against day-varying summary measure of ozone. This simplistic approach has severe limitations (see, Section 3.2). In this Chapter, we adopt Functional Data Analysis (FDA) to treat all hourly measurements of a day as one function. The functional form of ozone incorporates all hourly discrete measurements accounting for the temporal variations, and aids to uncover important features in the daily concentration curve that might not be observed from the discrete hourly observations. Hence, we assume that a daily smooth concentration curve represents a measure of daily exposure to the concentration of that particular day. We then adopt functional regression techniques to estimate the effect of the exposure using the concentration curve as predictor and hospital admission as health outcome. We explore how the total daily admission counts depends on the specific features of the ozone profile of a day. We compare our approach with that of the traditional daily summary measures and other suggested approaches using out-of-sample predictive study. For application, we use data from city of Milan, Italy in the summer periods of the years 1996-2002.

4.2 The Milan Data

The data set used in this Chapter is from the city of Milan, Italy and comprises the concentrations of pollutants, daily hospital admission, seasonal and weather condition variables for the summer periods (June-July-August) of the years 1996-2002. The concentrations of Pollutants and

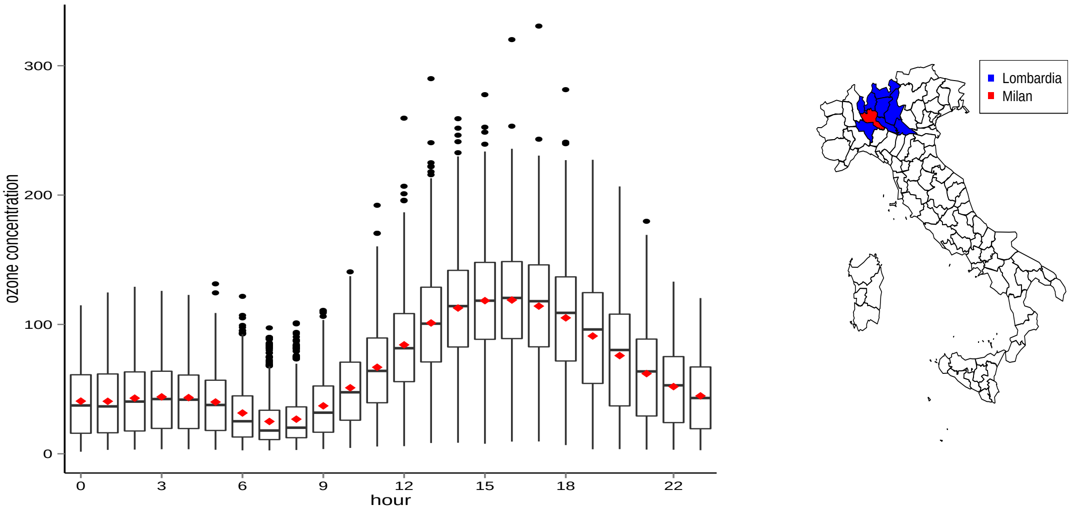

Figure 4.1: Boxplot of hourly ozone concentrations recorded for the summer months 1996-2002 (left) and the geographic map of region Lombardia, its capital city Milan (right).

weather conditions were obtained from the regional agency for environmental protection (ARPA) of Lombardia (Figure 4.2, right). The agency collects hourly concentrations of pollutants from the monitoring network stations in the region. Further details can be obtained from the agency web page (http://www.arpalombardia.it/qaria/). We only obtained the hourly ozone concentrations, the other pollutants and temperature data are given in the form of daily summary statistics. The left panel of Figure 4.2 displays summaries of the hourly ozone measures for the study period, we notice a clear daily pattern with the peak of the ozone concentrations observed in daily hours from 2 pm to 5 pm.

Daily hospital admission data was obtained from the regional health informative system for all hospitals located in the city of Milan. In order to consider events related to health episodes potentially connected to ozone, we ignored records related to surgical events, those scheduled to last less than one day and events for which the reason for admission was not specified. Figure 4.2 reports the distribution of hospital admission data across calendar year and month.

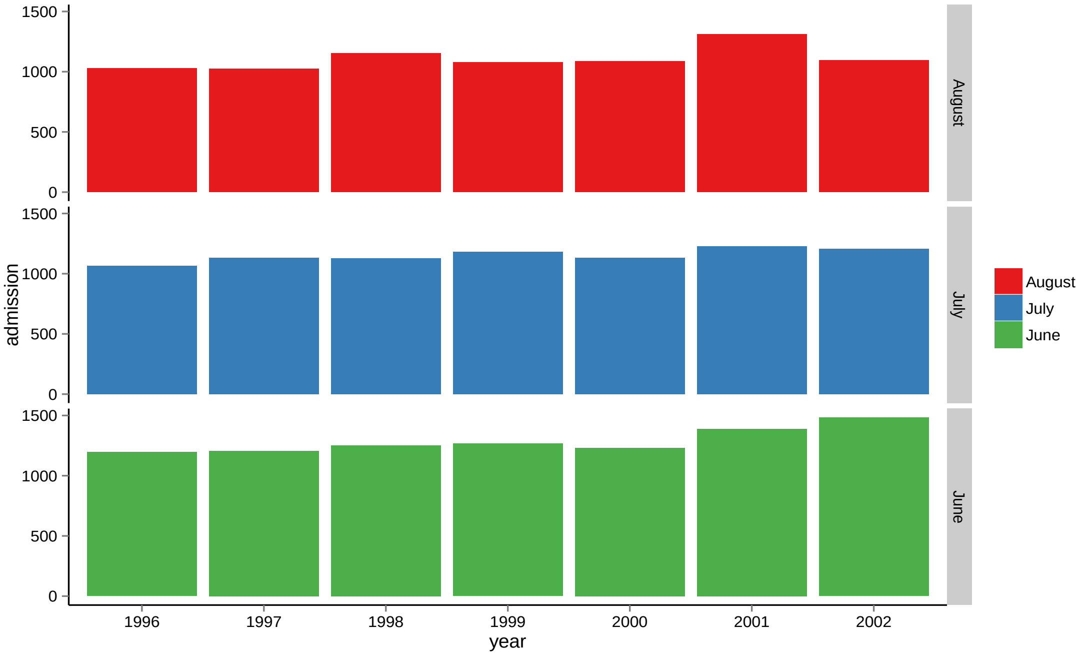

Figure 4.2: The distribution of hospital admission across calendar year and the summer months within the calendar year for city of Milan, Italy.

4.3 The Ozone Functional Data

Using the functional data analysis method (Section 2.2), we shall represent the discrete hourly measurements of ozone data as functions. Let $\tilde{X}_{t1}, \ldots, \tilde{X}_{tJ}$ be the discrete hourly ozone concentrations in the day t for $t = 1, \ldots, T$ measured at daily hours h_j for $j = 1, \ldots, J = 24$. Let the functional representation be denoted by $X_t(h)$, the function is estimated by

$$X_t(h) = \sum_{k=1}^{K} c_{tk}\phi_k(h), \tag{4.1}$$

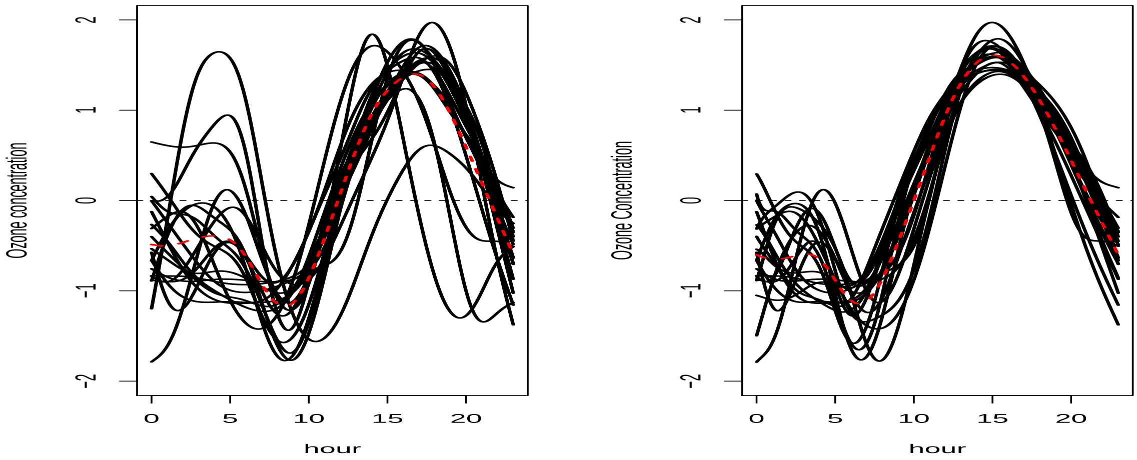

Figure 4.3: Sample of 20 ozone functional data (left) and the same ozone functional data after alignment using hour at which the ozone concentrations of a function is maximum (right).

where $\phi_1(h), \ldots, \phi_K(h)$ are B-spline basis functions, $c_{t1}, \ldots, c_{tK}$ are the coefficients of the B-spline basis associated to day t. Coefficients c_{tk} are estimated using the roughness penalty approach (O'Sullivan, 1986) (Section 2.2.2), selecting the smoothing parameter by generalized cross-validation (GCV, Golub et al. (1979)). We implemented the technique to the Milan ozone data using 18 B-spline of order 5 basis functions (K) and the smoothing parameter λ chosen by GCV is 10. Here, other number of basis functions and order can be selected, but the chosen λ could change to accommodate the difference. The total number of functional observations estimated is 599, which is the number of days included in the study. The left panel of Figure 4.3 shows a sample of 20 functions.

4.4 Aligning Ozone Functional Data

In Section 2.2.3, we presented general discussion about function alignment methods. In this section, we illustrate aligning of the ozone functional observations. Our sample of functions in the left panel of Figure 4.3 display observable common features or landmarks. These common features include the daily maximum and minimum ozone concentrations. The idea is to explore variation with respect to specific features of the ozone curves. We consider the daily maximum as feature of interest, the functions reach maximum concentrations at different daily hours. This indicates that while the overall process of ozone shares common features between days, the time point of the maximum varies from day to day. We are interested to compare the daily variability of ozone concentrations at their respective hour where the concentrations reach maximum. To achieve this, the functions must exhibit their respective peak at the same time point. For instance, if the functions $X_1(h)$ and $X_2(h)$ exhibit their peak at two different hours, then the functions can not be compared at a particular same hour. We then wish to align the functions to have the maximum ozone concentrations at the same time point. We require a target time point to align the maximum of each curve to the target time point. We define the target time point as the average of the hours at which the daily ozone concentrations reach maximum.

Formally, the maximum of each function is aligned to the target time point using $X_t^*(h) = X_t[W_t(h)]$, where $X_t^*(h)$ is the aligned function and $W_t(h)$ is called time-warping function. To ensure the alignment of the maximum of each curve to the target time point, conditions are imposed on $W_t(h)$ (see Section 2.2.3). The right panel of Figure 4 displays the same sample of 20 functional observations after alignment. The functions are aligned approximately at daily hour 3 pm which is the target location at which ozone concentrations reach maximum on the average. These aligned ozone functions will be used as a predictor in the next Section to assess the effect of the aligned ozone on health. In principle, other important features can be selected as a target time point instead of daily maximum. In pollutants and health studies, the average concentrations, threshold and the minimum concentrations can also be considered. In this context, we shall present a model for which the ozone functions are aligned at the daily minimum and average, then the models would be compared with the model which uses alignment at the daily maximum in order to recognize whether the functions are aligned fairly at the most plausible time point to capture the variation observed in the functions.

4.5 Functional Regression Models

The response variable y_t denotes hospital admission in the day t. The Poisson log-linear model which is specified in (2.2) assumes

$$Y_t \sim \text{Poisson}(\mu_t), \quad \text{for} \quad t = 1, \ldots, T \tag{4.2}$$

and uses daily summary of ozone concentration as exposure measure. Instead, we wish to measure daily exposure to the pollutant using a daily ozone function $X_t(h)$ as obtained in (4.1) to allow for the daily pattern of the concentrations. Thus, a functional generalized linear model (FGLM), the extension of the standard generalized linear model (McCullagh and Nelder, 1989), is adopted to the case in which the predictor is functional ozone and the response is scalar hospital admission. The model is given by

$$\log \mu_t = \int_1^J X_t(h)\beta(h)dh, \tag{4.3}$$

where $\beta(h)$ is a functional parameter which describes the association between exposure to ozone measured as function $X_t(h)$ and hospital admission at daily hour h. The parameter $\beta(h)$ gives the weight placed on the predictor $X_t(h)$ at each time of the day in determining the value of the response. For example, the estimate of the value of $\beta(h)$ evaluating at the daily hour 2 pm describes the contribution of the ozone concentrations measured at 2 pm on daily hospital admission. The function $\beta(h)$ is the analogue of the scalar coefficient β in generalized linear model (2.3). To estimate the model, we can specify $\beta(h)$ in the same way as the ozone function $X_t(h)$ using L dimensional B-spline basis

$$\beta(h) = \sum_{l=1}^{L} b_l \phi_l(h), \tag{4.4}$$

where b_1 for $l = 1, \ldots, L$ are the unknown coefficients to be estimated. Once b_l are estimated, $\beta(h)$ is directly recovered by multiplying the estimated b_l with the known basis functions. We estimate b_l using the P-spline penalty approach (Eilers and Marx, 1996), which incorporates combination of B-spline basis and difference penalty. The basic idea is instead of using the integrated second derivative as we have illustrated in the estimation of $X_t(h)$, a simple difference penalty on b_l is imposed. Formally, let the first difference denoted by Δ on b_l is defined as $\Delta b_l = b_l - b_{l-1}$, then the second difference penalty is $\lambda \sum_l \left(\Delta^2 b_l\right)^2$. The smoothing parameter λ still controls the weight of the penalty. Marx and Eilers (1998) suggested to implement P-spline for computational advantages and allows to choose number and position of knots more flexibly.

We now discuss confounding predictors. Particularly, the confounding effect of other functional pollutant or meteorological predictors can be considered. However, we do not pursue multiple functional predictor, since the hourly measurements of other pollutants were not available. We then explore the possible way to include daily summary measures of other pollutants, temperature and seasonal effects in the presence of functional ozone. The correlation matrix in the preliminary analysis suggests that the pair-wise correlations between the pollutants particulate matter smaller than $10\mu g/m^3$ denoted by (PM_{10}), Carbon Oxides (CO), Nitrogen Dioxides (NO_2) and Sulphur Dioxides (SO_2) are statistically significant. Including all or two of them in the same model could lead to multicollinearity problems. Levy et al. (2005) and Bell et al. (2005) recommended to include particulate matter in the study of ozone effect on health and the former suggested any observed relationship between ozone and health outcome may simply reflect particulate matter effect. Meteorological variables have also been known to confound the association between ozone and health. We then included daily maximum temperature as an additive non-linear smooth function $f(\text{Temp})$. The full functional generalized linear model is then given as:

$$\log \mu_t = \int_1^J X_t(h)\beta(h)dh + \alpha \text{PM}_{10(t)} + f(\text{Temp}_t) + \sum_{j=1}^{6} \delta_j \text{DOW}_{tj} + \sum_{k=1}^{6} \gamma_k \text{Year}_{tk}, \tag{4.5}$$

where,

$$\text{DOW}_{tj} = \begin{cases} 1 & \text{if day } t \text{ is day of the week } j\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

$$\text{Year}_{tk} = \begin{cases} 1 & \text{if day } t \text{ is in the year } k\,, \\ 0 & \text{otherwise.} \end{cases}$$

The smooth non-parametric function $f(\text{Temp})$ is modeled as a linear combination of a cubic B-spline basis. Detailed treatment of the estimation of this type of function is given in Wood (2006). The other non-functional covariates day of the week (DOW_j) and calendar year (Year_k) are included as linear terms. Since day of the week is defined as factor variable, it has given levels for its category, in which Monday $= 1, \ldots,$ Saturday $= 6$ and Sunday is the baseline. Similarly, calendar year at which the data was collected, 1996-2002 considered as factor and takes levels $1, \ldots 7$, and year 1996 considered as baseline. The parameter α measures the effect of PM_{10}, δ_j describes the effect of day of the week j relative to Sunday and γ_k is effect of year k compared to the effect in year 1996.

Often, the effect of exposure to ozone persists for some days from the date of exposure. To account such persistence, measures of pollutants are lagged by a number of days. In the context of functional ozone modeling, the use of lagged values to predict the current day admission is even more relevant, in the sense that the dependence of the same day hospital admission on ozone exposure measured hourly might not be logical if the time of hospital admission precedes the portion of the daily ozone function that is identified as potentially harmful. As for what number of days to lag, Zeger et al. (2000) suggested anywhere between zero and five days is appropriate. We consider the lagged values of hourly ozone measure for up to three days. As a remark, more complex modelling strategies were initially tried in order to identify the best model in terms of goodness of fit. These strategies include the use of lagged values over the previous three days for PM_{10} and daily maximum temperature. Such strategies did not significantly improve the goodness of fit. However, we included one day lag PM_{10} denoted by PM_{10}^{lag} to explain the current day hospital admission. The pair-wise correlations between PM_{10}^{lag} and the other explanatory covariates are relatively small.

The functional generalized linear model specified in (4.3) imposes a linearity assumption on the dependence of hospital admission on functional ozone which may be too restrictive. We wish to relax the linearity assumption and estimate a flexible non-linear shape of ozone effect. There are few examples of additive structures being used in a functional data setting. Müller and Yao (2008) and James and Silverman (2005) proposed a functional additive model in which the functional predictor is decomposed through functional principal components. Mclean et al. (2014) presented the same approach without decomposing the functional predictor. Model (4.3) can be specified as functional additive form

$$\log \mu_t = \int_1^J F\{X_t(h), h\} dh. \tag{4.6}$$

The function $F(X, h)$ is specified using tensor product B-splines (Wood, 2006). That is, $F(X, h)$ is treated as a bivariate function constructed using the marginal B-splines corresponding to X and h so that

$$F(X, h) = \sum_{k=1}^{K} \sum_{l=1}^{L} c_{kl} \theta_l(h) \phi_k(x),$$

where $\theta_l(h)$ for $l = 1, \ldots, L$ and $\phi_k(x)$ for $k = 1, \ldots, K$ are B-spline bases, c_{kl} are the unknown coefficients to be estimated. Here again the P-spline approach is adopted to estimate $F(X, h)$. The P-spline penalty is even more appealing in this case, since we want computational efficiency to smooth c_{kl} which requires row and column penalties in the direction of X and h respectively. Furthermore, the use of P-splines for additive models allow any degree of B-spline to be used with any order of difference for the penalties, offering greater flexibility (Marx and Eilers, 1998). A known issue in the use of tensor products of B-spline is that certain data regions may not have any observations, which prohibits estimating the coefficients (Fahrmeir et al., 2013). To avoid this issue, $X(h)$ is transformed using empirical cumulative distribution function (ecdf). A further advantage of the transformation is that we take the connection between quantile and ecdf to obtain interpretable estimate. Thus, we can interpret the estimated shape of $F(p, h)$ as the effect of $X(h)$ being at its pth quantile.

The analysis was performed using the freely available R software. The hourly ozone data was first smoothed and changed to functional data using *fda* package (Graves et al., 2009). The parameters of the functional generalized linear model and the additive version were obtained following the version of penalized iteratively re-weighted least squares (P-IRLS).

4.5.1 Results

First we fitted the model specified in (4.5) using the summer time functional ozone exposure measure as a predictor and the number of daily admissions as scalar response. For the selected non-functional confounding predictors, we fitted four models with different numbers of lagged days for the functional ozone. The parameter of interest $\beta(h)$ is estimated and 95% point-wise confidence intervals are computed to describe the uncertainty in the estimate. The estimate of the parameter $\beta(h)$ together with 95% point-wise confidence intervals for each model is shown in Figure 4.4. These estimates are achieved using the P-splines approach using 8 B-spline basis functions of order 3 and second order difference penalty. The confidence intervals of the estimated coefficient curve for the current day ozone exposure (lag 0) involves zero almost throughout the day, which suggest the effect of the pollutant is not significant. The estimate for one day lagged values of ozone (lag 1) is significant mainly after the daily hour 3 pm, which indicates that exposure to ozone in the previous day is associated to the current day hospital admission. These 95% point-wise confidence intervals for the estimate of $\beta(h)$ from two days lagged values of ozone (lag 2) and three days lagged ozone (lag 3) involve zero in the majority of the regions, but they show a significant effect in the night hours. The overall patterns of the estimate of $\beta(h)$ for lag 2 and lag 3 ozone are the same. This may indicate that the persistence of ozone effect may last up to three days from date of exposure. We fitted a model using 7 days lagged ozone exposure to investigate if the estimate would be different from the estimate of lag 2 and lag 3. The resulting estimate is very close to that of the estimate from lag 3 ozone.

The estimated values (associated standard deviation in brackets) of confounding predictors that were included as linear components are shown in Table 4.1. The current day PM_{10} and its one day lag PM_{10}^{lag} are significantly associated with daily number of hospital admission. The factor covariates day of the week and calendar year are both globally significant. To select the best lag among the estimated models, we compared the goodness of the models using the Unbiased Risk Estimate (UBRE) criterion (see Section 3.3.2). The UBRE scores given in Table 4.1 show that there is slight difference among the models, lag 1 ozone exposure produced lower UBRE score compared to the other lags. Thus, the dependence of hospital admission on

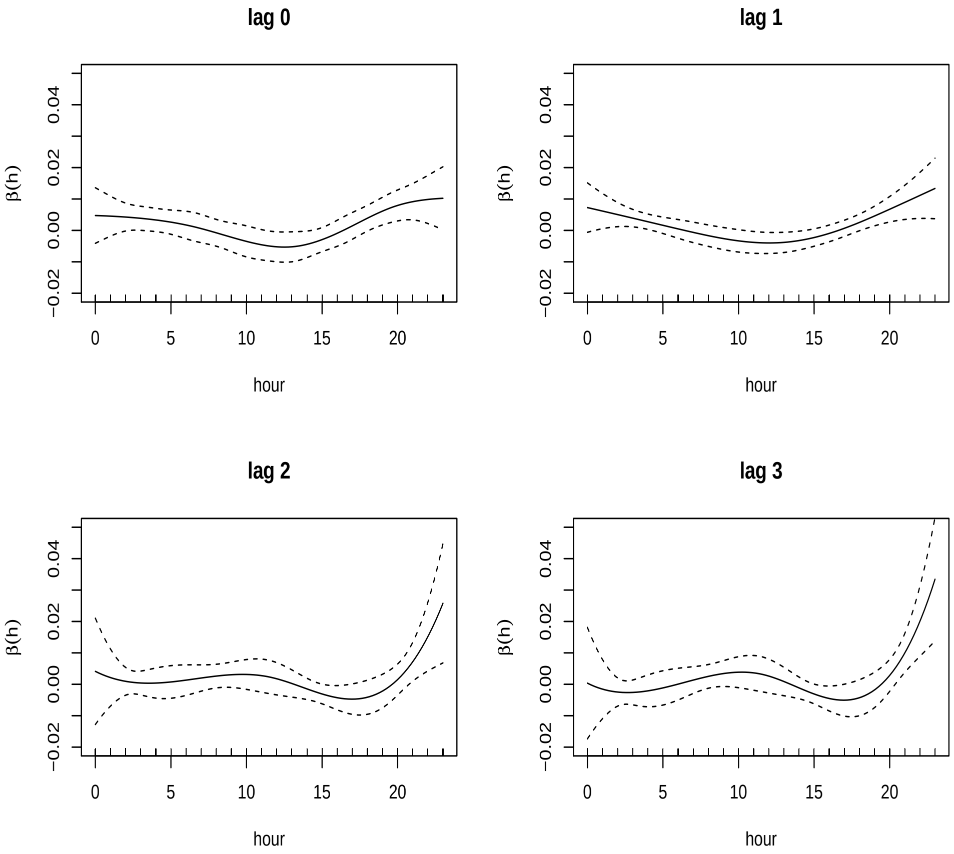

Figure 4.4: Estimate of ozone functional coefficient, $\beta(h)$, under the functional generalized linear model. The response is the current day hospital admission for each lag and the predictor is the functional ozone with different lag controlling other confounding predictors.

	The functional ozone with the following lags:			
	lag 0	lag 1	lag 2	lag 3
PM_{10}	0.025^{*}	0.024^{*}	0.024^{*}	0.026^{*}
	(0.008)	(0.007)	(0.008)	(0.008)
PM_{10}^{lag}	0.019^{*}	0.020^{*}	0.020^{*}	0.021^{*}
	(0.008)	(0.075)	(0.007)	(0.008)
Observations	599	598	597	596
UBRE	0.511	0.504	0.516	0.523

Table 4.1: Results for predictors with linear components under the functional linear regression models using different lag for the functional ozone. The estimates (associated standard deviation in the bracket) are given for particulate matter (PM_{10}) and its one day lag (PM_{10}^{lag}).

* indicates the significance of the predictor .

UBRE: Unbiased Risk Estimate.

one day lag functional ozone is selected as best. We then use the lag 1 model for further model improvement, particularly to estimate the aligned ozone effect and the functional additive model.

A second model estimate is considered for the aligned ozone functions. Aligning the functions allows us to explore variation at specific time point and we are interested to capture the portion of ozone curves near maximum that can be potentially harmful to health (see Section 4.4), and improve the estimate of the functional coefficient $\beta(h)$ using the aligned ozone curves at the daily maximum as predictor. Figure 4.5(a) presents the estimate of

$\beta(h)$ for which the ozone curves were aligned at the daily maximum. some portion of the confidence intervals does not involve zero. Particularly, the region of the estimated curve after the daily maximum (3 pm) is significant. Interestingly, this indicates the region of the estimate of $\beta(h)$ starting at the daily maximum, where the original ozone curves were aligned, is significantly associated to hospital admission.

According to epidemiological studies of exposure, the maximum level reached during the day is likely to cause health issues. Examining the association between ozone functions which are aligned at the daily maximum is therefore quite reasonable. Nevertheless, to verify whether results would have been better if the curves had been aligned at different points, we estimated $\beta(h)$ by aligning the ozone functions at the daily minimum ozone concentration and at the daily average concentration. The idea is to check whether there is a feature which is better than the daily maximum to capture the portion of the concentrations curve that is harmful to health. The estimate of $\beta(h)$ in panel b and c of Figure 4.5 are obtained by aligning the original functions at the daily minimum and at the daily average respectively. The former estimate leads to wide confidence intervals in the whole region and the latter estimate is significant mostly the evening hours. Apparently, the daily maximum is the best feature to align the ozone concentrations curves and it leads to obtain an estimate of ozone effect $\beta(h)$ which clearly points out relevant features of the daily pattern as compared to the daily minimum or daily average.

We notice that the results in this Section are coherent with those obtained using the standard methods. The classical generalized additive model is fitted using each hourly measurements of ozone concentrations as predictor and hospital admission as response. Since the model is fitted at each daily hour, we have 24 models which constitute various model information including the estimated coefficient and the unbiased risk estimate (UBRE) for each model.

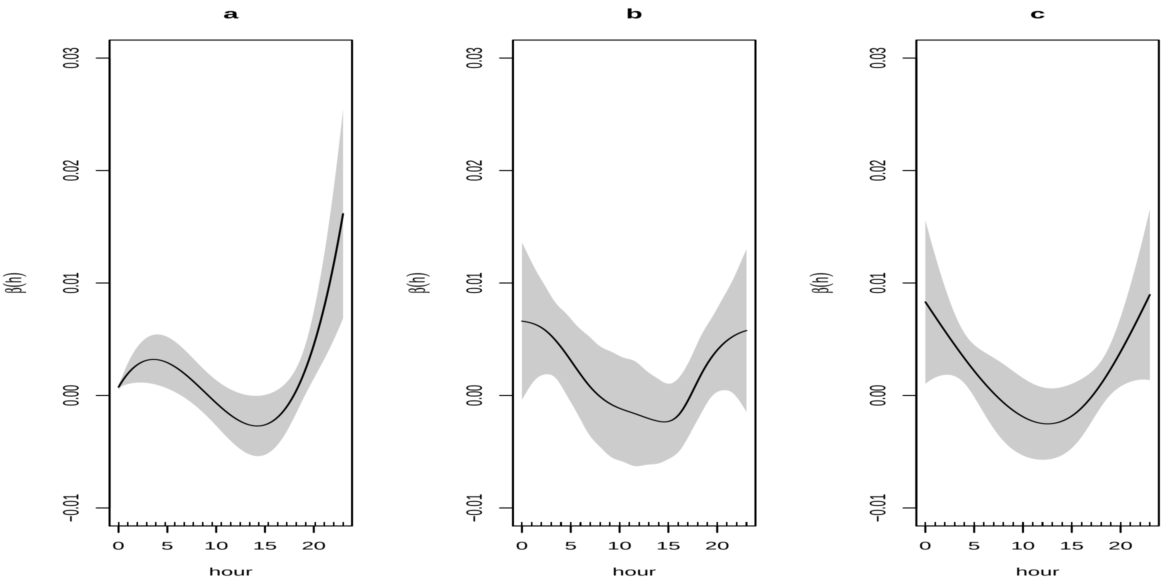

Figure 4.5: Estimated coefficient curves of ozone under the functional generalized linear model using aligned ozone curves at daily maximum (a), at daily minimum (b) and at daily average(c).

These results are shown in Figure 4.6, the UBRE score is minimum at the evening hour which suggests the model fitted at the evening hour come out as best. The left panel shows the estimated coefficients with associated 95% confidence intervals. The estimated coefficient is significant at each daily hour, but the estimated effect is maximum at evening hour. The Figure includes two more models fitted to the daily average and daily maximum of ozone concentrations (the vertical lines, in the far right) to compare with the results of the hourly estimates. The daily maximum produced higher estimate than the hourly measures, but once again, the UBRE indicates the model with the daily maximum may not be the preferred model. More empirical comparisons of the classical approaches and the functional regression method are given in Section 4.5.2.

The model we estimated so far assumes linearity in the dependence of func-

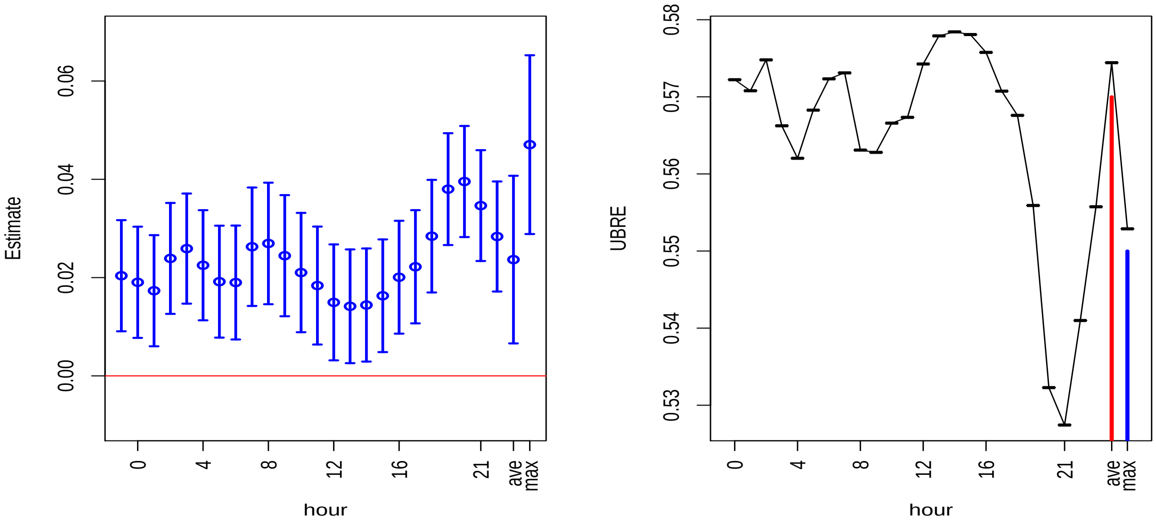

Figure 4.6: The estimates of ozone effect at each daily hour (left) and the Unbiased Risk Estimates (UBRE) (right) under the standard generalized additive model. For comparison, results for the daily average (ave) and maximum (max) are displayed at the far right side using vertical lines.

tional ozone $X(h)$ on daily hospital admission. We now estimate a non-linear flexible shape of $F(X, h)$ as given in (4.6). Following our discussion of Section 4.5, we assign 6 cubic B-spline with second order difference penalty for the ozone axis, and 5 cubic B-spline with second order difference penalty for the h axis. These marginal B-spline bases are used to construct tensor product spline bases with dimension 30. Figure 4.7 presents the estimated shape $\hat{F}(p, h)$ where the ozone concentrations X is given at its quantile p. High level of ozone effect (red color) observed as the quantile increases from 0.6 on wards and as the daily hour goes to the evening. This is generally coherent with the one dimensional coefficient estimate $\beta(h)$ where the region of the evening hours are more detrimental (Figure 4.4). We verified that the results are insensitive to the change in number of B-spline used. However, different orders of the penalties produce slightly different results. To find the best model based on UBRE criterion, we refitted the model varying the order of row and column penalties. The model with order 2 for both the ozone and daily hour penalties produced smaller UBRE score, and as such is chosen as the final model.

The non-functional confounding predictors are also estimated. The daily maximum temperature is estimated as non-linear smooth function using the nonparametric method. This estimate is shown in Figure 4.8, and it was achieved using cubic regression spline basis with dimension 10 for which the actual effective degrees of freedom estimated from the model are 8.254. The shape of the estimate is non-linear U-shaped in the full range, but wide confidence bands are observed in the low temperature region where the effect might not be significant. The estimates of the factor predictors day of the week and calendar year are given in Table 4.2. The parameters are significantly different from zero.

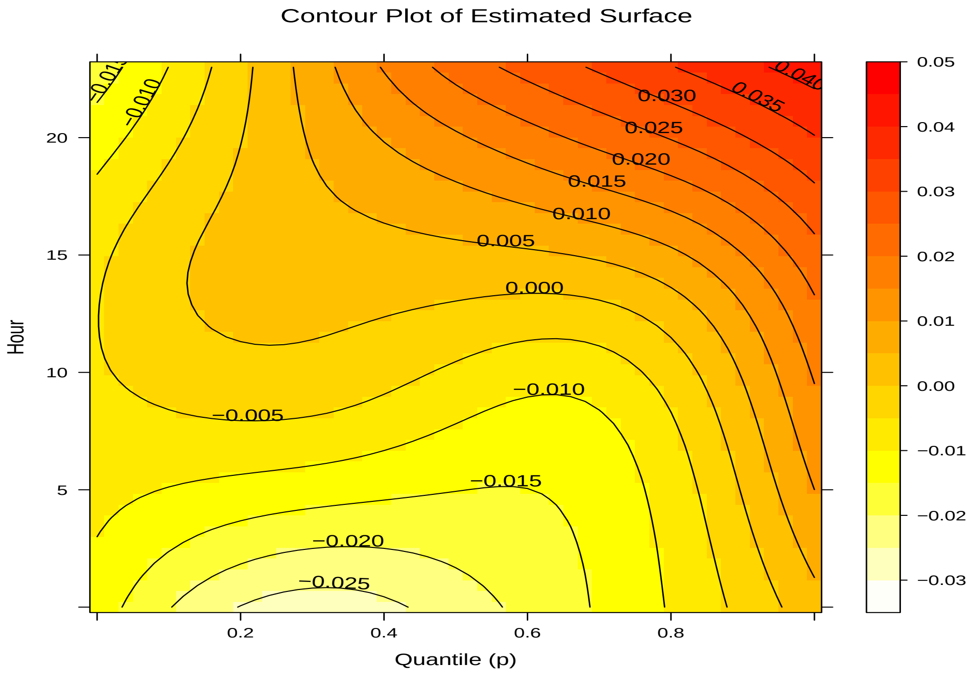

Figure 4.7: The estimated contour shape of $F(p,h)$ under functional additive model.

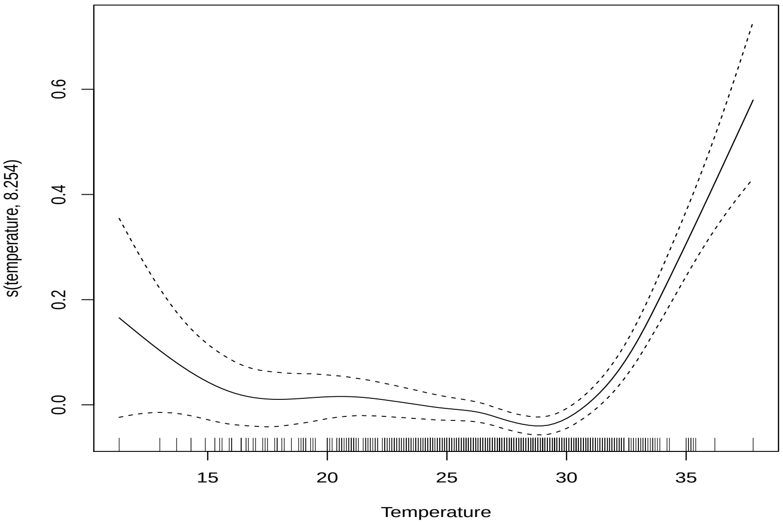

Figure 4.8: The effects of daily maximum temperature estimated as non-linear smooth function.

	Estimate	S.E	Estimate/S.E	P-value
PM_{10}	0.025	0.009	2.828	0.004
PM_{10}^{lag}	0.020	0.008	2.367	0.018
Day of the week				
Monday	-0.056	0.025	-2.238	0.025
Tuesday	-0.120	0.025	-4.758	0.000
Wednesday	-0.074	0.025	-2.960	0.003
Thursday	-0.0714	0.025	-2.812	0.004
Friday	-0.161	0.026	-6.192	0.000
Saturday	-0.266	0.026	-10.248	0.000
Calendar Year				
1997	0.011	0.027	0.396	0.692
1998	0.055	0.026	2.110	0.034
1999	0.104	0.027	3.859	0.000
2000	0.043	0.027	1.589	0.112
2001	0.158	0.025	6.166	0.000
2002	0.139	0.026	5.181	0.000

Note: S.E=standard deviation.

Table 4.2: Estimated coefficients and standard deviations of non-functional linear components.

4.5.2 Out-of-Sample Predictive Performance

The objective in this Section is to investigate if our approach to measure exposure in terms of function is superior to the standard generalized additive models which are simpler approaches that use daily summaries. The functional regression method we employed is more complex and computationally intensive, thus, we compare the predictive accuracy of the proposed method against simpler approaches. We selected four exposure models to compare with our approach, two of these use the daily average and maximum exposure measure. For more broader comparison, we included two more ozone exposure measure of Chiogna and Pauli (2011), these are (1) the difference between day time maximum hourly concentrations and the threshold level, and (2) night time average concentrations. The authors showed that their approaches capture the daily variability of ozone better than the standard daily summary measures. However, such approaches still rely on some form of daily summaries of the pollutant, since the approaches are computed from the hourly measurements. We denote these two measures as GAM-CP1 and GAM-CP2 respectively for reference. Our approach of functional regression exposure model is given in three scenarios, these are functional regressions based on: un-aligned ozone functions (we refer as FGLM-O), aligned ozone function (FGLM-A) and additive structure (FGAM).

For prediction, we split the period of study 1996-2002 in two sub periods: data for years 1996-1999 is used as training set and data for 2000-2002 is used as validation set. The basic idea is to use the first four years data to predict daily hospital admission for the latter three years and then to compute the out-of-sample residual mean squared error (RMSE), which is given in the form

$$\text{RMSE} = \left[254^{-1} \sum_{t \in \{\text{valid}\}} (y_t - \hat{y}_t)^2 \right]^{1/2} .$$

Exposure model	RMSE
GAM-Average	33.391
GAM-Maximum	33.392
GAM-CP1	28.415
GAM-CP2	28.419
FGLM-O	19.034
FGLM-A	19.021
FGAM	19.001

Table 4.3: Out-of-sample RMSE for different ozone exposure models including the functional regressions approach.

The training set contains 345 days and the validation set has 254 days. These sets of days correspond to the number of curves for functional regression models. We report results of RMSE in Table 4.3 for each model under consideration. In this context, the RMSE of a model indicates the performance of the model in predicting hospital admission in the validation group using the training group. The smaller the RMSE, the better is the model in predicting the response. The functional regression models in Table 4.3 have better predicting ability compared to the other approaches. Here, functional regression based on aligned functions has improved the forecasting slightly in terms of RMSE measure. Apparently, exposure models based on daily summary measures perform worse than other approaches.

4.6 Discussion

In this Chapter, we propose modeling the short-term effect of daily ozone concentrations on health using a functional representation of daily ozone concentration as a predictor using data from the city of Milan, Italy. This method is based on the principles of functional data analysis. The approach has various advantages, particularly, the daily temporal variability of ozone concentrations is accounted and the model aids to detect underlying patterns and features such as the hours at which ozone reaches minimum or maximum. The observed ozone functions are aligned using the hour at which the daily ozone measurements is maximum to capture the important variability observed in the functions as discussed in Section 4.4.

A functional regression model in which the response is scalar daily hospital admission and the predictor is the functional ozone is fitted controlling for pollutant particulate matter, day of the week, calendar effects and weather conditions. The persistent effect of exposure to ozone is modeled by taking the lagged values of the hourly measurements for up to three days. The results show the dependence of the current day hospital admission on previous day ozone exposure. The linear functional form of ozone coefficient is estimated for both un-aligned and aligned ozone functions. The estimated coefficient from the un-aligned ozone functions is significant mainly in the evening hours while the aligned ozone functions produce significant estimate in the day hours when the daily ozone concentrations level reaches near maximum. The region of the estimate near daily maximum is therefore identified as potentially harmful to health. A flexible non-linear shape of the ozone effect is estimated by relaxing the linearity assumption imposed on the dependence of hospital admission on ozone measured by functional form. The proposed methods improve the simpler alternatives based on scalar summaries of daily ozone concentration in terms of prediction.

Before closing this Chapter, we illustrate limitations. Spatial variability in addition to temporal variability that may exist in ozone presents another modeling challenge that we did not pursue. In the next Chapter, we shall present a model to address the limitation of this Chapter. Particularly, we propose a functional regression model which is suitable to deal with the issue of spatial heterogeneity.

Chapter 5

Functional Hierarchical Model for Multi-city Data

In the previous Chapter we considered a functional regression model to study the effect of daily pattern of pollutant ozone on health for the city of Milan. In this Chapter, we extend the functional regression method to model multi-city data which come from different geographic locations allowing for spatial heterogeneity. Thus, we wish to provide a picture of regional variation in the health effect of ozone over the study region. The Chapter is organized as follows, Section 5.1 provides a brief motivation of the work. Section 5.2 describes the multi-city data set. Section 5.3 presents the modeling approach: two model frameworks are presented. The first is the overall model, which is used to fit a functional regression model by pooling the different location data together, thus ignoring geographic variability in the pollutant effect. The second framework illustrates the more general functional hierarchical model which accounts for spatial heterogeneity. Section 5.3.3 discusses the results from the fitted models and finally Section 5.4 presents concluding discussion.

5.1 Motivation

In this Chapter, we propose a functional hierarchical modeling approach in the Bayesian paradigm using Markov Chain Monte Carlo to estimate ozone effect allowing for the daily variation of the concentration and the spatial heterogeneity in the estimated effect at once. The approach allows us to estimate the overall and location-specific ozone effect as a function of daily hour. our approach is easily generalized to a generic case of regression with multilevel structure and functional covariate.

5.2 Data

The database contains information about mortality, hourly ozone measurements, meteorological variables, and seasonal data for 15 cities in the USA for the summer periods of the years 1987-2000 (Table 5.1). Hourly ozone concentration measurements (expressed in parts per billion, ppb) were obtained from the American Environmental Protection Agency's (APA) Aerometric Information Retrieval System (AIRS) and AirData System. For those cities where measurements are available in more than one site, we take the average. Some typical daily patterns of ozone concentration are shown in figure 5.2. Daily mortality, weather and seasonal variables were obtained from the NMMAPSdata R package (Peng et al., 2004), originally assembled as part of the National Mortality, Morbidity, and Air Pollution Study (NMMAPS).

For the response variable, we consider daily death counts for each city, excluding accidental deaths. As far as confounders are concerned, it is customary to allow for meteorological conditions, in particular the temperature, and possibly other pollutants concentrations. After extensive preliminary analysis, which includes consideration of various daily summaries of meteorological variables, we used the 24 hour maximum temperature for each day. Other pollutants are often included in the study of ozone exposure,

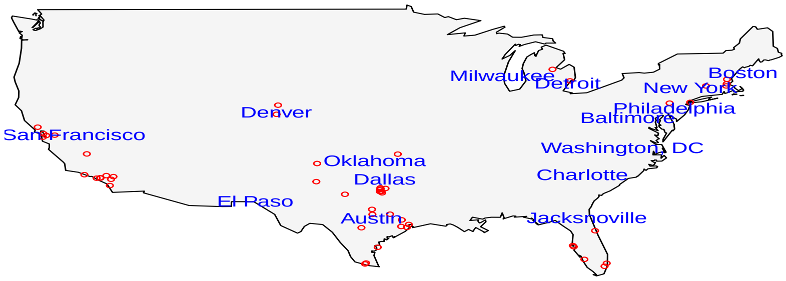

Figure 5.1: Map of the 15 USA cities selected for the study. Positions of the names of the cities correspond to their geographical locations.

particularly, Particulate Matter (PM) may act as confounder of the association between ozone and health outcome. However, we do not consider the effect of Particulate Matter, because measurements for Particulate Matter were available once every six days, which led to missing measurements in the majority of the summer periods under consideration.

5.3 Modeling Approach

We consider two modeling approaches, the first approach is an overall model which fits the same functional regression model for all the cities. That is the model is estimated by pooling all cities together, assuming a spatially homogeneous association between exposure to ozone and mortality. In the second approach, the estimate of ozone effect is assumed to vary across the cities, and then city-specific effects are estimated.

City	symbol	daily deaths on average	total deaths	number of days
Austin	aust	2.398	2087	870
Baltimore	balt	7.170	6439	898
Boston	bost	3.497	2087	907
Charlotte	char	2.783	3181	1143
Dallas	dal	15.841	18534	1170
Washington dc	dc	5.401	4888	905
Denver	den	4.986	6079	1219
Detroit	det	14.466	10445	722
El Paso	elp	2.283	2270	994
Jacksonville	jack	3.958	4240	1071
Milwaukee	mil	4.805	6036	1256
New York	new	55.256	47023	851
Oklahoma	okl	3.499	4434	1267
Philadelphia	phil	11.850	13462	1136
San Francisco	sanf	5.365	6905	1287

Table 5.1: Summary information of mortality for 15 USA cities collected in the summer periods (June-July-August) of the years 1987-2000.

5.3.1 Overall Model

Let y_{tc} represent the number of mortality caused by pollutant related issues of day t in city c for a pooled data from 15 different cities. The daily ozone concentration curve $X_{tc}(h)$ is estimated using the smoothing technique from its discrete counterparts $\tilde{X}_{tc}(h_1), \ldots, \tilde{X}_{tc}(h_J)$ as in Section 4.3

$$X_{tc}(h) = \sum_{l=1}^{L} c_{tcl}\theta_l(h),$$

where coefficients $c_{tcl}, \ldots, c_{tcL}$ are estimated using smoothing techniques

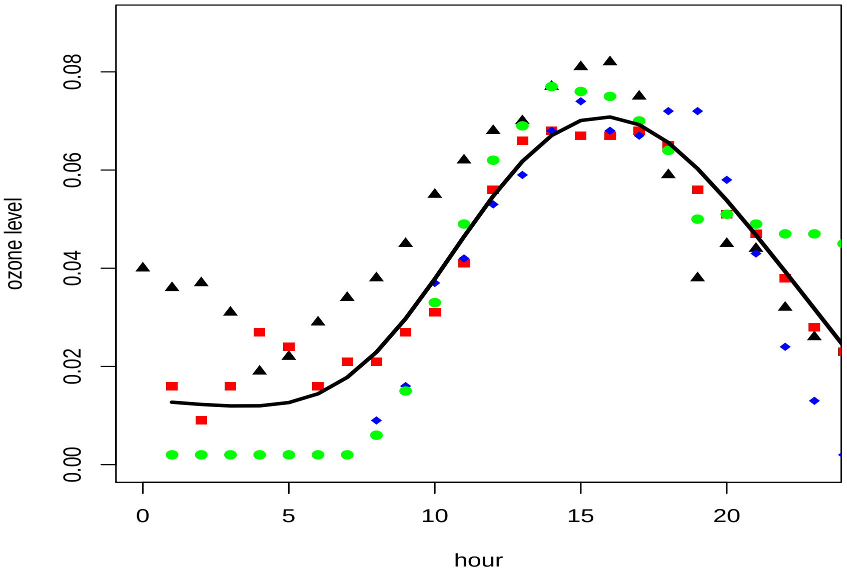

Figure 5.2: The hourly ozone measurements of 4 days for Washington DC in the summer of years 1990-2000. The average over the 4 days is the black solid line.

and $\theta_1(h)\ldots,\theta_L(h)$ are known B-spline basis functions. We employ the functional regression model described in Section 4.5 as

$$\begin{aligned} Y_{tc} &\sim \text{Poisson}(\mu_{tc}) \quad \text{for} \quad t=1,\ldots,T \\ \log\mu_{tc} &= \int_1^J X_{tc}(h)\beta(h)dh + f_c(\text{Temp}_{tc}) + \sum_{s=1}^{6}\delta_s\text{DOW}_{ts} + \sum_{k=1}^{13}\gamma_k\text{Year}_{tk}. \end{aligned} \tag{5.1}$$

The parameter of interest $\beta(h)$ measures the effect of the pollutant at specific time point h, and is estimated from the pooled data. Daily maximum temperature is included as a non-linear smooth function $f_c(\text{Temp}_{tc})$ for city c. Day of the week and calendar year are denoted by a sets of indicator variables (DOW) and (Year) respectively. To estimate the model, we need to represent $\beta(h)$ as a linear combination of known B-spline basis functions $\phi_1(h), \ldots, \phi_K(h)$

$$\beta(h) = \sum_{k=1}^{K} b_k \phi_k(h).$$

The expression for $X_{tc}(h)$ and $\beta(h)$ are substituted in equation (5.1), leading to

$$\log \mu_{tc} = \int_1^J \sum_{l=1}^{L} c_{tcl}\theta_l(h) \sum_{k=1}^{K} b_k\phi_k(h)dh + f_c(\text{Temp}_{tc}) + \sum_{s=1}^{6} \delta_s \text{DOW}_{ts} + \sum_{k=1}^{13} \gamma_k \text{Year}_{tk}.$$

With this model we obtain a marginal effect $\beta(h)$ and ignore the possibility of spatial heterogeneity in the estimated effect.

5.3.2 Functional Hierarchical Model

In this Section, we illustrate the functional hierarchical regression model designed to estimate coefficients $\beta(h)$ specific for each city and a pooled functional coefficient averaged over the cities. The model shall be fitted in the Bayesian paradigm using Markov Chain Monte Carlo simulation. We then modify specification (5.1) in the form

$$\log \mu_{tc} = \int_1^J X_{tc}(h)\beta_c(h)dh + f_c(\text{Temp}_{tc}) + \sum_{s=1}^{6} \delta_s \text{DOW}_{ts} + \sum_{k=1}^{13} \gamma_k \text{Year}_{tk}, \tag{5.2}$$

where $\beta_c(h)$, for $c = 1, \ldots, 15$, are location-specific functional coefficients. In order to estimate the model, we need a representation for the concentration curve $X_{tc}(h)$ and the associated parameter $\beta_c(h)$. To this end, we employ Functional Principal Component Analysis (FPCA, see Section 2.2.4) to express $X_{tc}(h)$ as a linear combination orthogonal eigenfunctions which in turn is obtained by spectral decomposition of the covariance function. Let

$K(s, h)$ be the overall covariance function of summer time ozone concentration for all the days in the study. We assume that $K(s, h)$ is the same across the cities, and can be given by $K(s, h) = \mathbb{C}ov\{X_t(s), X_t(h)\}$. Where $X_t(h)$ is a smooth ozone concentration curve common to all the cities. Then, there is an orthogonal expansion of $K(s, t)$ using eigenfunctions $\{\psi_l(.)\}_{l=1,2...}$ and non-increasing eigenvalues $\{\lambda_l\}_{l=1,2...}$ in the form

$$K(h, s) = \sum_{l=1}^{\infty} \lambda_l \psi_l(h)\psi_l(s), \quad h, s \in [1, J]. \tag{5.3}$$

The covariance function $K(h, s)$ can be consistently estimated using different methods, for instance, the method of moments (Di et al., 2009) and iterative penalized spline (Yao and Lee, 2006). Figure 5.3 shows the smooth estimate of the overall covariance function $K(h, s)$ obtained by the method of moments for the summer time ozone concentration of USA cities. Using Karhunen-Loeve representation (Karhunen (1947); Loeve (1965)), the orthogonal eigenfunctions $\{\psi_l(.)\}_{l=1,2...}$ are assumed as basis functions to express the function $X_{tc}(h)$ as

$$X_{tc}(h) = \sum_{l=1}^{\infty} \left(\int_1^J \psi_l(h) X_{tc}(h) dh \right) \psi_l(h) = \sum_{l=1}^{\infty} \xi_{tcl} \psi_l(h), \tag{5.4}$$

where the coefficients

$$\xi_{tcl} = \int_1^J \psi_l(h) X_{tc}(h) dh, \tag{5.5}$$

are uncorrelated random variables with zero mean and variances λ_l. Within principal component analysis, these random variables are called principal component scores or loadings. Hence, city-specific ozone concentration curves $X_{tc}(h)$ are linear combinations of the overall orthogonal eigenfunctions $\psi_l(h)$ with city-specific weights ξ_{tcl}. In practice, only a small number of eigenfunctions are sufficient to approximate $X_{tc}(h)$. Based on the share of explained variance, the first 8 eigenfunctions capture more than 99% of the observed functional variability (Table 5.2).

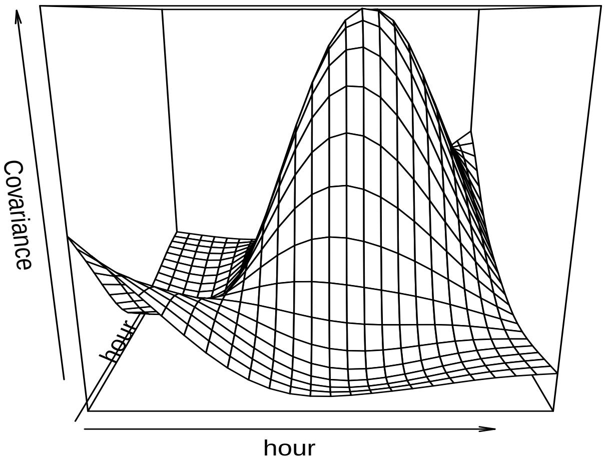

Figure 5.3: Estimate of the covariance function $K(s,h)$ using data from 15 USA cities for the summer ozone concentration.

	Eigenvalues							
	1	2	3	4	5	6	7	8
Var (10^{-3})	4.8	1.18	0.64	0.32	0.19	0.11	0.06	0.04
%Var	65.20	15.83	8.68	4.29	2.57	1.44	0.83	0.54
Cum.Sum	65.20	81.02	89.7	93.99	96.56	98	98.83	99.37

Table 5.2: The estimated eigenvalues for functional ozone computed from the hourly summer time ozone concentration.

The city-specific functional parameters $\beta_c(h)$ are specified in terms of a cubic K dimensional B-spline basis functions $\phi_1(h),\ldots,\phi_K(h)$ with equally spaced knots, common to all the cities. It follows

$$\beta_c(h) = \sum_{k=1}^{K} b_{ck}\phi_k(h) \quad \text{for} \quad k = 1,\ldots,K, \tag{5.6}$$

Thus, using specifications (5.4) for $X_{tc}(h)$ and (5.6) for $\beta_c(h)$, the expression in the integral of (5.2) can be further modified as

$$\begin{aligned} \log(\mu_{tc}) = \int_1^J X_{tc}(h)\beta_c(h)dh = \int_1^J \left[\sum_{l=1}^{L} \xi_{tcl}\psi_l(h)\right]\left[\sum_{k=1}^{K} b_{ck}\phi_k(h)\right] dh \\ = \boldsymbol{\xi}'_{tc}\mathbf{J}\mathbf{b}_c \quad \text{for} \quad \mathbf{J} = \int_1^J \psi_l(h)\phi_k(h)dh, \end{aligned} \tag{5.7}$$

where $\boldsymbol{\xi}'_{tc} = (\xi_{tc1}, \dots, \xi_{tcL})$ denote the $T \times L$ dimensional matrix of principal scores for city c, $\mathbf{J}$ is $L \times K$ dimensional matrix computed by numeric integration prior to model fitting and $\mathbf{b}_c = (b_{c1}, \dots, b_{cK})'$ are city-specific parameters of interest. For identifiability constraint, $K \leq L$. We assume the prior distribution for the coefficients

$$\begin{cases} \boldsymbol{b}_c \sim \text{Normal}(\boldsymbol{u}, \sigma_b^2 I) \\ \boldsymbol{u} \sim \text{Normal}(0, \sigma_u^2 I), \end{cases} \tag{5.8}$$

where $\mathbf{u} = (u_1, \dots, u_K)$ are pooled parameters, representing the overall effect. With this modeling approach, each city has its own coefficients, and their mean represent the overall effect, inferences can be made for the different city coefficients and the population average of the coefficients along with variance parameters. The variance parameters, σ_b^2, σ_u^2 are measures of between city variability and the average variability respectively. For these variance components, a gamma distribution with mean one and large variance is used as a weakly informative prior. The weakly informative priors are also used to specify the prior distributions of the parameters associated to the confounding covariates, in particular, normal prior distributions with mean zero and large variance are used. Two different methods can be employed to deal with $\boldsymbol{\xi}_{tc}$: the first approach is to obtain a relevant set of principal components from the covariance operator $K(s, h)$ (as outlined in Section 2.2.4) and then use the estimate as an input to the hierarchical model. This approach is the most commonly used and requires to estimate

$\boldsymbol{\xi}_{tc}$ by numerical integration. The second approach suggested by Crainiceanu and Goldsmith (2010) is to estimate at once both the principal component scores and the coefficients from the model. Thus, we incorporate estimation of the principal scores by treating them as additional parameters with prior distribution $\boldsymbol{\xi}_{tc} \sim \text{Normal}(0, \text{diag}(\lambda_1, \ldots, \lambda_L))$.

5.3.3 Results

Posterior distributions of model parameters are explored using Markov Chain Monte Carlo (MCMC) simulation. We first consider estimates from the overall model (Section 5.3.1). Relatively short chains are deemed sufficient to grant convergence. Figure 5.4 depicts the estimate of functional coefficient $\beta(h)$ together with 95% credible intervals, and shows a strong exposure effect for concentrations in the late afternoon, coherent with our findings (Arisido, 2014) for the city of Milan.

For the second model (Section 5.3.2), which allows for heterogeneous coefficients, 10,000 iterations are needed to converge, we discarded the first 5,000 iterations as burn-in and used the reminder 5,000 for posterior inference. The convergence of the model is assessed using the standard Bayesian diagnostic tools and there were no apparent convergence or mixing issues. Initially, some autocorrelations were detected, but when the simulation number was increased, the autocorrelations were lowered to the minimum.

Figure 5.5 reports the average of the posterior samples and the associated 95% credible intervals for the city-specific functional coefficients $\beta_c(h)$ and the pooled functional coefficient. The pooled coefficient is a synthesis of information from the 15 cities and it shows some interesting features. Particularly, the 95% credible bands do not involve zero from 3 pm on wards, coherent with estimate from the overall model (Figure 5.4). Here, the pooled functional coefficient estimated from functional hierarchical model produced

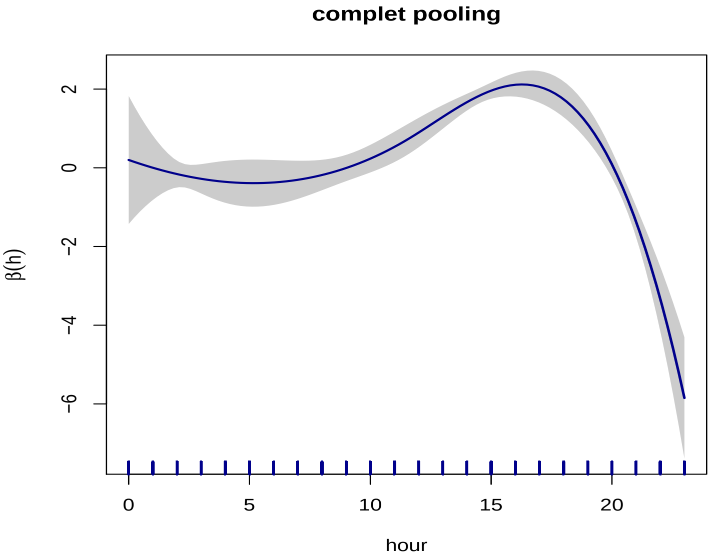

Figure 5.4: The estimate of functional ozone coefficient, $\beta(h)$, estimated by pooling data from 15 USA cities for the summer periods of the years 1987-2000.

wider credible intervals as compared to the estimate from the overall model which was estimated by pooling all cities together. As for city-specific estimates, except New York and Philadelphia, the other cities showed strong ozone effect in the afternoon and evening hours. These daily hours in the summer time are very hot for which ozone level can reach maximum to cause health issues. This is a general characteristic of ozone irrespective of difference in the estimated shapes.

The between-city variability is captured by the parameter σ_b^2. The median of the posterior samples for σ_b^2 is 0.686 and the associated 95% credible interval is $(0.547, 0.878)$. The interval is far from zero, this indicates the presence of relatively strong between-city variability. As far as the confounding variables are concerned, the distribution of the posterior samples and the associated summary information for days of the week and calendar year are reported in Figure 5.6.

Often, the health effect of exposure to pollutants can persist for some days from the date of exposure. We take into consideration the persistent effect of ozone using the time lagged ozone values to explain the current day mortality. Thus, the one day lag curve $X_{t-1}(h)$ is used instead of $X_t(h)$ as a functional predictor, the estimated coefficients are reported in Figure 5.7. The shape of the pooled functional estimate from lag 0 and lag 1 are generally the same, the credible band for estimates of lag 1 are slightly wider.

5.4 Discussion

This Chapter presents a method to estimate health effect of exposure to ozone from multi-city data accounting for temporal and spatial heterogeneity. The functional generalized linear model has been extended to functional hierarchical model using Bayesian paradigm to estimate location-specific and pooled pollutant effect controlling for confounders. Measuring exposure to pollutant in the form of function allows to include the daily temporal variability of the pollutant while a hierarchical structure is used to model spatial heterogeneity. The Bayesian paradigm is employed because it is convenient to deal with the hierarchical structure. We estimated both a pooled functional coefficient and coefficients that can vary across the cities. The pooled estimate showed that the most relevant hours are those around afternoon where ozone reaches maximum, and we also note that the shape of the pooled functional coefficient is coherent with the estimate of the pollutant effect from the overall functional regression model where city level variability is ignored. A similar shape of functional pooled coefficient is estimated from the one day lag ozone but with wider credible intervals. City specific estimates suggest that the effect of ozone is widely variable according to location, from being minimal exposure effect (New York, San Francisco, Philadelphia) to exhibiting a strong effect. The shapes of the functional coefficients also vary widely to the point of being opposite (Dallas, El Paso). This circumstance counts as evidence that a model allowing for location-dependent estimates is relevant.

The approach discussed in this Chapter can easily be extended to perform other air pollutants and health outcomes. This can be done for single city studies as the overall model in Section 5.3.1 or for multi-city studies as the functional hierarchical model in Section 5.3.2.

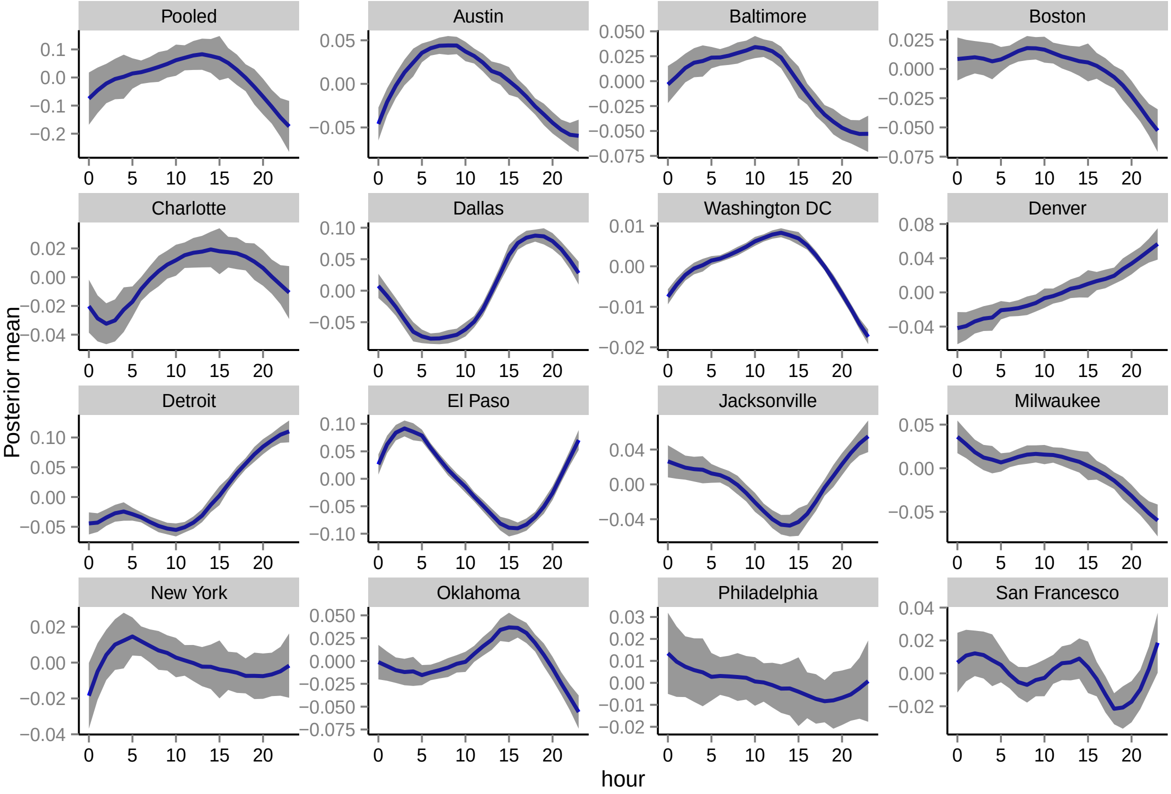

Figure 5.5: The mean of the simulated values for city-specific functional coefficients including the pooled functional coefficient estimate using data for 15 USA cities for the summer periods of the years 1987-2000. The confidence bands are the 95% point-wise credible bands and the mean of the posterior estimates are shown in solid line.

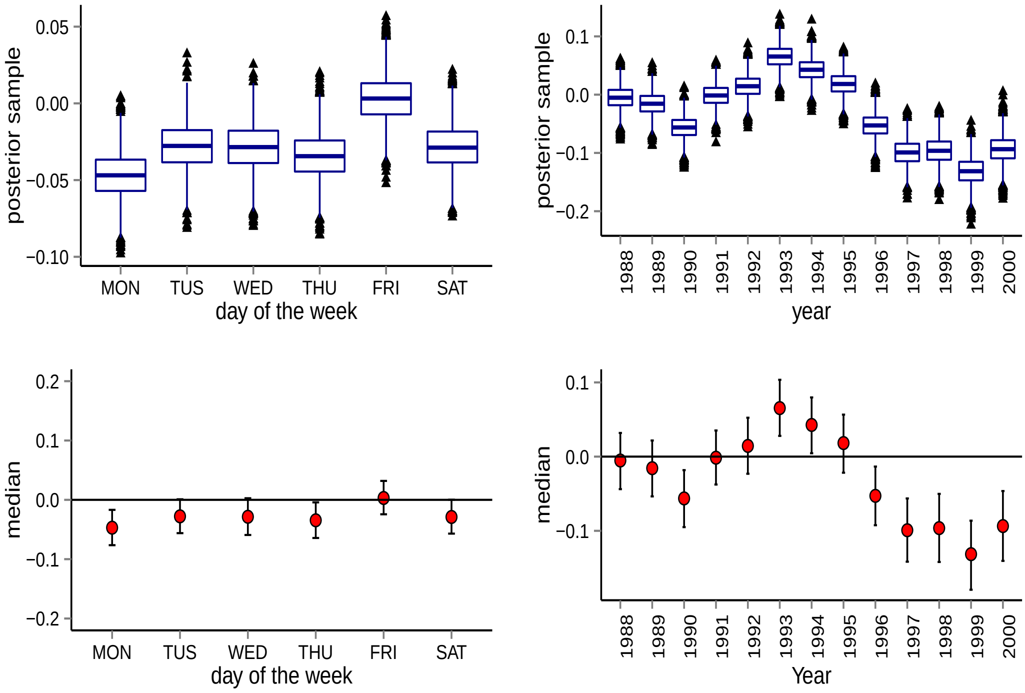

Figure 5.6: The upper panel shows the distributions of the simulated values for day of the week and calendar year parameters. The lower panel shows the median and 95% credible end points summarized from the simulated values of the same parameters.

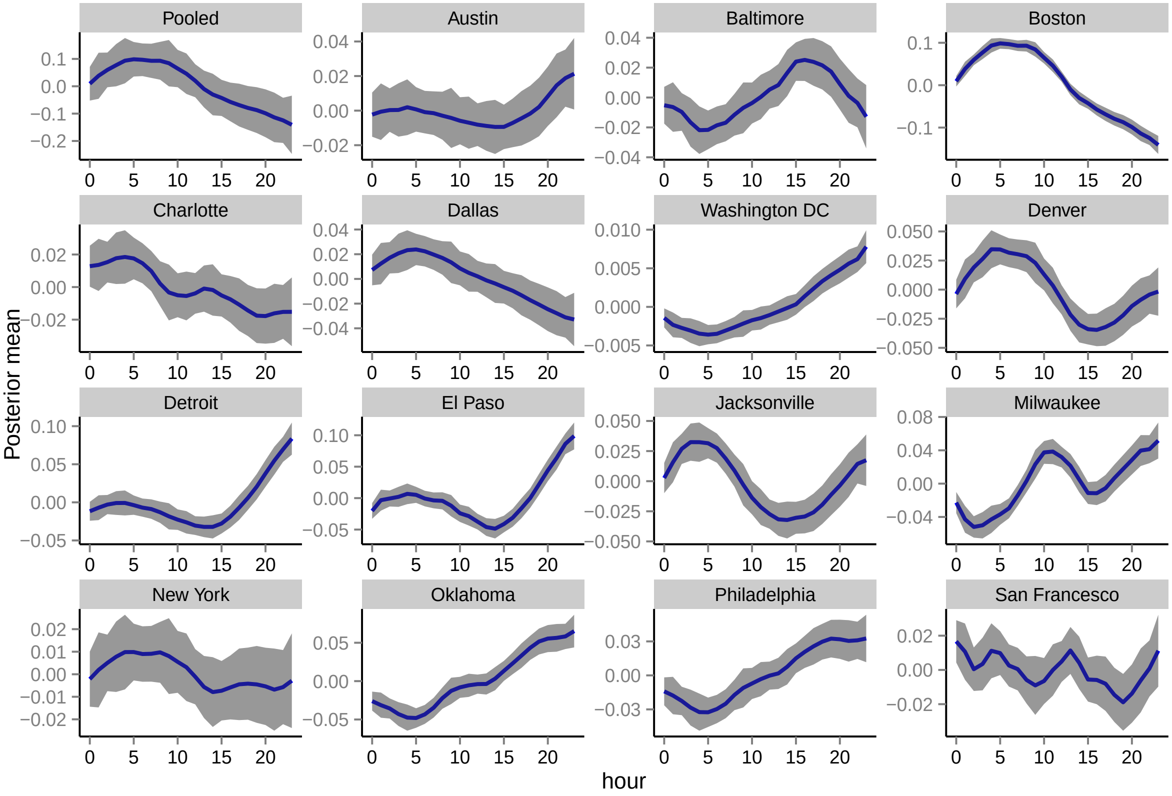

Figure 5.7: The mean of the simulated values for city-specific functional coefficients including the pooled functional coefficient estimate using one day lagged data for 15 USA cities for the summer periods of the years 1987-2000. The confidence bands are the 95% point-wise credible bands and the mean of the posterior estimates are shown in solid line.

Chapter 6

Principal Scores to Measure Exposure to Pollutants

In Chapter 4, we provided a functional measure of pollutant concentrations and used the measure to explain the health effect of ozone using functional regression model. The approach was used for single city data and then extended to the functional hierarchical model in Chapter 5 to model multi-city data and account for spatial heterogeneity in the estimated effect. In this Chapter, we provide an alternative approach to measure exposure to the pollutant concentrations by computing a principal scores from the hourly measures of ozone via functional principal component analysis. We aim to obtain an estimate of the effect allowing for the daily variation of concentration, but with a more parsimonious model. The approach is a special case of the methods in Chapter 5, in the sense that rather than using a full functional ozone exposure, we extract the important components of the daily curve in terms of principal scores. The remainder of the Chapter is organized as follows, Section 6.1 provides the motivation to measure exposure using principal scores. Section 6.2 illustrates the models. Section 6.3 discusses the results, and finally Section 6.4 presents concluding discussion.

6.1 Motivation

The functional data analysis technique allows to treat all hourly ozone records of a day as a function instead of collapsing the hourly records into a single summary measures. An alternative measure of exposure is to use principal scores to capture the most important portion of the daily ozone curve. The principal scores are produced from the hourly ozone records using functional principal component analysis. The main advantage of the approach is dimension reduction and parsimony. The infinite dimensional functions can be represented in terms of a few dimensional principal components. This low-dimensional principal scores will be used suitably as predictors to assess the effect of exposure to ozone on health controlling for confounders. The method is applied on 15 USA city data for the summer periods of the years 1987-2000 as discussed in Section 5.2. First, the method is used to estimate the overall effect by pooling all cities data together, then we fit a Bayesian hierarchical model to account the spatial heterogeneity in the estimate of the score effect. The model estimates city-specific ozone effect measured in the form of scores and the pooled effect averaged over the study cities.

6.2 Principal Score to Measure Exposure

The daily number of deaths in day t of city c, y_{tc} are assumed to be distributed as a Poisson variable with mean μ_{tc}, whose logarithm is modeled as an additive function of the covariates: the ozone concentration and, as a control, the temperature, the day of the week and the calendar year using the functional regression model which is given in (5.1). Proceeding similarly to Section 5.3, the concentration curve $X_{tc}(h)$ is specified as a linear combination of an overall orthogonal eigenfunctions $\{\phi_l(h)\}_{l=1,2...}$, with coefficients ξ_{tcl}, which are uncorrelated functional principal scores with zero mean and variances λ_l. The eigenfunctions are obtained using the overall empirical covariance function $K(s,h)$ (see Section 5.3) and normalized in the sense that $\int \psi_l^2(h)dh = 1$ to be used as basis functions. The main difference between model specification in this Chapter and the previous lies in the specification of the coefficient parameter $\beta(h)$. Rather than expressing $\beta(h)$ in terms of B-spline basis functions, we use the estimated orthonormal basis $\psi_l(h)$ as basis functions

$$\beta(h) = \sum_{k=1}^{K} \beta_k \psi_k(h). \tag{6.1}$$

Using the same orthonormal basis to represent the concentration curve $X_t(h)$ and the coefficient $\beta(h)$, the linear predictor (5.1) has reduced to

$$\log \mu_t = \boldsymbol{\xi}_{tc}\boldsymbol{\beta} + f_c(\text{Temp}_{tc}) + \sum_{j=1}^{6} \delta_j \text{DOW}_{tj} + \sum_{k=1}^{13} \gamma_k \text{Year}_{tk}, \tag{6.2}$$

which is a generalized linear model with a functional covariate using principal component scores $\boldsymbol{\xi}_{tc}$ as main predictors. These principal scores are estimated by plugging the estimator of $\psi_l(h)$ in (5.5) and evaluating the integral over a grid of points. The advantage is twofold: dimension reduction by reducing the number of coefficients and avoids possible multicollinearity issues among the predictors. Independence between the scores is evident from Figure 6.1(b) where score 2 is plotted against score 3.

We choose the first 3 principal scores, hence the dimension of covariate matrix $\boldsymbol{\xi}_{tc} = (\xi_{tc1}, \xi_{tc2}, \xi_{tc3})$ is $T \times 3$ with coefficients $\boldsymbol{\beta} = (\beta_1, \ \beta_2, \ \beta_3)'$. The first three eigenfunctions capture 90% of the total variability, suggesting that the selected scores sufficiently describe important aspects of the daily ozone curve. The corresponding eigenfunctions are displayed in Figure 6.1(a). We fit the model using Bayesian paradigm, specifying weakly informative priors for all model parameters, in particular, normal priors with mean zero and large variance are assigned for $\boldsymbol{\beta}$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_6)$ and $\boldsymbol{\gamma} = (\gamma_1 \ldots, \gamma_{13})$. We now focus on extending specification (6.2) to model multi-city data and account for spatial heterogeneity. Rather than estimating one marginal co-

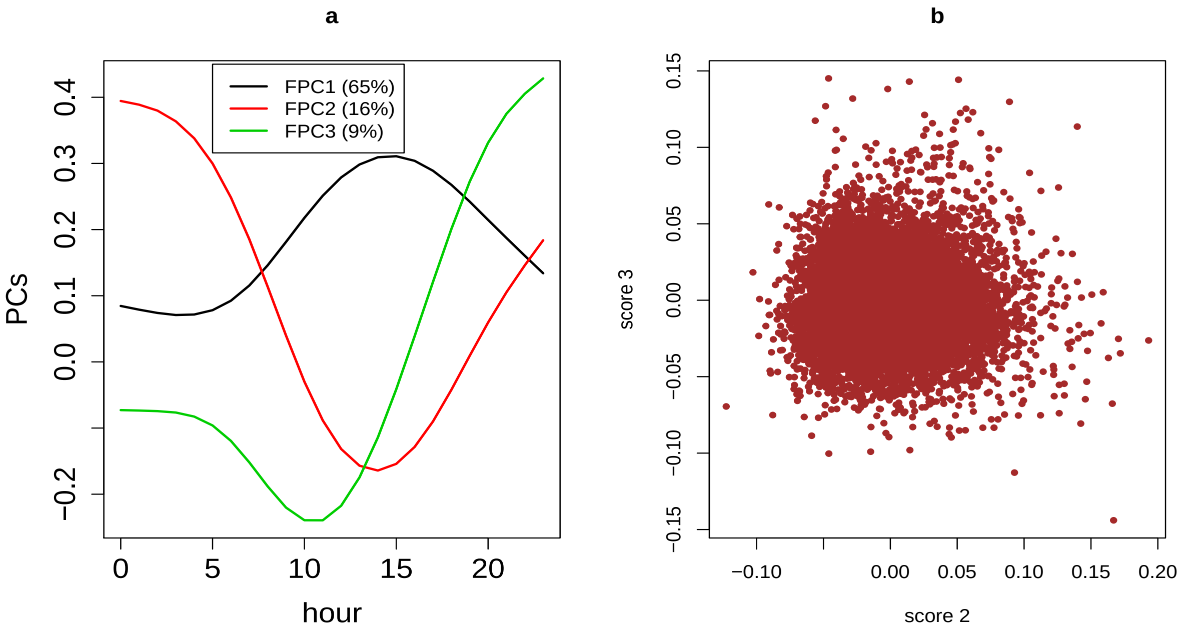

Figure 6.1: The three functional principal components with share of explained variance in the bracket (a) and bivariate plot of principal score 2 against score 3 (b). The principal scores are independent of each other.

efficient parameter $\boldsymbol{\beta}$, we estimate exposure to ozone effect that can vary across cities and estimate the overall effect. The model is

$$\log \mu_{tc} = \boldsymbol{\xi}_{tc}\boldsymbol{\beta}_c + f_c(\text{Temp}_{tc}) + \sum_{j=1}^{6} \delta_j \text{DOW}_{tj} + \sum_{k=1}^{13} \gamma_k \text{Year}_{tk}, \qquad (6.3)$$

where $\boldsymbol{\beta}_c = (\beta_{c1},\ \beta_{c2},\ \beta_{c3})'$ indicate the three city-varying score coefficients for $c = 1, \ldots, 15$. This model is the same as model (5.7) where matrix $\mathbf{J}$ becomes the identity matrix I. The prior specification of $\boldsymbol{\beta}_c$ is similar to the one adopted in Chapter 5

$$\begin{cases} \boldsymbol{\beta}_c \sim \text{Normal}(\boldsymbol{\mu}_\beta, \sigma_\beta^2 I) \\ \boldsymbol{\mu}_\beta \sim \text{Normal}(\mathbf{0}, \sigma_\mu^2 I). \end{cases} \qquad (6.4)$$

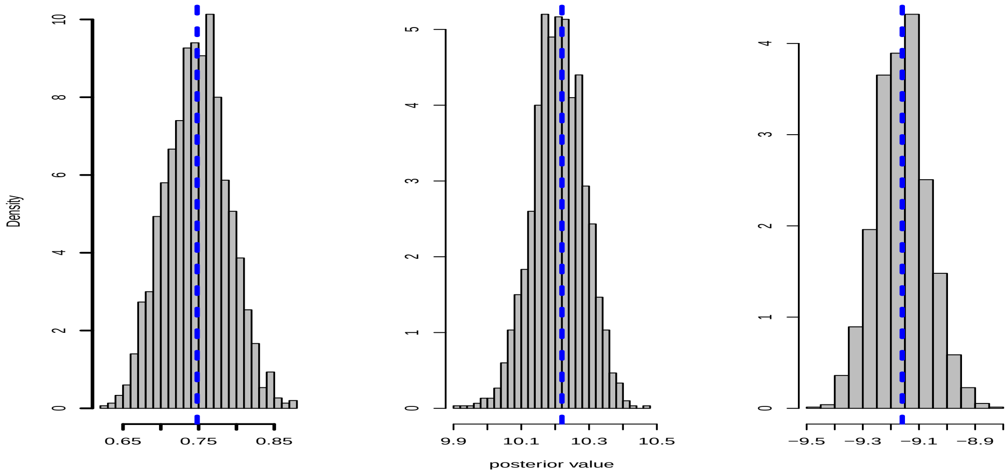

Figure 6.2: Distribution of samples from the posterior distribution of associated to coefficients; the first principal score (left), the second principal score (middle) and the third principal score (right). The median for each score coefficient is indicated by a thick broken line.

The parameter $\boldsymbol{\mu_{\beta}} = (\mu_{\beta_1},\ \mu_{\beta_2},\ \mu_{\beta_3})$ denote the overall score effect associated to each score coefficient $\boldsymbol{\beta}_c$. The variance component $\sigma^2_{\beta}I$ is a 3×3 diagonal matrix with the diagonal elements $\sigma^2_{\beta_1}, \sigma^2_{\beta_2}$ and $\sigma^2_{\beta_3}$ which measure the between city heterogeneity with respect to each estimate of $\boldsymbol{\beta}_c$ and σ^2_{μ} measures the average variability. For these variance components, a weakly informative prior is used, particularly, a gamma distribution with mean one and large variance.

6.3 Results

We estimated the models using the Gibbs sampler of Markov Chain Monte Carlo techniques. Relatively short chains are required to grant convergence, and the posterior draws are used to summarize the models. First we obtained estimates of the parameters by pooling the cities together as in model (6.2). Figure 6.2 displays the posterior distributions of the three score coefficients.

	Mean	S.E	lower	Median	upper
score[1]	0.781	0.090	0.680	0.757	1.0259
score[2]	10.208	0.087	10.052	10.220	10.3555
score[3]	-9.125	0.135	-9.332	-9.137	-8.8737
Day of the week					
Monday	-0.110	0.024	-0.134	-0.115	-0.0371
Tuesday	0.073	0.027	0.046	0.067	0.1591
Wednesday	0.028	0.027	-0.001	0.023	0.1084
Thursday	-0.129	0.029	-0.156	-0.135	-0.0411
Friday	0.061	0.028	0.036	0.057	0.1559
Saturday	0.039	0.024	0.015	0.034	0.1077

Note: S.E: standard deviation

Table 6.1: Posterior summaries of a model for which separate city data are pooled. The three principal scores are computed from hourly ozone data using functional principal component analysis.

Relevant summary measures including 95% credible intervals computed from the posterior samples are reported in Table 6.1. The credibility intervals for the score coefficients do not include zero, thus suggesting a statistically significant effect. The first three eigenfunctions grasp the variation in the afternoon hours where concentration reaches maximum, hence the strong effect shown by the associated principal scores may be an expected event and it appears broadly coherent with the estimate of $\beta(h)$ according to the overall model in Chapter 5 (5.1).

A second model is considered to obtain city-varying score effects accounting for spatial heterogeneity as given in (6.3). Figure 6.3 presents the distribution of the posterior values (upper) for the parameters $\boldsymbol{\beta}_c$ and the associated median estimate along with 95% credible set (lower). The Figure also depicts the pooled parameters for each score. No apparent outliers are observed in the distribution of the posterior values for β_{c_1} . It appears that the three principal scores are not significant for the city of Austin (aust, for short notation of cities name, see Table 5.1). For the second principal scores, the posterior values for Dallas (dall) and Milwaukee (mil) are slightly far from the others. Among the three scores, the third principal score is not significant for many cities.

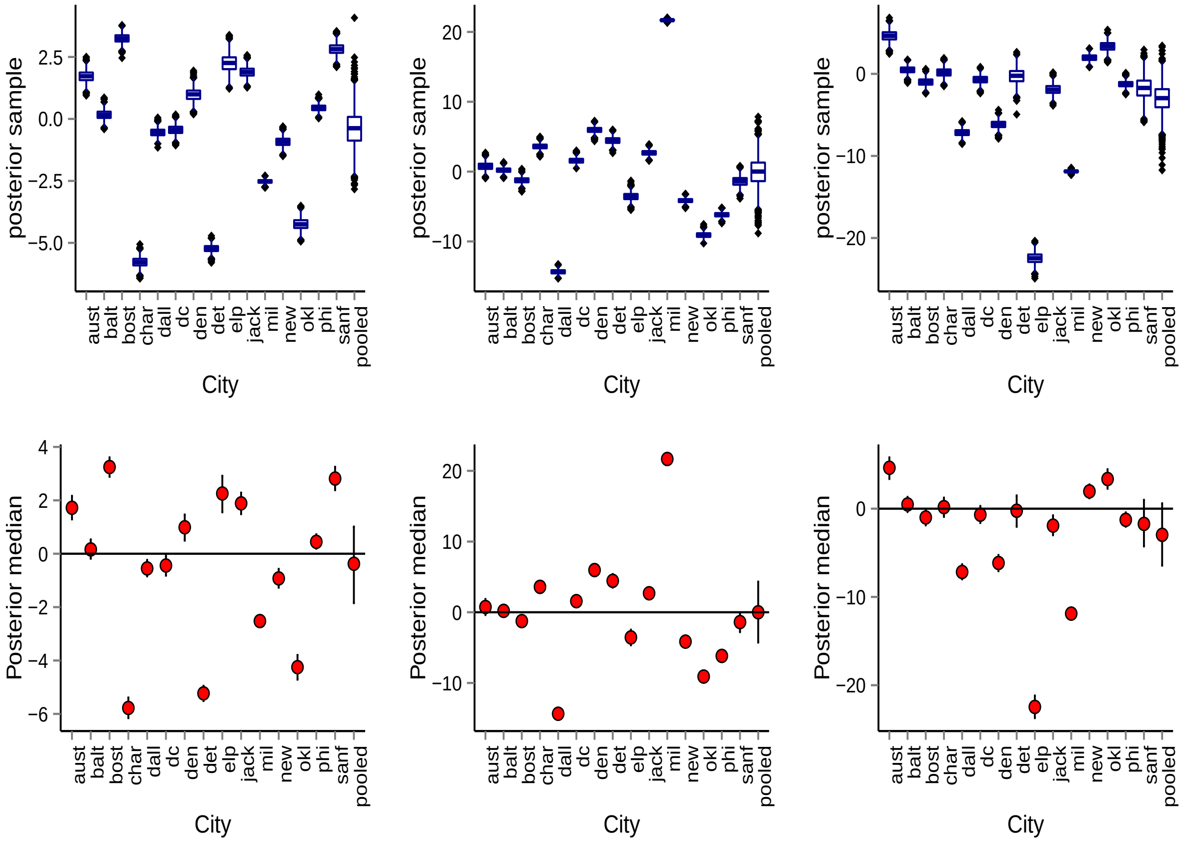

Figure 6.3: The distribution of the posterior values (upper) and the associated median along with 95% credible intervals (lower) of the parameters β_{c1}, β_{c2} and β_{c3} from left to right. The posterior information of the pooled parameters ($\mu_{\beta1}$, $\mu_{\beta2}$, $\mu_{\beta3}$) are given in the far right hand side of each plot.

To measure the persistence of the exposure, up to two days lagged values of ozone concentrations are used to explain the current day mortality. Figure

6.4 displays the median and associated 95% credible sets of the posterior samples for the first two scores coefficients across the cities and their overall effect. Exposure estimates for lag 1 effect are similar to the same day exposure effect (Figure 6.3), while lag 2 estimates seem away from zero, particularly with respect to score 1.

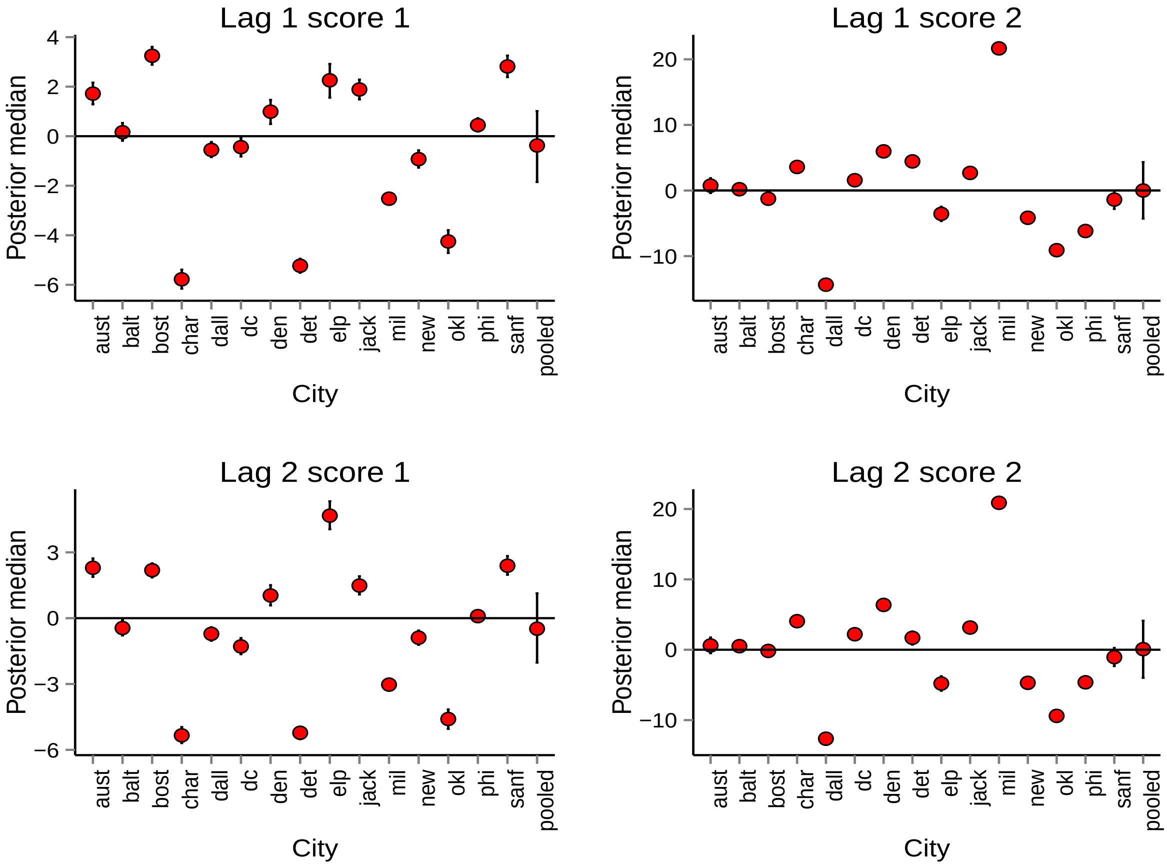

Figure 6.4: The medians and associated 95% credible intervals of posterior samples for lag 1 and lag 2 exposure effect. The pooled information is shown in the far right hand side of each plot.

The results of the variance components for the same day and lag 1 exposure are summarized in Table 6.2. The median of the three variances $\sigma^2_{\beta_1}$, $\sigma^2_{\beta_2}$ and $\sigma^2_{\beta_3}$ are quite different from zero. The fact that the 95% credible intervals of

	Mean	2.5%	Median	97.5%
$\sigma^2_{\beta_1}$	2.831	1.978	2.758	4.060
$\sigma^2_{\beta_2}$	7.931	5.571	7.699	11.410
$\sigma^2_{\beta_3}$	6.724	4.737	6.522	9.852
$\sigma^2_{\beta_1}$.lag1	2.931	1.978	2.757	4.060
$\sigma^2_{\beta_2}$.lag1	8.226	5.570	7.699	11.410
$\sigma^2_{\beta_3}$ .lag1	7.036	4.737	6.522	9.850

Table 6.2: The posterior summaries of the variance component under the city-specific model for lag 0 and lag 1 exposure. These variances describe the between city variability in the estimate of the three score coefficients.

these between city variabilities are away from zero suggests the existence of between location variability. Thus employing regression models by pooling the data may not be appropriate.

6.4 Discussion

In this Chapter, we consider modeling the health effect of ozone for which its exposure is measured in terms of principal scores which are computed from the hourly pollutant measurements. Functional principal component analysis method is used to generate a fixed number of principal scores. We have selected the first three principal scores that grasp approximately 90% of the daily ozone variation. These scores also capture important aspects of the daily ozone curve, particularly, the region of ozone where believed to be potentially harmful such as the daily maximum region. Two models are proposed to investigate the effect of the principal scores on health. In the first model, no attempt were made to consider the heterogeneity effect of the cities. In this case the model was fitted on pooled data from 15 USA cities. The fact that the principal scores lead to an estimated ozone effect that is significantly different from zero, may be a confirmation that the proposed exposure measures reflect important aspects of exposure. In the second approach, we fit a Bayesian hierarchical model to account for spatial heterogeneity which assumes that the association between exposure to the pollutant and mortality can vary across the cities. The framework allows to estimate an overall effect and it can be considered as a combined information across the cities to make inferences regardless of locations.

We used up to two days lagged values of hourly ozone concentration to measure the persistent effect of exposure. The exposure effect of one day lag is not quite dissimilar to the same day exposure effect and the effect of two day exposure seems more influential. A measure of between city variation among the three scores is estimated, and suggests strong presence of geographic heterogeneity. This may confirm that estimated environmental exposure effect is spatially heterogeneous. Finally, the models can be used to study the health effect of other pollutants or joint pollutants effect.

Chapter 7

Conclusion

7.1 Description

The adverse health effect of exposure to environmental pollutants has become a global issue since early twentieth century when a series of severe air pollution episodes occurred in different parts of the world. The 1930 Meuse Valley fog in Belgium and the London 'Great Smog' in 1952 were some of the episodes which were associated to a rise in the number of hospital admission and premature death. These events led to formulation of strategies to reduce the environmental pollution levels. However, pollution persists at high levels, and studies continued to detect effect on human health from the exposure. Studies use data typically consists of health outcome data, the measurements of pollutants and various confounding variables for a certain study region. The health outcome data contains daily counts of mortality (death) or morbidity (hospital admission) for the population residing within the study area. There are several pollutants studied for their effect on health, with ozone being the main pollutant studied in this thesis. Concentrations of pollutants are obtained from a network of monitoring stations located throughout the study region. Measurements are typically taken at various times throughout the day, often on hourly basis.

To represent the daily exposure to a pollutant concentration, the daily patterns of the pollutant have been reduced to a single summary figures. Then, the health effect of the exposure has been estimated by regressing day-varying health outcome against day-varying daily summary measure of exposure using generalized linear models (GLM) or generalized additive models (GAM). Such exposure effect has been estimated at various geographic regions (Gryparis et al., 2004; Zhang et al., 2006), and it has been shown through multi-city studies that the estimated effect is spatially heterogeneous (Katsouyanni et al., 1996; Samet et al., 2000). However, the use of daily summaries to represent daily exposure to the pollutant have been unsatisfying.

In this thesis, we have proposed approaches to measure exposure to pollutants to address the limitations of the traditional exposure measures and improve effect estimation. The work presented in this thesis has been centred around two major themes. The first theme proposes two related representative measures of daily exposure to pollutants using functional data analysis. The first exposure measure attempts to resolve the issue of measuring exposure by treating all hourly measures of a day as one function accounting for temporal variation of the pollutant. The predictive efficiency of our approach is superior as compared to the predictive accuracy of models based on the daily summaries and other related approaches from literature. A second exposure measure considered is principal scores computed from hourly ozone measurements using functional principal component analysis technique. The computed principal scores grasp the most important portion of the daily ozone variation. The second theme is an extension of the first theme, implemented for data which come from multiple geographic locations. Both the functional form and principal scores exposure measures have been embedded into a hierarchical model to allow for spatial heterogeneity of the effect. We assumed the exposure effect to vary across the study location and estimated both location-specific and an overall effect.

7.2 Synthesis of empirical findings

Pollutants are a potent health and environmental problem. Ground level Ozone (O3) is one of the main component of pollutants produced by chemical reaction in the presence of sunlight. The process of ozone formation is a continuous process, it follows different daily fluctuations exhibiting strong daily patterns. The concentrations of ozone are usually measured by a network of monitoring sites in a particular study region. Each monitoring site typically measures continuously throughout the day. Despite the advances in ozone monitoring and recording, there continue to be gaps in analyzing the monitored data. One main issue in this regard is the method to measure human exposure to ozone. Systematic models that take into account the daily patterns of ozone are particularly rare. Despite the availability of the hourly measurements from monitoring networks, most studies collapse these measurements into single daily summaries. The most frequently used daily summaries are the average and maximum of the 24 hour measurements.

The daily summaries are rough synthesis of the daily pattern of pollutant concentrations and totally disregard the temporal variation observed in the daily concentrations. Further, high ozone concentrations which can be harmful to health can be recorded at a particular hour of the day, but this would not necessarily mean that a daily summary is predictive for health outcome. The daily summaries may not be representative of the actual personal exposure to individuals, since they ignore the portion of the time spent outdoor. For instance, Figures 3.2 and 4.2 display a clear pattern for which the concentrations reach maximum in the afternoon hours, and low measurements in the morning and night hours. It may be argued that a good measure of exposure should contain information on the presence of high concentration during the day. We therefore proposed that the daily exposure of a person is the concentration level as a function of daily hour. While it is true that any measure of concentration may not be a precise measure of exposure, we argue that allowing for daily pattern is important improvement as compared to single summaries. Thus, we remark that estimates of exposure effect on health contain these inherent facts of the pollutant.

In Chapter 4, we used functional data analysis techniques to describe hourly ozone measurements as a function accounting for temporal variation of the pollutant. We implemented a functional regression models considering the functional exposure measure as a predictor and hospital admission counts as the response, with pollutants measured in the monitoring network of city of Milan, Italy for the summer periods of years 1996-2002. We estimated the effect of exposure as a function of daily hour which allows to examine the influence of the pollutant throughout the day, as opposed to a single scalar estimate associated to daily summaries. Subsequently, the portion of daily ozone function potentially linked to health has been recognized. Particularly, the region of the estimate from afternoon hours on wards, where the concentration level reaches maximum, is identified as potentially harmful to health. A drawback to the functional regression model in Chapter 4 is that it has to be applied to a single study region. It is likely that the estimated effect is spatially heterogeneous across different study regions.

In Chapter 5, we have proposed a functional hierarchical modeling approach to estimate pollutant exposure effect allowing for the daily variation of the concentration and spatial heterogeneity of the effect at once. The approach was developed using functional regression model which has been discussed in Chapter 4 and the Bayesian hierarchical model paradigm. We estimated both city-varying and overall exposure effect as a function of time of the day for 15 USA cities for the summer periods of the years 1987-2000. As shown in Figure 5.5, the health effect of exposure to the pollutant is widely variable according to geographic locations, both in strength of the effect and shape. The pooled estimate is a synthesis of information from the study locations and used to make inferences in the overall effect, and it is broadly coherent with the overall estimate obtained by pooling cities together that assumed homogeneous exposure effect across locations (see Figure 5.4). However, the former approach produced wider credible intervals while the latter approach, which does not use variability across the locations, produced narrow credible intervals.

In Chapter 6, we have provided an alternative and simpler approach, which still meets the aim of producing a representative exposure measure allowing for daily variation. This approach uses a fixed number of principal scores derived from the hourly concentrations through functional principal component analysis. These scores capture the important portion of the concentration curve to be used as a potential exposure measure in the model. We exploited the orthogonality property of the eigenfunctions to obtain uncorrelated principal scores which avoids multicollinearity related issues in the use of the principal scores as covariates. A further advantage of the approach is dimension reduction; the infinite dimensional functions can be represented in terms of a few dimensional principal components. To estimate the model, we adopted Bayesian hierarchical model using data coming from 15 USA cities. The model estimated a strong exposure effect which may be a confirmation that the proposed exposure measures reflect important aspects of exposure. Further, the estimated effect broadly coherent with the estimate of a full functional exposure measure presented in Chapters 4 and 5. This may be interesting, in the sense that the principal scores exposure measure is a special case of the functional exposure measure, since the principal scores are obtained by using orthogonal eigenfunctions as basis functions instead of the B-splines to specify the effect of the exposure measure. The approach serves as an alternative pollutant exposure measures to other approaches such as Chiogna and Pauli (2011) and Arisido (2014).

7.3 Limitations

In this thesis, we have explored the challenges of exposure measure in environmental studies of pollutants and health and presented alternative approaches to address these challenges. The proposed methods are based on the principle of functional data analysis and Bayesian methods. The approach is first used for a single study location as illustrated in Chapter 4, and then extended to study multiple geographic locations as in Chapters 5 and 6.

In Chapter 4, we have compared the predictive performances of different exposure measures using out-of-sample prediction method. This may not be a definitely perfect means of comparing different methods, but it is more of an assessment means to get how the proposed approach works against the standard approaches before embarking to extend the method to the more complex approach as done in Chapter 5. The Chapter studied the confounding effect of the daily average of particulate matter in the ozone functional model. We used the daily average since hourly recorded data for particulate matter were not available. It would have been interesting to consider additional functional predictors to assess which predictors play the most important role in predicting the response.

Chapter 5 illustrated a model accounting for spatial heterogeneity of estimated exposure effect in the presence of temporal variability in the pollutant concentrations. It is a more general model to estimate the health effect of a functional exposure measure for a multi-location study. The relevance of the model is quite clear as it accounts for spatio-temporal variation and used to make inference for both location-specific and pooled estimates. However, two criticisms can be directed at the model, first the model did not account for spatial correlation that may be present between closest cities. In the future studies, we shall assume that the functional coefficients $\beta_c(h)$ and

$\beta_{c'}(h)$ estimated from the proposed model might be correlated as a function of distance between two cities c and c'. This type of work is not uncommon in multivariate setting (see for example Dominici et al. (2000)). In the context of this thesis, investigation of the correlation between two functional object will be an additional issue. The second limitation is related to comparing the method with other approaches. It is rather difficult to compare the results with other studies, in fact the functional coefficient is not directly comparable to the scalar coefficient of a single summary measure. Thus, the functional hierarchical model do not allow the possibility of summarizing the exposure effect in a single quantity in order to compare with other studies.

Chapters 5 and 6 have illustrated a picture of regional variation in the effect of exposure to ozone. However, we only included a sample of 15 cities from USA, and would have been included more number of cities to obtain a more national representative results. This could also allow to investigate whether the results be the same if more number of cities were included in the study. In the future, we shall continue the study by including as many cities as possible in order to obtain a national health effect of ozone. Further, the Chapters did not include the confounding effect of other pollutants in any form, since measurements for all pollutants other than ozone in the summer of 1987-2000 were obtained only once for every six days. Therefore, the majority of study days did not have measurements. In the future, we aim to pursue the joint functional effects of ozone and particulate matter in other study regions where data are accessible. This would potentially include the functional confounding effect of weather variables such as temperature.

References

Albert, J. (2007). *Bayesian computation with R.* Springer.

Arisido, M. (2014). Functional data modeling to measure exposure to ozone. *21st International Conference on Computational Statistics (Geneva, 19-23 August).*

Armstrong, B. (2006). Models for the relationship between ambient temperature and daily mortality. *Epidemiology*, 17(6):624–631.

Bell, M. L., Dominici, F., and Samet, J. M. (2005). A meta-analysis of time-series studies of ozone and mortality with comparison to the national morbidity, mortality, and air pollution study. *Epidemiology (Cambridge, Mass.)*, 16(4):436.

Bell, M. L., McDermott, A., Zeger, S. L., Samet, J. M., and Dominici, F. (2004). Ozone and short-term mortality in 95 us urban communities, 1987-2000. *Jama*, 292(19):2372–2378.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis.* Springer.

Bosq, D. (2000). *Linear processes in function spaces: theory and applications.* Springer.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to time series and forecasting.* Taylor & Francis.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*.

Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics and Probability Letters*, 45(1):11–22.

Chiogna, M. and Pauli, F. (2011). Modelling short-term effects of ozone on morbidity: an application to the city of milano, italy, 1995–2003. *Environmental and Ecological Statistics*, 18(1):169–184.

Clayton, D., Hills, M., and Pickles, A. (1993). *Statistical models in epidemiology*. IEA.

Corradi, C. (1977). Smooth distributed lag estimators and smoothing spline functions in hilbert spaces. *Journal of Econometrics*, 5(2):211–219.

Crainiceanu, C. M. and Goldsmith, A. J. (2010). Bayesian functional data analysis using winbugs. *Journal of statistical software*, 32(11).

Crainiceanu, C. M., Staicu, A.-M., and Di, C.-Z. (2009). Generalized multilevel functional regression. *Journal of the American Statistical Association*, 104(488):1550–1561.

De Boor, C. (1976). Splines as linear combinations of b-splines. a survey.

De Sario, M., Katsouyanni, K., and Michelozzi, P. (2013). Climate change, extreme weather events, air pollution and respiratory health in europe. *European Respiratory Journal*.

Dewey, S. H. (2000). *Don't breathe the air: Air pollution and US environmental politics, 1945-1970*. Texas A&M University Press College Station, TX.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *The annals of applied statistics*, 3(1):458.

Dobson, A. J. (2001). *An introduction to generalized linear models.* CRC press.

Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris Jr, B. G., and Speizer, F. E. (1993). An association between air pollution and mortality in six us cities. *New England journal of medicine*, 329(24):1753–1759.

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(3):193–203.

Dominici, F., Samet, J. M., and Zeger, S. L. (2000). Combining evidence on air pollution and daily mortality from the 20 largest us cities: a hierarchical modelling strategy. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(3):263–302.

Dominici, F., Sheppard, L., and Clyde, M. (2003). Health effects of air pollution: a statistical review. *International Statistical Review*, 71(2):243–276.

Dumouchel, W. (1995). Meta-analysis for dose—response models. *Statistics in medicine*, 14(5-7):679–685.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science.*

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing.* CRC press.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: Models, methods and applications.* Springer.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability.* CRC Press.

Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics*, 39(3):254–261.

Ferraty, F. (2011). *Recent advances in functional data analysis and related topics*. Springer.

Gao, H. O. (2007). Day of week effects on diurnal ozone/nox cycles and transportation emissions in southern california. *Transportation Research Part D: Transport and Environment*, 12(4):292–305.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). Bayesian data analysis, (chapman & hall/crc texts in statistical science).

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*.

Geweke, J. et al. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department.

Gill, J. (2007). *Bayesian methods: A social and behavioral sciences approach*. CRC press.

Goldberg, M. S., Burnett, R. T., Brook, J., Bailar, J. C., Valois, M.-F., and Vincent, R. (2001). Associations between daily cause-specific mortality and concentrations of ground-level ozone in montreal, quebec. *American journal of epidemiology*, 154(9):817–826.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Graves, S., Hooker, G., and Ramsay, J. (2009). Functional data analysis with r and matlab.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press.

Gryparis, A., Forsberg, B., Katsouyanni, K., Analitis, A., Touloumi, G., Schwartz, J., Samoli, E., Medina, S., Anderson, H. R., Niciu, E. M., et al. (2004). Acute effects of ozone on mortality from the "air pollution and health: a european approach" project. *American journal of respiratory and critical care medicine*, 170(10):1080–1087.

Härdle, W. and Simar, L. (2007). *Applied multivariate statistical analysis*. Springer.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*. CRC Press.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hazucha, M. J., Bates, D. V., Bromberg, P. A., et al. (1989). Mechanism of action of ozone on the human lung. *J Appl Physiol*, 67(4):1535–1541.

Indritz, J. (1963). *Methods in analysis*. Macmillan New York:.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):411–432.

James, G. M. and Silverman, B. W. (2005). Functional adaptive model estimation. *Journal of the American Statistical Association*, 100(470):565–576.

Ji, M., Cohan, D. S., and Bell, M. L. (2011). Meta-analysis of the association between short-term exposure to ambient ozone and respiratory hospital admissions. *Environmental Research Letters*, 6(2):024006.

Jolliffe, I. T. (2002). Introduction. *Principal Component Analysis*.

Karhunen, K. (1947). *Über lineare Methoden in der Wahrscheinlichkeitsrechnung*, volume 37. Universitat Helsinki.

Katsouyanni, K., Schwartz, J., Spix, C., Touloumi, G., Zmirou, D., Zanobetti, A., Wojtyniak, B., Vonk, J., Tobias, A., Pönkä, A., et al. (1996). Short term effects of air pollution on health: a european approach using epidemiologic time series data: the aphea protocol. *Journal of epidemiology and community health*, 50(Suppl 1):S12–S18.

Kneip, A. and Ramsay, J. O. (2008). Combining registration and fitting for functional models. *Journal of the American Statistical Association*, 103(483):1155–1165.

Levy, J. I., Chemerynski, S. M., and Sarnat, J. A. (2005). Ozone exposure and mortality: an empiric bayes metaregression analysis. *Epidemiology*, 16(4):458–468.

Li, K.-C. (1986). Asymptotic optimality of cl and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*.

Loeve, M. (1965). *Fonctions aléatoires du second ordre...*

Malfait, N. and Ramsay, J. O. (2003). The historical functional linear model. *Canadian Journal of Statistics*, 31(2):115–128.

Marr, L. C. and Harley, R. A. (2002). Spectral analysis of weekday–weekend differences in ambient ozone, nitrogen oxide, and non-methane hydrocarbon time series in california. *Atmospheric Environment*, 36(14):2327–2335.

Marx, B. D. and Eilers, P. H. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis*, 28(2):193–209.

McCullagh, P. and Nelder, J. A. (1989). Generalized linear models.

Mclean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F., and Ruppert, D. (2014). Functional generalized additive models. *Journal of Computational and Graphical Statistics*, 23(1):249–269.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Moolgavkar, S. H., Luebeck, E. G., Hall, T. A., and Anderson, E. L. (1995). Air pollution and daily mortality in philadelphia. *Epidemiology*, 6(5):476–484.

Mudway, I. and Kelly, F. (2000). Ozone and the lung: a sensitive issue. *Molecular aspects of medicine*, 21(1):1–48.

Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics.*

Müller, H.-G. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association*, 103(484).

Myers, R. H., Montgomery, D. C., Vining, G. G., and Robinson, T. J. (2012). *Generalized linear models: with applications in engineering and the sciences*, volume 791. John Wiley & Sons.

Neidell, M. (2009). Information, avoidance behavior, and health the effect of ozone on asthma hospitalizations. *Journal of Human Resources*, 44(2):450–478.

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS.* John Wiley & Sons.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical science.*

Pauli, F. and Rizzi, L. (2006). Statistical analysis of temperature impact on daily hospital admissions: analysis of data from udine, italy. *Environmetrics*, 17(1):47–64.

Peng, R. D. and Dominici, F. (2008). *Statistical methods for environmental epidemiology with R: a case study in air pollution and health.* Springer.

Peng, R. D., Welty, L. J., and McDermott, A. (2004). The national morbidity, mortality, and air pollution study database in r.

Pfanzagl, J. (1994). *Parametric statistical theory.* Walter de Gruyter.

Pope III, C. A. and Dockery, D. W. (2006). Health effects of fine particulate air pollution: lines that connect. *Journal of the Air and Waste Management Association*, 56(6):709–742.

Ramsay, J. O. (2006). *Functional data analysis.* Wiley Online Library.

Ramsay, J. O. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological).*

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies.* Springer.

Richardson, S. and Best, N. (2003). Bayesian hierarchical models in ecological studies of health–environment effects. *Environmetrics*, 14(2):129–147.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods.* Citeseer.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression.* Cambridge university press.

Samet, J. M., Zeger, S. L., Dominici, F., Curriero, F., Coursac, I., Dockery, D. W., Schwartz, J., and Zanobetti, A. (2000). The national morbidity, mortality, and air pollution study. *Part II: morbidity and mortality from air pollution in the United States Res Rep Health Eff Inst*, 94(pt 2):5–79.

Samoli, E., Schwartz, J., Wojtyniak, B., Touloumi, G., Spix, C., Balducci, F., Medina, S., Rossi, G., Sunyer, J., Bacharova, L., et al. (2001). Investigating regional differences in short-term effects of air pollution on daily mortality in the aphea project: a sensitivity analysis for controlling long-term trends and seasonality. *Environmental health perspectives*, 109(4):349.

Seinfeld, J. (1991). Rethinking the ozone problem in urban and regional air pollution.

Shah, A. S., Langrish, J. P., Nair, H., McAllister, D. A., Hunter, A. L., Donaldson, K., Newby, D. E., and Mills, N. L. (2013). Global association of air pollution and heart failure: a systematic review and meta-analysis. *The Lancet*, 382(9897):1039–1048.

Shang, H. L. (2014). A survey of functional principal component analysis. *AStA Advances in Statistical Analysis*, 98(2):121–142.

Silverman, B. and Ramsay, J. (2005). *Functional Data Analysis*. Springer.

Sousa, S., Alvim-Ferraz, M., and Martins, F. (2013). Health effects of ozone focusing on childhood asthma: What is now known–a review from an epidemiological point of view. *Chemosphere*, 90(7):2051–2058.

Staniswalis, J. G., Yang, H., Li, W.-W., and Kelly, K. E. (2009). Using a continuous time lag to determine the associations between ambient pm2. 5 hourly levels and daily mortality. *Journal of the Air and Waste Management Association*, 59(10):1173–1185.

Stern, A. C. (1973). *Fundamentals of air pollution*. Elsevier.

Stieb, D. M., Chen, L., Eshoul, M., and Judek, S. (2012). Ambient air pollution, birth weight and preterm birth: a systematic review and meta-analysis. *Environmental research*, 117:100–111.

US Environmental Protection Agency, C. U. C. F. A. P. (2007). The plain english guide to the clean air act.

Violato, M., Petrou, S., and Gray, R. (2009). The relationship between household income and childhood respiratory health in the united kingdom. *Social Science and Medicine*, 69(6):955–963.

Wahba, G. (1990). *Spline models for observational data.* Siam.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3):439–447.

Wood, S. (2006). *Generalized additive models: an introduction with R.* CRC press.

Yao, F. and Lee, T. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):3–25.

Zanobetti, A., Wand, M., Schwartz, J., and Ryan, L. (2000). Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics*, 1(3):279–292.

Zeger, S. L., Thomas, D., Dominici, F., Samet, J. M., Schwartz, J., Dockery, D., and Cohen, A. (2000). Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environmental health perspectives*, 108(5):419.

Zhang, Y., Huang, W., London, S. J., Song, G., Chen, G., Jiang, L., Zhao, N., Chen, B., and Kan, H. (2006). Ozone and daily mortality in shanghai, china. *Environmental health perspectives.*