University of Cincinnati

Date: 5/24/2019

I, Mohammad AN Bhuiyan, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Biostatistics (Environmental Health).

It is entitled:

Bayesian Shape Invariant growth curve model for longitudinal data

Student's name: **Mohammad AN Bhuiyan**

This work and its defense approved by:

Committee chair: Marepalli Rao, Ph.D.

Committee member: Monir Hossain, Ph.D.

Committee member: Jane Khoury, Ph.D.

Committee member: Heidi Sucharew, Ph.D.

Committee member: Rhonda Szczesniak, Ph.D.

Committee member: Jessica Woo, PhD

33097

Bayesian shape invariant growth curve model for longitudinal data

A dissertation submitted to the

Graduate School

of the University of Cincinnati

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy in Biostatistics - Big Data Track

in the Division of Biostatistics and Bioinformatics

of the Department of Environmental Health

of the College of Medicine

by

Mohammad Alfrad Nobel Bhuiyan

June, 2019

M.S. University of Cincinnati

Committee Chair: Marepalli B. Rao, Ph.D.

Abstract

The analysis of growth curves has played a vital role in estimating the growth trajectory of populations as well as identifying critical factors corresponding to various shapes of those trajectories. In recent years, shape invariant modeling has become an active area of research for non-parametric growth curve modeling, where a single function is transformed by scaling and shifting it to fit each subject usually through affine transformations. Lawton, first proposed SIM called it self-modeling regression; in their approach, the function for the underlying shape is illustrated for various parametric functions. Later, Beath developed a model to explain longitudinal growth patterns and extended the SIM to include time-dependent covariates. As a type of SIM, the regression spline expressed as a basis function consisting of a different set of knots; the resulting structure fitted as a nonlinear mixed effects model and parameters are typically estimated using maximum likelihood. This allows estimating the parameters for the between-subjects variation. The research in longitudinal growth curve modeling utilizing the Bayesian inferential procedure is limited, and the wider application is hindered by the computational complexities involved in such models. Cole proposed Super Imposition by Translation and Rotation model and expressed individual growth curves through three subject-specific parameters; named as size, tempo, and velocity.

It is an important inferential problem to test no association between two binary variables based on data. A Test-based on sample odds ratio is commonly used. We bring in a competing test based on the Pearson correlation coefficient. An Odds ratio does not extend to higher order contingency tables, whereas the Pearson correlation does. It is a useful exercise to understand how the Pearson correlation stacks against the odds ratio in $2x2$ tables. Another measure of association is the canonical correlation. In my second chapter we used power comparisons in $2x2$ Contingency Tables: Odds Ratio versus Pearson Correlation versus Canonical Correlation to understand how Pearson correlation stacks against the odds ratio in $2x2$ tables in the test of association.

Air pollution is a growing global challenge and may have a moderate to the severe negative impact on human health. Vehicles, households, and industries emit a complex mixture of air pollutants, within which ambient particulate matter smaller than 2.5 micro m $PM_{2.5}$ are thought to have the greatest effect on human health. Prior epidemiologic evidence suggests short-term $PM_{2.5}$ exposure is associated with the development and exacerbation of several health problems. Children are more susceptible to $PM_{2.5}$ related health effects due to their immature immune system and ongoing development and growth. The relationship of $PM_{2.5}$ with asthma emergency department visit between 2011 and 2015 was identified within the Cincinnati Children's Hospital Medical Center electronic medical record based on International Classification of Disease (ICD-9) and we used a data-driven clustering algorithm to find any clustering patterns existed by day within the study duration. In finding the impact of$PM_{2.5}$ on stoke. we used a case-crossover design, to examine the association of exposure to $PM_{2.5}$ and onset of incident stroke for the calendar year 2010.

Acknowledgments

I want to thank Dr.Md Monir Hossain, Ph.D. for guiding me throughout the Ph.D. process and serving as a personal and career mentor. He has shown me by example, what a successful and responsible scientific researcher is, and I am very thankful for the opportunities that he has created for me. I also want to thank Dr.MB Rao for serving as my academic chair and inspiring me to maintain an enthusiastic pursuit of learning forever. Thanks also to Dr. Rhonda Szczesniak, Dr. Hiedi Sucharew and Dr. Jane Khoury and Dr. Jessica Woo for serving on my qualifying and dissertation examination committee, my dissertation committee, and the faculty and staff in the UC Department of Environmental Health and the CCHMC Department of Biostatistics and Epidemiology for their time and support. I also want to thank Michael Wathen, Noell

Dedication

This work is dedicated to my parents and my brothers, who were my first teachers and mentors. I will always be grateful for the sacrifices they have made to help me succeed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Bayesian growth curve modeling: The analysis of growth curves has played a vital role in estimating the growth trajectory of populations as well as identifying critical factors corresponding to various shapes of those trajectories. Examples, which have motivated statistical developments ranging from hierarchical linear models to multivariate analysis methods, include psychological change over time, cardiovascular studies, and associations between adolescent moderate-vigorous physical activity and depressive symptoms in young adulthood and examination of longitudinal associations among cognition, function, and depression in Alzheimer's disease patients. In recent years, shape invariant modeling (SIM) has become an active area of research for non-parametric growth curve modeling, where a single function (or, curve) is transformed by scaling and shifting it to fit each subject usually through transformations. Lawton [126] , first proposed SIM called it self-modeling regression; in their approach, the function for the underlying shape is illustrated for various parametric functions. Later, Beath [23] developed a model to explain longitudinal growth patterns and extended the SIM to include time-dependent co-variates. As a type of SIM, the regression spline is expressed as a basis function consisting of a different set of knots; the resulting structure fitted as a nonlinear mixed effects model and parameters are typically estimated using maximum likelihood. This allows estimating the parameters for the between-subjects variation.

Research in longitudinal growth curve modeling utilizing the Bayesian inferential procedure is limited, and the wider application is hindered by the computational complexities involved in such models. Cole [51] proposed Super Imposition by Translation and Rotation (SITAR) model and expressed individual growth curves through three subject-specific parameters; named as size, tempo, and velocity. We propose a Bayesian Shape invariant model for longitudinal data and the performance of the model is evaluated with real and simulated data. Currently, instead of using a fixed knot for the spline function, we are trying to extend our model as "Bayesian free knot growth curve fitting" using reversible jump MCMC.

Power Comparisons in 2x2 Contingency Tables: Odds Ratio versus Pearson Correlation versus Canonical Correlation: It is an important inferential problem to test no association between two binary variables based on data. A Test-based on sample odds ratio is commonly used. We bring in a competing test based on the Pearson correlation coefficient. Odds ratio does not extend to higher order contingency tables, whereas Pearson correlation does. It will be a useful exercise to understand how Pearson correlation stacks against the odds ratio in $2x2$ tables. Another measure of association is the canonical correlation. In my work, we examine how competitive Pearson correlation is vis-à-vis odds ratio in terms of power in the binary context, contrasting further with both the Wald Z and Rao Score tests. We generate an extensive collection of joint distributions of the binary variables and estimate the power of the tests under each joint alternative distribution based on random samples. The consensus is none of the tests dominates the other.

Creating statistical computing tools for geo-coding and environmental exposure assessment:

Air pollution is a growing global challenge and has a severe, negative impact on human health. Vehicles, households, and industries emit a complex mixture of air pollutants, within which ambient particulate matter smaller than 2.5 μm ($PM_{2.5}$) are thought to have the greatest effect on human health. Prior epidemiologic evidence suggests short-term $PM_{2.5}$ exposure is associated with the development and exacerbation of several health problems. Children are more susceptible to $PM_{2.5}$ related health effects due to their immature immune system and ongoing development and growth. In my study, I am trying to create statistical computing tools for geo-coding and environmental exposure assessment. Daily ambient concentrations of $PM_{2.5}$ were estimated using residential addresses using a previously developed and validated the spatiotemporal model. Briefly, our $PM_{2.5}$ model is based on satellite-derived measurements of aerosol optical depth (AOD), a measure of the scattering of electromagnetic radiation due to aerosols in the atmosphere. These measurements calibrated using ground-based $PM_{2.5}$ monitoring and meteorological and land use data. Spatio-temporal data sets harmonized to a 1 x 1 km grid, and random forests were used to train a model to predict $PM_{2.5}$ concentrations.

1. Source-specific contributions of particulate matter to asthma-related emergency department utilization

In my first study, we downloaded all emergency department (ED), and urgent care (UC) visits for asthma between 2011 and 2015 within the Cincinnati Children's Hospital Medical Center's (CCHMC) electronic medical record (EHR) based on International Classification of Disease (ICD-10). Daily estimations of the source-specific contributions of different $PM_{2.5}$ sources were estimated using a chemical mass balance source apportionment model, and then we used a model-based clustering method to group days with similar source profiles. Using daily counts of pediatric, asthma-related hospital utilization for one urban county in Cincinnati, Ohio, USA, we then tested whether or not the type $PM_{2.5}$, as determined by cluster membership, significantly modified the effect of $PM_{2.5}$ on utilization.

2. Differential impact of acute $PM_{2.5}$ exposure on risk of stroke by stroke sub-type, sex and race: A case-crossover study

In my second study, we downloaded all stroke patients from the electronic medical record (EHR) based on the International Classification of Disease ICD9. Daily ambient concentrations of $PM_{2.5}$ were estimated using residential addresses and used a case-crossover design to investigate the association between short term $PM_{2.5}$ exposure and incidence of different sub types of stroke.

Chapter 2

Bayesian shape invariant model for longitudinal growth curve data.

Bayesian Superimposition by Translation and Rotation model for longitudinal growth curve data

Mohammad Alfrad Nobel Bhuiyan[1,2], Heidi Sucharew[2,3], Rhonda Szczesniak[2,3], Marepalli Rao[1,2], Jessica Woo[2,3], Jane Khoury[2,3], Md Monir Hossain[2,3]

[1]Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio, USA

[2]Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[3]Department of Pediatrics, University of Cincinnati, Cincinnati, Ohio, 45267, USA

Corresponding Author:

MD Monir Hossain

University of Cincinnati

E-mail: Md.Hossain@cchmc.org

Abstract

Growth curve modeling should ideally be flexible, computationally feasible, and allow for the inclusion of co-variates for better predictability and mechanistic explanations. The original Super Imposition by Translation and Rotation (SITAR) growth curve model, motivated by epidemiological research on the evolution of pubertal heights over time, fits the underlying shape function for height over age and estimates subject-specific deviations from this curve in terms of size, tempo, and velocity using maximum likelihood. This approach is effective in subsequent applications, but the estimation method does not provide uncertainty estimates for unknown parameters, and predictive ability has been largely unexplored. A more recent Bayesian implementation undertaken for multivariate SITAR modeling, and this approach requires multiple longitudinal outcomes and has added a computational burden. Modern growth curve studies of height measurements from children with attention-deficit hyperactivity disorder (ADHD) have gained importance in epidemiological research due to potentially adverse effects from stimulant medications. Motivated by a particular longitudinal study on the heights of 197 pediatric ADHD patients who began stimulant treatment at varying ages, we describe a Bayesian extension of the original SITAR model. We incorporate co-variate effects, assess mixing properties, and examine different spline formulations to model the underlying growth shape. As demonstrated by the real data application and simulation study, the Bayesian SITAR approach provides a natural, computationally feasible way to generate uncertainty estimates for treatment-outcome associations. We also discussed the future extensions to the approach.

Keywords: Bayesian inference, functional data analysis, growth curves, shape invariant

Introduction

The analysis of childhood growth curves has played a vital role in estimating the growth trajectory of populations as well as identifying critical factors corresponding to various shapes of those trajectories, such as sex. Indeed, early origins of growth curve modeling utilized cross-sectional growth curves, in which the population data were used to derive the growth patterns for various age and gender groups. The two widely known references of using cross-sectional data are the CDC growth chart [120] And WHO growth standards [153], which are mainstays in clinical care. More recently, growth curve analyses based on longitudinal data have allowed more accurate identification of growth patterns, since the longitudinal data allows incorporating within- and between-subject effects simultaneously [196].

Earlier work to fit growth curves was based on one of two parametric assumptions, namely logarithmic or exponential. Logarithmic curves assume a quick growth increase at the beginning, but the gains slowly disappear as time passes. Logarithmic growth curves are broadly applied in bacterial growth [211, 19], biodegradation [175], fitness and strength training, and learning ability. By contrast, the exponential curve assumes that growth is slower at the beginning with gains that are more rapid over time. [112] explained growth curve modeling through exponential curves.[111] proposed a different model adding a linear term which was fitted by [31] and found a poor fit of the data based on the systematic variation of the residual [23]. Moreover,[32] also analyzed the model proposed by Count [56] and found an exponential model as a better model. Different parametric growth curve models proposed over the last two decades, such as the linear model, reciprocal model, logistic model, Gompertz model, and the Weibull model. [142], [197] explored other approaches of growth curves and found a poor fit.[198] Explained growth data analysis using polynomial regression. One limitation of polynomial regression that it requires a higher degree of polynomial to provide an adequate fit with the resulting coefficients without having any significant interpretation. Alternative approaches to analyze growth curve data were proposed by [84] , [152] [146] ,[167]. Nonparametric models using regression splines to model the underlying shape function have been shown to decrease bias, thereby improving the estimation of subject-specific effects [188]. Furthermore, regression splines, such as natural splines, have been shown to provide a better-localized fit to the mean response, compared to global polynomials [122].

Equally important to finding an appropriate model to depict growth patterns is understanding the risk factors that contribute to adverse growth. The growth model is essential for designing an intervention trial or to increase public health awareness surrounding potential benefits and adverse effects of various environmental exposures and growth patterns. Furthermore, having interpretable estimates for growth characteristics, such as peak height velocity, can ameliorate confounding in epidemiologic studies.[178]. Other examples, which have motivated statistical developments ranging from hierarchical linear models to multivariate analysis methods, include psychological change over time [101], cardiovascular changes [131], associations between adolescent moderate-vigorous physical activity and depressive symptoms in young adulthood [42], associations among timing of sexual victimization and timing of drinking behavior[92], examination of longitudinal associations among cognition, function, and depression in Alzheimer's Disease patients [204],and describing change in personality trait [108]

In recent years, shape invariant modeling (SIM) has become an active area of research for non-parametric growth curve modeling, where a single function (or, curve) is transformed by scaling and shifting it to fit each subject usually through affine transformations.[126], Who first proposed SIM called it self-modeling regression; in their approach, the function for the underlying shape illustrated for various parametric functions. Later,[23] developed a model to explain longitudinal growth patterns and extended the SIM to include time-dependent covariates.[51] Extended the model by changing the sign of the velocity parameter and named it SITAR (Superimposition by Translation and Rotation). As a type of SIM, the regression spline expressed as a basis function consisting of a different set of knots; the resulting structure fitted as a nonlinear mixed effects model and parameters are typically estimated using maximum likelihood. This allows estimating the parameters for the between-subjects variation. Based on the underlying pattern of the data, various shape invariant models have been proposed. When the data has logarithmic growth, SIM uses for the log-transformed data.

Bayesian growth curve modeling has also seen similar progress with many applications to real datasets as well as longitudinal growth datasets [19],[10], [69], [159], [177]. One of the main advantages of a Bayesian approach is that it generates the uncertainty estimates (i.e., the estimate for the variance) for all unknown parameters naturally since each parameter explained by a probability distribution. Other advantages include the use of prior probability distributions to assimilate information from previous studies or experts opinion and allows control of confounding; having posterior probabilities is an easily interpretable alternative to p-values; in hierarchical modeling, incorporating latent variables such as an individual's true disease status in the presence of a diagnostic error. Moreover, MCMC methodology facilitates the implementation of Bayesian analyses of complex data sets containing missing observations and multidimensional outcomes [69]. Due to this flexibility and better prediction of the exposure-outcome relationship, researchers are becoming more interested in Bayesian modeling. Notable work in Bayesian growth curve modeling includes the multivariate extension by [196] the original SITAR model.

A brief outline of the paper is as follows. In Section 2, a brief description of the original SITAR model and the interpretation of various model parameters provided. In a subsection, we provide some descriptions for the spline function used in SITAR and how this model connected to GAM. Section 3 describe the Bayesian implementation of the SITAR model with and without subject-specific covariates, and also the DAG representations of these models. Section 4 illustrates the MCMC implementation; the specification for the prior distributions; the full conditional distribution and the posterior distribution for each model parameter; and how the assessment of model performance. The full derivation of posterior distributions is given in the Appendix. Applications with the real data are provided in Section 5, and with simulated data in Section 6. The Final Section includes the discussion and some proposals to the future extension.

The Superimposition by Translation and Rotation model

Following the notation from [23], the SITAR model can be expressed as:

$$y_{ij} = \gamma_{i2} + \boldsymbol{h}(\frac{t_{ij} - \gamma_{i1}}{exp(-\gamma_{i3})}) + \epsilon_{ij}; i = 1, \ldots N, \text{and } j = 1, \ldots T_i. \tag{2.1}$$

Here, the y_{ij} is the growth measure of i^{th} child at the j^{th} time points which corresponds to age (in years) in our motivating example.Subject-specific coefficients $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2}, \gamma_{i3})$ enable each individual's growth trajectory to be aligned to a common growth curve, $h(\cdot)$, via transformations to the $x-$ and $y-$axes.In this formulation, we will estimate $h(\cdot)$ using a spline function and let ϵ_{ij} be measurement error.

The goal is to estimate the subject-specific vector γ_i such that the corresponding individual growth curve form deviations from the average curve $h(\cdot)$. Following previously described work by Cole and others in equation (1), γ_{i2} is termed as *Size.* γ_{i2} can also be interpreted as subject-specific shift up or down in the spline curve along the response axis. γ_{i2} is a random intercept term; when the response measure is height, γ_{i2} is larger for taller children and smaller for shorter children. γ_{i1} termed *Tempo*, which is a random time intercept and corresponds to differences in the timing of the growth spurt. This subject-specific left-right shift in the growth curves positive for late puberty and negative for early. The scaling factor within the spline function, γ_{i3} is termed *Velocity* and corresponds to differences in the duration of the growth spurt between individuals. The Velocity parameter shrinks or stretches the time scale [51].

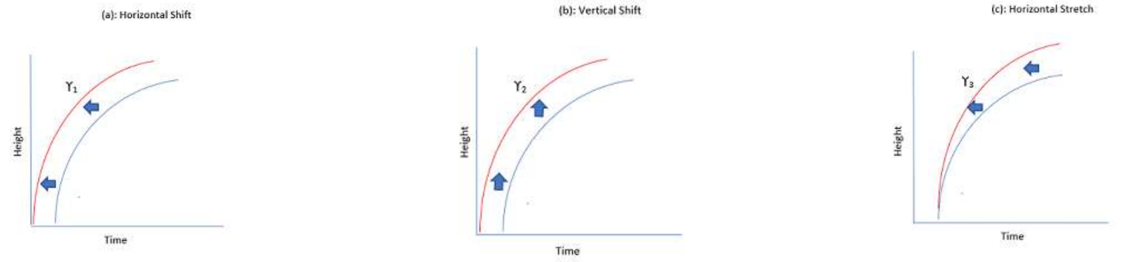


Figure 2.1: Schematic representation of Superimposition by Translation and Rotation model with Horizontal shift(γ_1) , vertical shift(γ_2) and stretch(γ_3)

Bayesian Implementation of Superimposition by Translation and Rotation model

The above SITAR model presented in equation (1) can be written with basis representation as follows;

$$y_{ij} = \gamma_{i2} + \boldsymbol{Z}_{ij}^T \boldsymbol{\beta}_{(\kappa+2)} + \epsilon_{ij}; i = 1, \ldots N, \text{and } j = 1, \ldots T_i. \tag{2.2}$$

Where,

$$\boldsymbol{Z_{ij}} = \boldsymbol{B}(exp(\gamma_{i3})(t_{ij} - \gamma_{i1})),$$

$$\boldsymbol{\gamma_i} = (\gamma_{i1}, \gamma_{i2}, \gamma_{i3}), \boldsymbol{\gamma_i} \sim N_3(\boldsymbol{0}, \boldsymbol{\Sigma_\gamma}), \text{and } \boldsymbol{\epsilon_{ij}} \sim N(0, \sigma^2).$$

$\boldsymbol{Z_{ij}}$ is the basis of the natural cubic spline, evaluated at $(exp(\gamma_{i3})(t_{ij} - \gamma_{i1}))$. Thus, $\boldsymbol{z}_{ijk}$ is a vector of length $\boldsymbol{\kappa} + 2$, and $\boldsymbol{\beta}$ is the regression coefficient vector of same length. Here, $\boldsymbol{\kappa}$ is the number of inner knots and 2 represents the boundary knots. so the natural cubic spline has $\boldsymbol{\kappa} + 2$ independent coefficients. Subject-specific $\boldsymbol{\gamma}_i$ is assumed to have multivariate normal of order 3 with 0 mean vector and $\boldsymbol{\Sigma}_\gamma$ variance-covariance matrix.We assume that $\boldsymbol{\epsilon}_{ij}$ is independently normally distributed with mean 0 and variance $\boldsymbol{\sigma}^2$, and also independent of $\boldsymbol{\gamma}_i$.

The other distributions for ϵ_{ij} such as Student's-t can also be considered depending on the type of growth data.

With subject-specific covariates: In many growth curve analyses, there is need for the inclusion of covariates for better predictability, as well as for better explanation of growth mechanisms. For example, it may be of interest to know how the gender difference affects the growth patterns, or how the medication at early age for a specific disease condition affects the growth at later ages specifically by size, tempo, and velocity. The subject-specific covariates can be included in the model specified for $\boldsymbol{\gamma}_i$ with a non-zero mean vector. If we assume (p-1) subject-specific covariates, the mean and the variance of $\boldsymbol{\gamma}_i$ becomes,

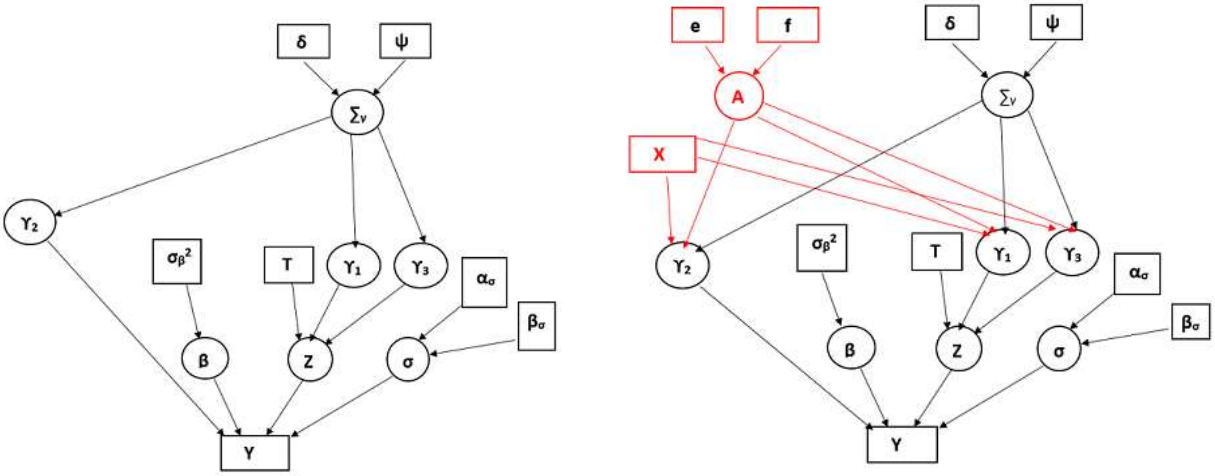
$$\boldsymbol{\gamma_i} \sim N_3(\boldsymbol{AX_i}, \boldsymbol{\Sigma}_\gamma),$$

where,

$$\boldsymbol{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1p} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2p} \\ \alpha_{31} & \alpha_{32} & \dots & \alpha_{3p} \end{pmatrix}, \boldsymbol{X_i} = \begin{pmatrix} x_{i1}, & x_{i2}, & \dots, & x_{ip} \end{pmatrix}^T, \text{ and } \boldsymbol{\Sigma}_\gamma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix}.$$

The first column of the regression coefficient matrix A is for the intercept, and the remaining columns are for each $(p-1)$ covariate coefficients. Similarly, the first column of the design matrix X contains the value 1.

Directed acyclic graph (DAG): DAG is a graphical representation of a hierarchical model which shows how the observed data and the unobserved parameters are conditionally dependent on each other. In the graph, the circle indicates the stochastic node (or, the unobserved parameters that need to be estimated), and the rectangle indicates observed data or the hyper-parameters where they were assigned to fixed values apriori. When the Bayesian hierarchical model has a complex dependency structure, DAG help to better visualize the model as a whole, and the derivation of the posterior distribution for each stochastic node.



(a) Without Covariates

(b) With Covariates

Figure 2.2: Graphical representation of the model without and with subject-specific covariates. The subject-specific covariates are shown in red.

MCMC Implementation

Prior distributions :

The prior distribution for all model parameters are assumed to be independent apriori, and follow an uninformative flat probability distribution in general. The residual variance was assumed to follow an inverse gamma distribution with fixed shape and scale parameters such that $\sigma^2 \sim IG(0.001, 0.001)$. Alternative prior distributions for residual variance parameter can also be used following [82]. For covariate coefficients and the basis coefficient, we have assumed $\alpha \sim N(0, 1000)$ and $\beta \sim N(0, 1000)$, respectively. The vector α was defined after stacking the A matrix. The variance-covariance matrix for $\boldsymbol{\gamma}_i$ is assumed to follow an inverse Wishart distribution with 3 degree of freedom and 0.01 scale parameter, $\Sigma_\gamma \sim IW(3, 0.01)$.

Full conditional distributions :

In the Bayesian framework the joint posterior distribution is proportional to the product of the likelihood function and the prior distributions. Therefore, the full posterior distribution for the model in (2) with subject-specific covariates is as follows;

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_\gamma|y) \propto \prod_{i=1}^{N}\prod_{j=1}^{T_i} N(y_{ij}|\gamma_{i2} + z_{ij}(\gamma)^T\beta, \sigma^2) \prod_{i=1}^{N} N_3(\boldsymbol{\gamma}_i \mid Ax_i, \Sigma_\gamma)$$

$$\prod_{i=1}^{3p} N(\alpha \mid 0, \sigma_\alpha I_{n_p}) \times \prod_{k=1}^{k+2} N(\beta_k|0, \sigma_\beta^2) \times IG(\sigma^2|\alpha_\sigma, \beta_\sigma) \times IW(\Sigma_\gamma|\delta, \psi) \quad (2.3)$$

We follow the block update procedures in MCMC iterations whenever the posterior distribution has a standard form. It improves the convergence and mixing as well, and also saves computational time. Following the Gibbs sampling procedure, the updates for each parameters are as shown below.For the sake of simplicity, covariate effects and subject-specific notation (i) are omitted for portions of the updates.

Updating $\boldsymbol{\alpha}$:

$$p(\boldsymbol{\alpha}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_\gamma, y) \propto \mathbf{N}(\bar{\Sigma}_\alpha[(X^T \otimes \Sigma_\gamma^{-1})\gamma_{vec}], [X^T X \otimes \Sigma_\gamma^{-1} + I_{n_p}\sigma_\alpha^2]^{-1})$$

where, $\bar{\Sigma}_\alpha = [X^T X \otimes \Sigma_\gamma^{-1} + I_{n_p}\sigma_\alpha^2]$, and γ_{vec} is a vector of stacked $\boldsymbol{\gamma}^T$.

Updating $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}|\boldsymbol{\alpha}, \sigma^2, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_\gamma, y) \propto \mathbf{N}(\sigma^{-2}\bar{\Sigma}_\beta \mathbf{Z}^{\mathbf{T}}(\mathbf{y} - \gamma_{\mathbf{2}}), \sigma^{-2}\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \sigma_\beta^{-2} I_{k+2})$$

where, $\bar{\Sigma}_\beta = [\sigma^{-2}\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \sigma_\beta^{-2} I_{k+2}]$.

Updating $\boldsymbol{\sigma}^2$:

$$p(\sigma^2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_\gamma, y) \propto \boldsymbol{\mathit{Inverse\text{-}Gamma}}\left(a + \frac{\sum T_i}{2}, b + \frac{\sum_{ij}\{y_{ij} - \gamma_{i2} - \mathrm{B}(exp(\gamma_{i3})(t_{ij} - \gamma_{i1}))\}^2}{2}\right).$$

Updating $\mathbf{\Sigma}_\gamma$(without covariate):

$$p(\mathbf{\Sigma}_\gamma|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\sigma^2,y) \propto \boldsymbol{Inverse\text{-}Wishart}\left(\delta+n,\left(\psi^{-1}+\gamma^{\mathrm{T}}\gamma\right)^{-1}\right)$$

where, $\boldsymbol{\delta}$ is degrees of freedom, and $\boldsymbol{\psi}$ is the scale matrix.

Updating $\mathbf{\Sigma}_\gamma$ (with covariate):

$$p(\mathbf{\Sigma}_\gamma|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\sigma^2,y) \propto \boldsymbol{Inverse\text{-}Wishart}\left(\delta+n,\left(\psi^{-1}+(\gamma-XA^T)^{\mathrm{T}}(\gamma-XA)\right)^{-1}\right)$$

where, $\boldsymbol{\delta}$ is degrees of freedom and $\boldsymbol{\psi}$ is the scale matrix.

Moreover, for the full conditional of the subject-specific parameter, $\boldsymbol{\gamma}_i$ is given by,

$$\boldsymbol{\gamma}_i|y_i,x_i,\boldsymbol{\alpha},\boldsymbol{\beta},\sigma^2,\mathbf{\Sigma}_\gamma \propto N_3(\gamma_i|Ax_i,\Sigma_\gamma)\times\prod_j N(y_{ij}|\gamma_{i2}+\boldsymbol{B}(exp(\gamma_{i3})(t_{ij}+\gamma_{i1}))\beta,\sigma^2)).$$

For updating $\boldsymbol{\gamma}_i$, we used a random walk Metropolis-Hastings (M-H) algorithm to generate posterior samples for the subject-specific effect. The candidate samples are generated from multivariate $Student's-t$ distribution with 5 degrees of freedom and mean at the current value. The variance parameter of the multivariate $Student's-t$ was appropriately tuned to ensure that the acceptance rate of candidate samples to posterior samples in M-H step was around 20-30%.

Convergence, Mixing, and Identifiability

Before summarizing the MCMC samples (or, posterior samples) as posterior mean, median or highest posterior region; it is important to check that the posterior samples for each parameter are converging, mixing well and have less auto-correlation. The convergence of

MCMC samples indicates how close we are to the true posterior distribution, and mixing indicates how well the parameter space is explored. There are different ways of checking these, namely, trace plot, autocorrelation plot, QQ plot, Brooks plot, Gelman-Rubin test, etc. A simple exploration of the trace plot gives an insight into the characteristics of the MCMC samples. Trace plots are produced for each parameter and checked whether different starting values lead to better mixing and convergence, saving the computational time as well.

Sensitivity Analysis

In Bayesian inference, it is recommended to check the validity of the posterior estimates for the choice of other prior distributions, or other hyper-parameters values. Sensitivity analysis ensures this purpose. Even though we assigned a noninformative flat prior distributions to all parameters, we checked the sensitivity of the estimates by changing the values for hyper-parameters. We also did the sensitivity analysis following the method proposed by [113] called *perturbation function*. We perturbed the model with different spline function (with and without covariates) and analyzed the sensitivity of the model. Two spline functions: basis spline and natural cubic spline functions were considered for this purpose.

Real Data Example: ADHD Children

Longitudinal data from a retrospective chart review of heights and weights for 197 ADHD (Attention-deficit Hyperactivity Disorder) children who visited a community-based pediatric primary care practices in Cincinnati (Ohio, USA) was collected. The children were in the age range: 1-17 years, and under stimulant medications to effectively reduce symptoms of ADHD. The objective of the original study was to evaluate how the age at the start of stimulant medication may have impacted child growth trajectories. The original ADHD data had 6,134-time points recorded on 197 patients; among them, only 3084 height measures were available for the analysis. The longitudinal study design is appeared to be unbalanced. Figure 3 shows the growth heights of each patient at various ages. The study enrolled 138 males (70%). For the 197 unique patients the age range was (1.37, 16.76) with the mean age 9.3 years, and the height range (76.2, 183.2) with the mean height 134.4 cm (Table 2.1). Each patient was prescribed a stimulant medication at a certain visit. The mean age at stimulant medication start was 7.9 years and ranged from 4.2 to 12.3 years.

Table 2.1: Summary statistics for the 197 ADHD patients

Variable	Min	$1st.Q$	$Median$	$Mean$	$3rd.Q$	Max	SD
Age	1.363	7.249	9.443	9.263	11.369	16.764	3.001
Height	76.2	122.6	134.6	134.4	147.3	183.2	18.86
Wight	9.072	21.32	28.12	31.71	38.33	114.3	31.51
Start age of stimulant medication	4.17	6.75	7.86	7.949	9.120	12.330	1.6

We applied the Bayesian SITAR model to the ADHD data and sought to describe the size, tempo, and velocity parameters relevant to the height growth of these ADHD patients. And simultaneously, we describe the model fitting procedure and its relative performances in comparison to the frequentist SITAR model in details.

MCMC implementation: The SITAR model requires to specify the Spline function with a specific number of knots. We used both B-spline and the natural cubic spline functions for checking their relative performances. As for specifying the number of knots, we used eight equally spaced knots for both spline function, six interior and two exteriors. The number of knots was determined based on a compromise between optimizing a fit criterion and the computational burden — the prior distributions specified according to prior Section. The MCMC implementation followed a block update procedure with a mix of the Gibbs and M-H algorithm. In implementation, the inverse-Wishart distribution for the posterior distribution for $\mathbf{\Sigma}_\gamma$ was redefined as a scaled inverse-Wishart distribution for correct estimation of correlation matrix as well as for quick convergence.

We ran multiple chains with a relatively longer burn-in period. The model was runs for 500,000 iterations with 90% burn-in samples. This yielded 50,000 samples for posterior inference. We further reduced the posterior samples to size 5,000 after thinning by parameter 10 for reducing the autocorrelation in posterior samples. All the results reported in this manuscript were derived from these 5,000 posterior samples. Convergence was checked by examining the trace plot of the posterior samples for each parameter and by using [83] test. We also checked the autocorrelation from the autocorrelation plot, and it showed a much faster rate of decreasing towards zero with the increasing lag values. The variance-covariance matrix of the proposal density in the M-H algorithm was tuned accordingly so that the acceptance rate was approximately 20%.

Contour Plot: Mixing well of posterior samples is an important property to ensure that there is no specific trend in posterior samples among the parameters, and independence is maximized. We randomly selected three patients and their posterior samples for a horizontal shift ($\boldsymbol{\gamma}_1$), and stretch ($\boldsymbol{\gamma}_3$)parameters plotted in Figure 2.3. The plot shows there is no specific trend in posterior samples for these two parameters, and they scattered around the center, indicating a low correlation.

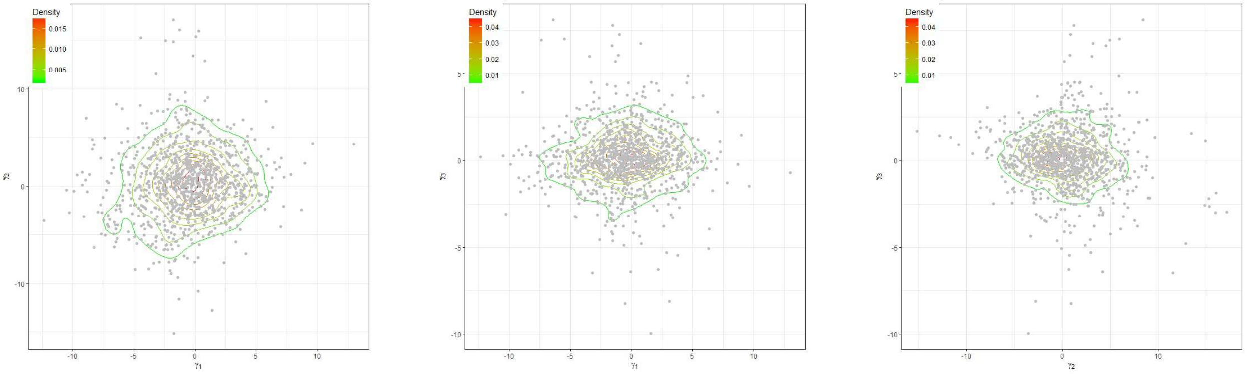


Figure 2.3: Contour plot of posterior samples for horizontal shift (γ_1) and stretch (γ_3) parameters for randomly selected three patients.

Results

SITAR growth curve analysis utilizes the biology of growth. This model relies on the concept of growth and assumes a linear relationship between chronological time and growth. An individual growth spurt can progress or slow down concerning time. This progress and delay is reflected in tempo parameter. Moreover, this processor delay can more or less proceed over time, which is reflected in the velocity parameter. The analysis is modeled on both the height scale (i.e., the size parameter) and the age scale, and in this sense, it mimics biology as the appropriate age scale is developmental age, not chronological age.

We compared the frequentist SITAR model with Bayesian SITAR model applying on ADHD patient data.

Table 2.2: SITAR Fixed parameter value

Parameter	SITAR without	Bayesian SITAR without
	Covariates	Covariates
γ_1	4.46	4.42
γ_2	0.54	0.69
γ_3	0.07	0.06
	SITAR with	Bayesian SITAR with
	Covariates	Covariates
γ_1	4.47	4.44
$\gamma_{1.Gender}$	0.05	0.07
$\gamma_{1.ageFirstMed}$	0.01	0.03
γ_2	0.57	0.53
$\gamma_{2.Gender}$	1.19	1.16
$\gamma_{2.ageFirstMed}$	0.18.	0.20
γ_3	0.08	0.06
$\gamma_{3.Gender}$	-0.07	-0.09
$\gamma_{3.ageFirstMed}$	-0.01	-0.04

We were interested to see the pattern of the growth velocity of the ADHD patient children and the effect of medication on their growth. Here, age at peak velocity (APV) is 12.38 years, and peak velocity(PV) is 8.76 cm.

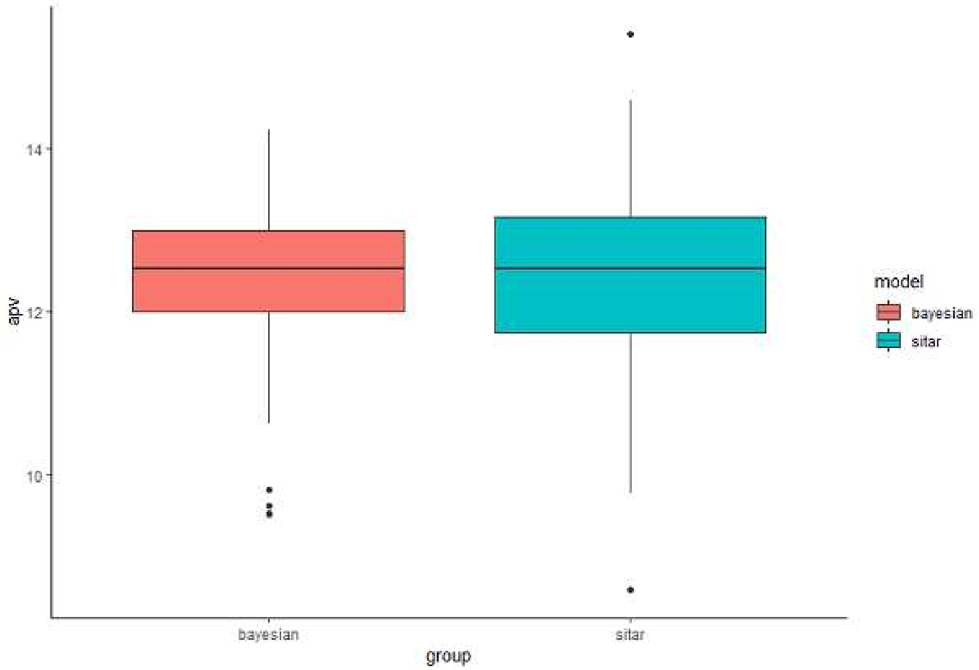


Figure 2.4: Comparison between Age at peak velocity in bayesian and frequentist method

Model Selection and Predictive ability:

We implemented both the original SITAR model and the Bayesian SITAR model on the ADHD data. Two spline functions, B-spline and natural cubic spline, was used in the Bayesian SITAR model. In all spline functions, eight knots used. We calculated the root mean square prediction error (RMSPE) for both the SITAR model and the Bayesian SITAR model to check their predictive ability. It appeared that the Bayesian SITAR model with natural cubic spline was performing better in both of these criteria (Table 2.3).

Table 2.3: Root mean square prediction error (RMSPE) values for the ADHD data

	SITAR without Covariates	Bayesian SITAR without Covariates	SITAR with Covariates	Bayesian SITAR with Covariates
$RMSPE$	0.20	0.19	0.18	0.15

Cross-validation:

For checking the predictive ability of Bayesian SITAR model, we adopted the cross-validation method where four patients with two early start ages of stimulant medication from each sex and two late start ages from each sex were randomly left out from the analysis as a validation sample. The remaining 193 patients constructed the derivation sample were analyzed using the Bayesian SITAR model to generate the mean predictive curve. Figure 2.5 plots the mean predictive curve with 95% credible interval from the derivation sample. The four patients from the validation sample plotted in the same figure with red color for females and blue for males. The results ensure that the randomly picked four patients with various start ages of stimulant medication and sex to lie within 95% credible interval of the mean predictive curve.

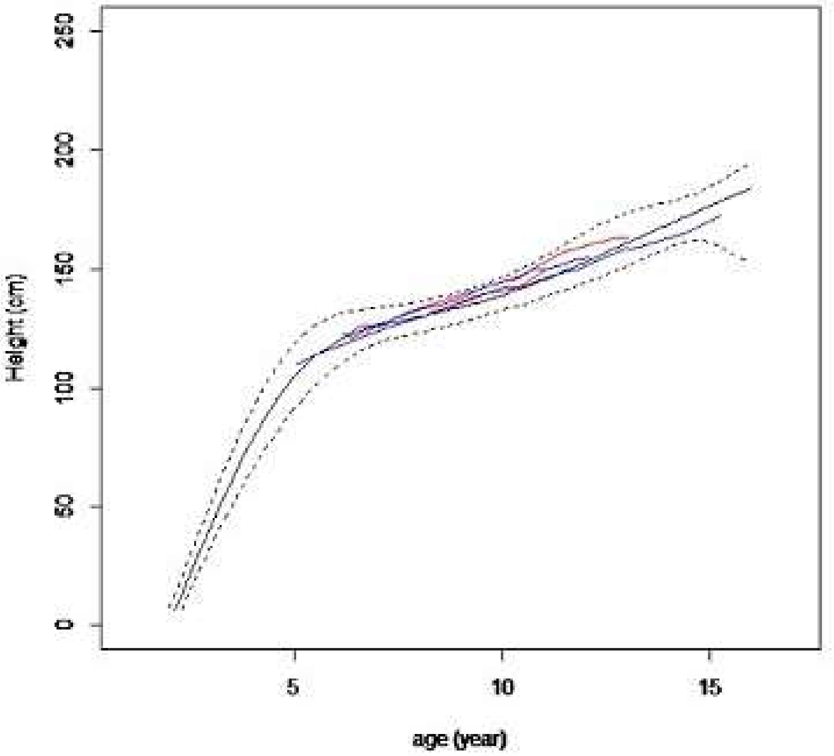


Figure 2.5: Predicted mean growth curve for the 193 ADHD patients with 95% credible interval. The remaining four patients were plotted with red color for females and blue for males

Simulated Data Example

The simulated experiment was designed to assess the relative performances of the Bayesian SITAR model concerning its frequentist counterpart. An R package "AGD (Analysis of Growth Data)" Version 0.35 was used to simulate the height data using the LMS method to obtain normalized growth. Using the growth charts data for the reference population, AGD computes heights in cm conditioning on age, sex, and the growth percentile. The current version of 'AGD' includes data for three reference populations: the United States, The Netherlands, and WHO data from multiple countries. We used the United States population as our reference population in simulating heights. To ensure wide variability and some levels of grouping patterns existed in the simulated heights, we generated data from three groups. In the first group, growth percentile range was set to (70 - 90)%; growth percentile range for the second group set to (40 - 60)%; and the third group set to (10 - 30)%. In each group, the proportions for males and females were equal.

The steps involved in the data generation process are:

1. Generate the height, y_{ij} using AGD package in R with the United States as a reference population. Generate 150 subjects such that: a) 50 from each group, and b) 25 males and 25 females in each group

2. , All subjects height data, were simulated for ages 5-20 years with six months increase

3. The simulated data in steps 1 and 2 was used in the SITAR package in R to get estimates of γ^{SITAR}, β^{SITAR}, $\sigma^{2,SITAR}$, and Σ_{γ}^{SITAR}

4. Using these estimates, a new y_{ij} for height was generated following the equation:

$$y_{ij}^{new} = \gamma_{i2}^{SITAR} + BS(exp(\gamma_{i3}^{SITAR})(t_{ij} - \gamma_{i1}^{SITAR}))\beta^{SITAR} + noise(0, \sigma^{2,SITAR}),$$

Where, the random noise generated from a normal distribution with mean 0 and variance

$\sigma^{2,SITAR}$, and BS is for the B-spline function with 8 knots. This random noise can be interpreted as a measurement error or that the errors that are still unaccounted for by the SITAR parameters. Step 4 can be run many times, depending on the desired number of realizations achieved for the simulated dataset. We run it five times to generate five realizations. Figure a) and b) plots the data generated from the AGD package and one realization after adding the random noise, respectively.

In the evaluation of model performance for the simulated datasets, we used the estimates from the SITAR package at Step 3: γ^{SITAR}, β^{SITAR}, $\sigma^{2,SITAR}$, and Σ_{γ}^{SITAR}, as true estimates of all model parameters.

As similar to the real data example, we implemented both the original SITAR model and the Bayesian SITAR model with Two spline functions, B-spline and natural cubic spline, on the simulated data. The number of knots was assigned to eight. We checked the RMSPE values for the most parsimonious model and for checking the predictive model ability, respectively. It appeared that the Bayesian SITAR model with natural cubic spline was performing better in both of these criteria (Table 2.4).

Table 2.4: Root mean square prediction error (RMSPE) for the simulated data

	SITAR without	Bayesian SITAR without	SITAR with	Bayesian SITAR with
	Covariates	Covariates	Covariates	Covariates
$RMSPE$	0.24	0.21	0.22	.18

Discussion

The original SITAR model is effective in explaining subject-specific deviations in terms of size, tempo, and velocity from the underlying shape curve. The model is currently unable to produce standard error estimates for subject-specific deviations for size, tempo, and velocity parameters. Although bootstrap methods are often used to estimate the standard error, this is a post-hoc analysis and often fails to estimate the true uncertainty specifically when the model has a complex structure, i.e., a non-linear mixed effects model. Estimating the standard error for estimates of subject-specific deviations for size, tempo, and velocity in Bayesian SITAR model comes naturally from the posterior samples for the subject-specific parameter (γ). Although we considered vague prior distributions in our analyses, the model is flexible to assimilate information from previous studies or expert opinion through specifying informative prior distributions.

We applied the Bayesian SITAR model with two spline functions and with an optimum number of knots (i.e., 8) to real and simulated data sets, and compared the results with the original SITAR model. We observed that for both data sets, the Bayesian SITAR model with natural cubic spline function has better predictive ability than the original SITAR model. With the real data application.

In our future work, we plan to examine the Bayesian SITAR model in finding clustering patterns in shape invariant parameters tempo (γ_1), size (γ_2), and velocity (γ_3). In many applications of shape invariant models in growth curve modeling, finding the group of subjects with similar growth patterns in terms of size, tempo, and velocity have real significance. Another extension can include but not limited to, extending the model to free knot natural cubic spline, where we will utilize the Bayesian adaptive regression spline (BARS). It has been shown earlier that BARS provide a parsimonious fits [63].

Acknowledgement

The project described was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health, under Award Number 5UL1TR001425-04. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Coauthor (RS) supported by grant K25 HL125954 from the National Heart, Lung and Blood Institute of the National Institutes of Health.

2.1 Appendix

Full Conditional Distribution:

Updating α :

$$
\begin{aligned}
p(\alpha|\beta,\sigma^2,\gamma,\Sigma_\gamma,y) &\propto \prod_{ij} N(y_{ij}|\gamma_{i2}+z_{ij}^T\beta_j,\sigma^2)\prod_i N(\gamma \mid AX,\Sigma_\gamma) \\
&\quad \times \prod_i N(\beta_i|0,\sigma_\beta^2 I_{k+2}) \times \prod_{i=1}^{n} N(\alpha \mid 0,\sigma_\alpha I_{n_p}) \times IG(\sigma^2|\alpha_\sigma,\beta_\sigma) \times IW(\Sigma_\gamma|\delta,\psi) \\
&\propto \prod_i N(\gamma \mid AX,\Sigma_\gamma) \times \prod_{i=1}^{n} N(\alpha \mid 0,\sigma_\alpha I_{n_p}) \\
&\propto \mathbf{N}_{3n}(\gamma_{vec}|[X\otimes I_3]\alpha, I_n\otimes\Sigma_\gamma)\mathbf{N}_2(\alpha|0,\sigma_\alpha^2 I_{n_p}) \\
&\propto exp\left\{-\frac{1}{2}[\gamma_{vec}-(X\otimes I_3)\alpha]^T(I_n\otimes\Sigma_\gamma^{-1})[\gamma_{vec}-(X\otimes I_3)\alpha]+\alpha^T\sigma^{-2}I_2\alpha\right\} \\
&\propto exp\left\{-2\alpha^T(X\otimes I_3)^T(I_n\otimes\Sigma_\gamma^{-1})\gamma_{vec}+\alpha^T(X\otimes I_3)^T(I_n\otimes\Sigma_\gamma^{-1})(X\otimes I_3)\alpha+\alpha^T\sigma^{-2}I_2\alpha\right. \\
&\propto exp\left\{-2\alpha^T(X^T\otimes\Sigma_\gamma^{-1})\gamma_{vec}+\alpha^T(X^TX\otimes\Sigma_\gamma^{-1})\alpha+\alpha^T\sigma^{-2}I_2\alpha\right\} \\
&\propto \mathbf{N}(\bar{\Sigma}_\alpha[(X^T\otimes\Sigma_\gamma^{-1})\gamma_{vec}],[X^TX\otimes\Sigma_\gamma^{-1}+I_{n_p}\sigma_\alpha^2]^{-1})
\end{aligned}
$$

Updating β :

$$\begin{aligned}
p(\beta|\alpha,\sigma^2,\gamma,\Sigma_\gamma,y) &\propto \prod_{ij} N(y_{ij}|\gamma_{i2}+z_{ij}^T\beta_j,\sigma^2)\prod_i N(\gamma \mid 0,\Sigma_\gamma)\\
&\quad\times \prod_i N(\beta_i|0,\sigma_\beta^2 I_{k+2}) \times \prod_{i=1}^{n} N(\alpha \mid 0,\sigma_\alpha I_{n_p}) \times IG(\sigma^2|\alpha_\sigma,\beta_\sigma) \times IW(\Sigma_\gamma|\delta,\psi)\\
&\propto \prod_{ij} N(y_{ij}|\gamma_{i2}+z_{ij}^T\beta_j,\sigma^2)\prod_i N(\beta_i|0,\sigma_\beta^2 I_{k+2})\\
&\propto \exp -\left\{||\mathbf{y}-\gamma_{\mathbf{2}}-\mathbf{Z}\beta||^2\right\}/2\sigma^2 \times \exp -\left\{||\beta||^2\right\}/2\sigma_\beta^2\\
&\propto \exp -\left\{\sigma^{-2}\beta^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\beta - 2\sigma^{-2}(\mathbf{y}-\gamma_{\mathbf{2}})^{\mathrm{T}}\mathbf{Z}\beta + \sigma_\beta^{-2}\beta^{\mathrm{T}}\beta\right\}/2\\
&\propto \exp -\Big\{\beta^{\mathrm{T}}\underbrace{\left[\sigma^{-2}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}+\sigma_\beta^{-2}I_{k+2}\right]}_{\bar\Sigma_\beta^{-1}}\beta - 2\sigma^{-2}(\mathbf{y}-\gamma_{\mathbf{2}})^{\mathrm{T}}\mathbf{Z}\beta\Big\}/2\\
&\propto \exp -\left\{\beta^{\mathrm{T}}\bar\Sigma_\beta^{-1}\beta - 2\sigma^{-2}\beta^{\mathrm{T}}\bar\Sigma_\beta^{-1}\bar\Sigma_\beta\mathbf{Z}^{\mathrm{T}}(\mathbf{y}-\gamma_{\mathbf{2}})\right\}/2\\
&\propto \exp -\left\{[\beta-\sigma^{-2}\bar\Sigma_\beta\mathbf{Z}(\mathbf{y}-\gamma_{\mathbf{2}})]^{\mathrm{T}}\bar\Sigma_\beta^{-1}[\beta-\sigma^{-2}\bar\Sigma_\beta\mathbf{Z}^{\mathrm{T}}(\mathbf{y}-\gamma_{\mathbf{2}})]\right\}/2\\
&\propto \mathbf{N}(\sigma^{-2}\bar\Sigma_\beta\mathbf{Z}^{\mathbf{T}}(\mathbf{y}-\gamma_{\mathbf{2}}),\sigma^{-2}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}+\sigma_\beta^{-2}I_{k+2})
\end{aligned}$$

Updating σ^2 :

$$\begin{aligned}
p(\sigma^2|\alpha,\beta,\gamma,\Sigma_\gamma,y) &\propto \prod_{ij} N(y_{ij}|\gamma_{i2}+z_{ij}^T\beta_j,\sigma^2) \times \prod_{i=1}^{n} N(\alpha \mid 0,\sigma_\alpha I_{n_p})\prod_i N(\gamma \mid 0,\Sigma_\gamma)\\
&\quad\times \prod_i N(\beta_i|0,\sigma_\beta^2 I_{k+2}) \times IG(\sigma^2|\alpha_\sigma,\beta_\sigma) \times IW(\Sigma_\gamma|\delta,\psi)\\
&\propto \prod_{ij} N(y_{ij}|\gamma_{i2}+z_{ij}^T\beta_j,\sigma^2) \times IG(\sigma^2|\alpha_\sigma,\beta_\sigma)\\
&\propto \mathbf{IG}\left((a+\frac{N}{2}, b+\frac{\sum\limits_{ij}\{y_{ij}-\gamma_{i2}-exp(\gamma_{i3})\mathrm{B}(t_{ij}+\gamma_{i1})\}^2}{2}\right)
\end{aligned}$$

Updating Σ_γ :

$$\begin{aligned}
p(\Sigma_\gamma|\alpha,\beta,\gamma,\sigma^2,y) &\propto \prod_{ij} N(y_{ij}|\gamma_{i2}+z_{ij}^T\beta_j,\sigma^2)\prod_i N(\gamma \mid AX_i,\Sigma_\gamma) \\
&\quad\times \prod_i N(\beta_i|0,\sigma_\beta^2 I_{k+2}) \times \prod_{i=1}^n N(\alpha \mid 0,\sigma_\alpha I_{n_p}) \times IG(\sigma^2|\alpha_\sigma,\beta_\sigma) \times IW(\Sigma_\gamma|\delta,\psi) \\
&\propto \prod_i N(\gamma \mid AX_i,\Sigma_\gamma) \times IW(\Sigma_\gamma|\delta,\psi) \\
&\propto |\Sigma_\gamma^{-1}|^{-\frac{n}{2}} exp -\frac{1}{2} tr[\Sigma_\gamma^{-1}(\gamma - XA^T)^T(\gamma - XA^T)]|\Sigma_\gamma^{-1}|^{-\frac{\delta+3+1}{2}} exp\left\{-\frac{1}{2}tr(\Sigma_\gamma^{-1}\psi)\right\} \\
&\propto |\Sigma_\gamma|^{-\frac{n+\delta+3+1}{2}} exp\left\{-\frac{1}{2}tr\Sigma_\gamma^{-1}(\gamma - XA^T)^T(\gamma - XA^T)+\psi\right\} \\
&\propto \mathrm{IW}\left(\delta+n, \left(\psi^{-1}+((\gamma - XA^T)^T(\gamma - XA^T)\right)^{-1}\right)
\end{aligned}$$

Normality Assumption

check the normality assumption of the subject-specific parameter, we checked the Q-Q plot. The Q-Q plot indicates the normality assumption is satisfied.

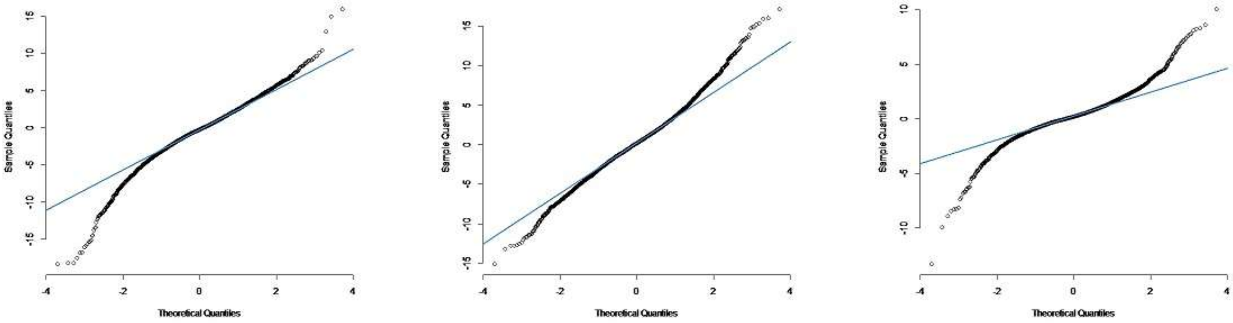


Figure 2.6: Normality assumptions plot for Basis spline for the subject specific parameters (a) γ_1 (b) γ_2 and (c) γ_3.

Autocorrelation

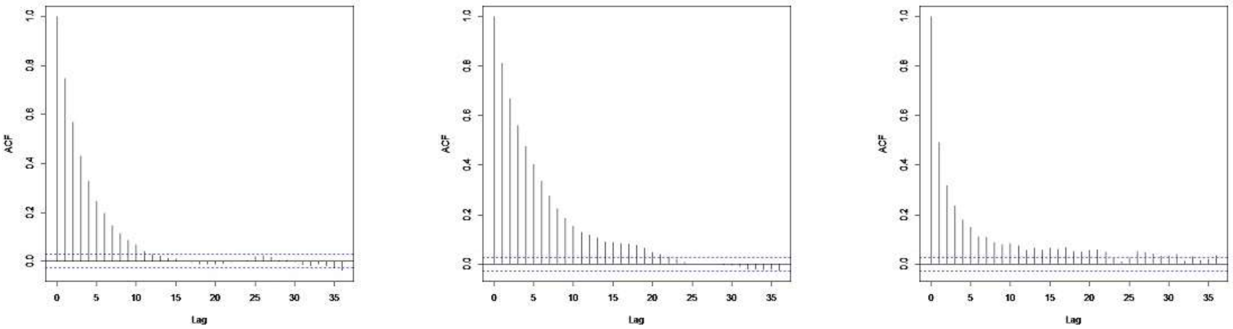


Figure 2.7: Auto correlation plot for Basis spline for the subject specific parameters (a) γ_1 (b) γ_2 and (c) γ_3.

Trace plot:

Trace plot indicates the convergence of the model parameters. We randomly selected a person and checked the trace plot, which indicates the model parameters converged.

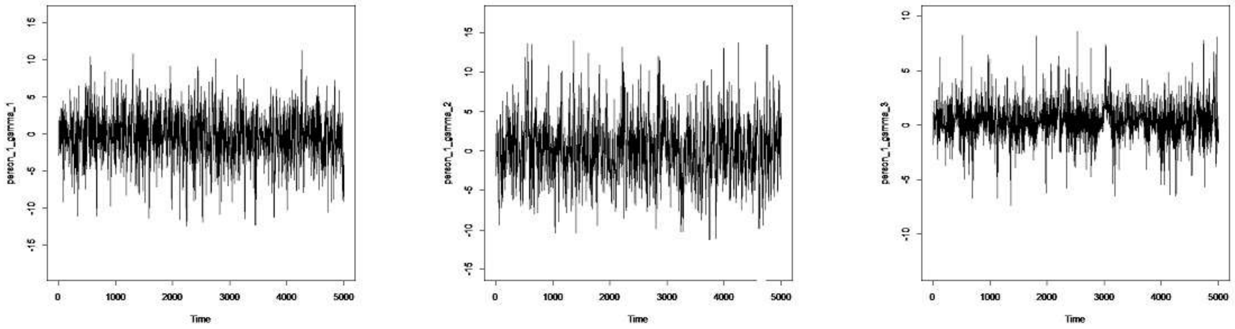


Figure 2.8: Auto correlation plot for Basis spline for the subject specific parameters (a) γ_1 (b) γ_2 and (c) γ_3.

Contour Plot

Mixing well of posterior samples is an important property to ensure that there is no specific trend in posterior samples among the parameters, and they are independent as much as possible. We randomly selected three patients and their posterior samples for a horizontal shift (γ_1), and stretch (γ_3) parameters plotted in Figure 4. The plot shows there is no specific trend in posterior samples for these two parameters, and they scattered around the center, indicating a low correlation.

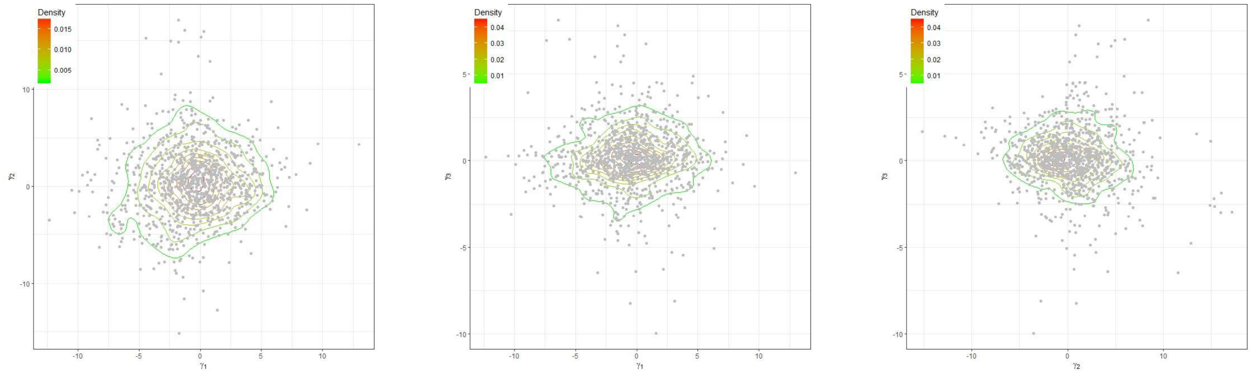


Figure 2.9: Contour plot of posterior samples for horizontal shift (γ_1) and stretch (γ_3) parameters for randomly selected three patients.

g

Chapter 3

Power Comparisons in 2x2 Contingency Tables: Odds Ratio versus Pearson Correlation versus Canonical Correlation.

Power Comparisons in 2x2 Contingency Tables: Odds Ratio versus Pearson Correlation versus Canonical Correlation

Mohammad Alfrad Nobel Bhuiyan[1,2], Michael J Wathen [1] ,MB Rao[1]

[1]Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio, USA

[2]Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

Corresponding Author:

Mohammad Alfrad Nobel Bhuiyan
University of Cincinnati
Kettering Lab Building, 160 Panzeca Way, Cincinnati, OH 45267-0056
E-mail: bhuiyama@mail.uc.edu

Abstract

It is an important inferential problem to test no association between two binary variables based on data. Tests based on the sample odds ratio are commonly used. We bring in a competing test based on the Pearson correlation coefficient. In particular, the Odds ratio does not extend to higher order contingency tables, whereas Pearson correlation does. It is important to understand how Pearson correlation stacks against the odds ratio in 2x2 tables. Another measure of association is the canonical correlation. In this paper, we examine how competitive Pearson correlation is vis-à-vis odds ratio in terms of power in the binary context, contrasting further with both the Wald Z and Rao Score tests. We generated an extensive collection of joint distributions of the binary variables and estimated the power of the tests under each joint alternative distribution based on random samples. The consensus is that none of the tests dominates the other.

keywords: Odds ratio, Pearson correlation, Canonical correlation, Contingency table, Power, Simulations

Introduction

Let X and Y be two binary random variables with joint distribution,

$$Q = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$$

Let the marginal probabilities be p_{1+}, p_{2+}, p_{+1}, p_{+2}. The odds ratio is defined by,

$$\theta = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

which is a measure of association between X and Y.

Facts :

- $0 \leq \theta \leq \infty$
- X and Y are independent if and only if $\theta = 1$
- Odds ratio measures to what extent the variables are away from independence.
- The ratio $\theta \geq 1$ means $Pr(X = 1, Y = 1) > Pr(X = 1)Pr(Y = 1)$. It is more likely to get $X = 1$ and $Y = 1$ than is possible under independence.

The joint distribution is unknown. Our test of Hypothesis is, Null Hypothesis (H_0): X and Y are independent.

vs

Alternative Hypothesis (H_1): X and Y are not independent.

Both null and alternative hypotheses are composite. Several tests can be built based on a random sample $\begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix}$ from the joint distribution.

Tests based on sample odds ratio

The likelihood estimator of θ is given by

$$\widehat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Let the marginal totals be n_{1+}, n_{2+}, n_{+1}, and n_{+2}. The asymptotic variance of $ln\left(\widehat{\theta}\right)$ is given (Courtesy: Delta method [57], [4], [3]) by

$$Asy\text{Var}\left(\ln \widehat{\theta}\right) = \frac{1}{np_{11}} + \frac{1}{np_{12}} + \frac{1}{np_{21}} + \frac{1}{np_{22}}$$

and it is estimated by

$$\widehat{Asy\text{Var}}\left(\ln\widehat{\theta}\right) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

The Wald's Z-statistic is given by

$$Z_1 = \frac{\ln\widehat{\theta} - \ln\theta}{\sqrt{\widehat{AsyVar}\left(\ln\widehat{\theta}\right)}} \tag{3.1}$$

which has a standard normal distribution, in large samples. In particular, $Z_1 = \frac{\ln\hat{\theta}}{\sqrt{Var\left(\ln\hat{\theta}\right)}}$ has the standard normal distribution N(0, 1) under the null hypothesis. An alternative to the Wald statistic is Rao's Score statistic. The variance of $\ln\widehat{\theta}$ is calculated under the null hypothesis and then estimated. The statistic is given by

$$Z_2 = \frac{\ln\hat{\theta}}{\sqrt{\widehat{AsyVar}_{H_o}\left(\ln\hat{\theta}\right)}} \tag{3.2}$$

The statistic Z_2 has a standard normal distribution under the null hypothesis for large samples. The formula for the asymptotic variance is given by:

$$Var_{H_o}\left(\ln\hat{\theta}\right) = \frac{1}{np_{1+}p_{+1}} + \frac{1}{np_{1+}p_{+2}} + \frac{1}{np_{2+}p_{+1}} + \frac{1}{np_{2+}p_{+2}}$$

and it is estimated by

$$\widehat{AsyVar}_{H_o}\left(\ln\hat{\theta}\right) = \frac{n}{n_{1+}n_{+1}} + \frac{n}{n_{1+}n_{+2}} + \frac{n}{n_{2+}n_{+1}} + \frac{n}{n_{2+}n_{+2}}.$$

Of course, we could have used the traditional chi-squared statistic for testing independence. However, unlike the odds ratio, there is no population chi-squared measure of association. We showed the relationship between chi-squared statistic and the likelihood estimate of Pearson correlation.

Tests based Pearson correlation

We are looking for a competitor to the odds ratio. One competitor is the Pearson Correlation [99]. The population correlation is given by

$$\phi = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{1+}p_{2+}p_{+1}p_{+2}}},$$

Where the entities under the square root are the marginal probabilities and it has the property $-1 \leq \phi \leq 1$. The random variables X and Y are independent if and only if $\phi = 0$. The likelihood estimate of ϕ is given by

$$\widehat{\phi} = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}}$$

A Z-statistic known as Wald can be built based on the likelihood estimator $\widehat{\phi}$ of ϕ. For the record, it is defined by

$$Z_3 = \frac{\widehat{\phi}}{\sqrt{\widehat{AsyVar}\left(\widehat{\phi}\right)}}. \tag{3.3}$$

The asymptotic variance of $\widehat{\phi}$ defined by the delta method is given in Appendix. For a description of the delta method, see [57]. Another alternative is the canonical correlation [123] defined by

$$\rho = \sqrt{p_{1+}p_{+1}p_{2+}p_{+2}}\left[\frac{p_{11}}{p_{1+}p_{+1}} - \frac{p_{12}}{p_{1+}p_{+2}} - \frac{p_{21}}{p_{2+}p_{+1}} + \frac{p_{22}}{p_{2+}p_{+2}}\right]$$

It can be shown that $\phi = \rho$. We use the notation ϕ and ρ interchangeably.Our motivation for including the canonical correlation into the mix goes a bit deeper. Canonical correlations arise from the singular value decomposition of a transform of the joint distribution. Several layers of dependence between X and Y can explained through (singular values) the canonical correlations. In the $2x2$, there is only one canonical correlation ρ and it is exactly the same as the Pearson ϕ.

As an alternative to Wald's Z statistic, we have Rao's score statistic based on $\widehat{\phi}$

$$Z_4 = \frac{\widehat{\phi}}{\sqrt{\widehat{AsyVar}_{H_o}\left(\widehat{\phi}\right)}}. \tag{3.4}$$

It turns out that, $AsyVar_{H_o}\left(\widehat{\phi}\right) = \frac{1}{n}$ [149]. It can be checked that $n\widehat{\phi}^2 = \chi^2$ ([150],[151]), the usual chi-squared statistic of the data in the $2x2$ contingency table [99]. We set the level of significance at 5%. Reject the null hypothesis at 5% level of significance if $|Z_i| > 1.96$.

The goals in this work are as defined below.

- Compare and contrast the properties of the measures of association: ϕ and θ (Sections 2 and 3).
- Make power comparisons between the Wald's test (Z_1) and Rao's score test (Z_2) based on the odds ratio and between Wald's test (Z_3) and Rao's score test (Z_4), which is the same as the chi-squared test, based on the Pearson correlation or canonical correlation (Section 4).

Description of Power comparisons using simulations.

1. Draw randomly 100 distributions from the space $\Omega = \{(p_{11,}p_{12,}p_{21,}p_{22})\,; p_{ij} \geq 0, sum = 1\}$. For sampling, we use the uniform Dirichlet distribution: Dirichlet $(p_{11}, p_{12}, p_{21}, p_{22}; 1, 1, 1, 1)$ whose joint density is given by $f\,(p_{11}, p_{12}, p_{21}, p_{22}) = 6$, $(p_{11}, p_{12}, p_{21}, p_{22}) \in \Omega$. Marginally, p_{ij}s are identically distributed. The marginal distribution of p_{11} is Beta$\,(1, 3)$ with $\mathrm{E}\,(p_{11}) = \frac{1}{4}$ and $\mathrm{Var}\,(p_{11}) = \frac{3}{80}$.

2. With probability one, under each joint distribution, X and Y are associated.

3. From each joint distribution $(p_{11}, p_{12}, p_{21}, p_{22})$ generated from the uniform Dirichlet distribution, generate a random sample $(n_{11}, n_{12}, n_{21}, n_{22})$ of 100 observations from the Multinomial$(n_{11}, n_{12}, n_{21}, n_{22}; \text{prob} = (p_{11}, p_{12}, p_{21}, p_{22}))$. The justification for the sample size to be 100 is that we can reasonably expect each $n_{ij} \geq 5$. All the tests methods are asymptotic in nature and the methods described by [50], [49] for the applicability of the asymptotic tests are being followed. For each Multinomial sample, we apply all the four tests defined by (1), (2), (3), and (4) at 5% level of significance. A counter was used for each test by: Counter = 1 if the null hypothesis is rejected, 0, if not rejected. Repeat the Multinomial sampling 1000 times. The estimated power under a test is the proportion of times the null hypothesis is rejected.

4. Present the results by tables and graphs.

In Section 2, we contrast Pearson ϕ and $ln(Oddsratio)$. In Section 3, we explain the background of canonical correlation. In Section 4, we present the results. In Section 5, we discuss the results. The asymptotic variance of $\widehat{\phi}$ is presented in the Appendix.

Canonical Correlation (Pearson correlation) versus Odds Ratio

A number of assumptions are as follows, .

- The event $\{X = 1, Y = 1\}$ is more likely than under the independence of X and Y if and only if $\theta > 1$ if and only if $ln\,(\theta) > 0$ if and only if $\rho > 0$.
- The event $\{X = 1, Y = 1\}$ is less likely than under the independence of X and Y if and only if $\theta < 1$ if and only if $\ln{(\theta)} < 0$ and if and only if $\rho < 0$.
- $-\infty \leq \ln{(\theta)} \leq \infty$

- $-1 \leq \rho \leq 1$

- The correlations are more attractive in that their ranges are bounded. However, the odds ratio has better interpretability than the correlation.

- If the joint distribution is $\begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$, $\ln(\theta) = \infty$ and $\rho = 1$.

- If the joint distribution is $\begin{pmatrix} 0 & 0.5 \\ 0.5 & 0 \end{pmatrix}$, $\ln(\theta) = -\infty$ and $\rho = -1$.

We introduce variables of the joint distribution: $A = \frac{p_{11}}{p_{1+}p_{+1}}$; $B = \frac{p_{12}}{p_{1+}p_{+2}}$; $C = \frac{p_{21}}{p_{2+}p_{+1}}$; $D = \frac{p_{22}}{p_{2+}p_{+2}}$. Another characterization in terms of the variables emerges as follows:

$$A > 1 \Leftrightarrow D > 1 \Leftrightarrow \theta > 1 \Leftrightarrow \rho > 0.$$

Let, $G_1 =$ geometric mean of A and $D = (AD)^{0.5}$,

$G_2 =$ geometric mean of B and $C = (B * C)^{0.5}$,

$A_1 =$ Arithmetic mean of A and D $= \frac{A+D}{2}$,

$A_2 =$ Arithmetic mean of B and C $= \frac{B+C}{2}$.

The measures θ and ρ are functions of these pillars through their arithmetic and geometric means.

Odds ratio $= \theta = \left(\frac{G_1}{G_2}\right)^2$ and $ln\theta = 2(lnG_1 - lnG_2)$

The canonical correlation ρ is connected to the pillars.

$$\begin{aligned}
\rho &= \sqrt{p_{1+}p_{+1}p_{2+}p_{+2}}\left[\frac{p_{11}}{\sqrt{p_{1+}p_{+1}}} - \frac{p_{12}}{\sqrt{p_{1+}p_{+2}}} - \frac{p_{21}}{\sqrt{p_{2+}p_{+1}}} + \frac{p_{22}}{\sqrt{p_{2+}p_{+2}}}\right] \\
&= 2\sqrt{p_{1+}p_{+1}p_{2+}p_{+2}}(A_1 - A_2) \\
&= \frac{(p_{11}p_{22} - p_{12}p_{21})}{\sqrt{p_{1+}p_{+1}p_{2+}p_{+2}}} \\
&= \sqrt{p_{1+}p_{+1}p_{2+}p_{+2}}(G_1^2 - G_2^2)
\end{aligned}$$

Genesis of Canonical Correlations and Pearson ϕ

Given any 2x2 matrix A there exist two orthogonal matrices L and M each of order 2x2 such that

$$\mathrm{LAM}^T = \begin{pmatrix} \rho_1 & 0 \\ 0 & \rho_2 \end{pmatrix}$$

where $\rho_1\,(\geq 0)$ and $\rho_2\,(\geq 0)$ are the singular values of the matrix A with a conventional ordering of $\rho_1 \geq \rho_2 \geq 0$ As a matter of fact, ${\rho_1}^2$ and ${\rho_2}^2$ are the eigenvalues of AA^T and the singular values are the non-negative square root of the eigenvalues. Let the bivariate binary distribution along with the marginals be given by

$$\mathrm{Q} = \begin{pmatrix} p_{11} & p_{12} & p_{1+} \\ p_{21} & p_{22} & p_{2+} \\ p_{+1} & p_{+2} & 1 \end{pmatrix}$$

Let

$$\mathrm{B} = \begin{pmatrix} \frac{p_{11}}{\sqrt{p_{1+}p_{+1}}} & \frac{p_{12}}{\sqrt{p_{1+}p_{+2}}} \\ \frac{p_{21}}{\sqrt{p_{2+}p_{+1}}} & \frac{p_{22}}{\sqrt{p_{2+}p_{+2}}} \end{pmatrix}$$

The singular values ρ_1 and ρ_2 of B are called canonical correlations of X and Y. It turns out that $\rho_1 = 1$ and $\rho_2 = \rho$ has the property that $1 \geq \rho \geq 0$. The canonical correlation ρ characterizes independence of X and Y. That is $\rho = 0$ if and only if X and Y are independent [123]. We do not follow this definition of canonical correlation. Technically, ρ is taken to be the non-negative square root of one of the eigenvalues of BB^T. Where, one of the eigenvalues is always equal to one. The other one is given by

$$\rho^2 = \frac{p_{11}^2}{\sqrt{p_{1+}p_{+1}}} + \frac{p_{12}^2}{\sqrt{p_{1+}p_{+2}}} + \frac{p_{21}^2}{\sqrt{p_{2+}p_{+1}}} + \frac{p_{22}^2}{\sqrt{p_{2+}p_{+2}}} - 1.$$

We want to show both positive and negative square roots of ρ^2. We have found that the following takes both positive and negative values in $[-1, 1]$ whose square is ρ^2:

$$\rho = \sqrt{p_{1+}p_{+1}p_{2+}p_{+2}} \left[\frac{p_{11}}{\sqrt{p_{1+}p_{+1}}} - \frac{p_{12}}{\sqrt{p_{1+}p_{+2}}} - \frac{p_{21}}{\sqrt{p_{2+}p_{+1}}} + \frac{p_{22}}{\sqrt{p_{2+}p_{+2}}} \right].$$

We keep the same notation ρ. One can check that $\rho = \phi$ [68].

Results

The power of the tests based on Z_1 and Z_2 is compareed graphically under the 100 randomly generated a bivariate distribution of X and Y (Figure 3.1). The numerical results are presented in Appendix.

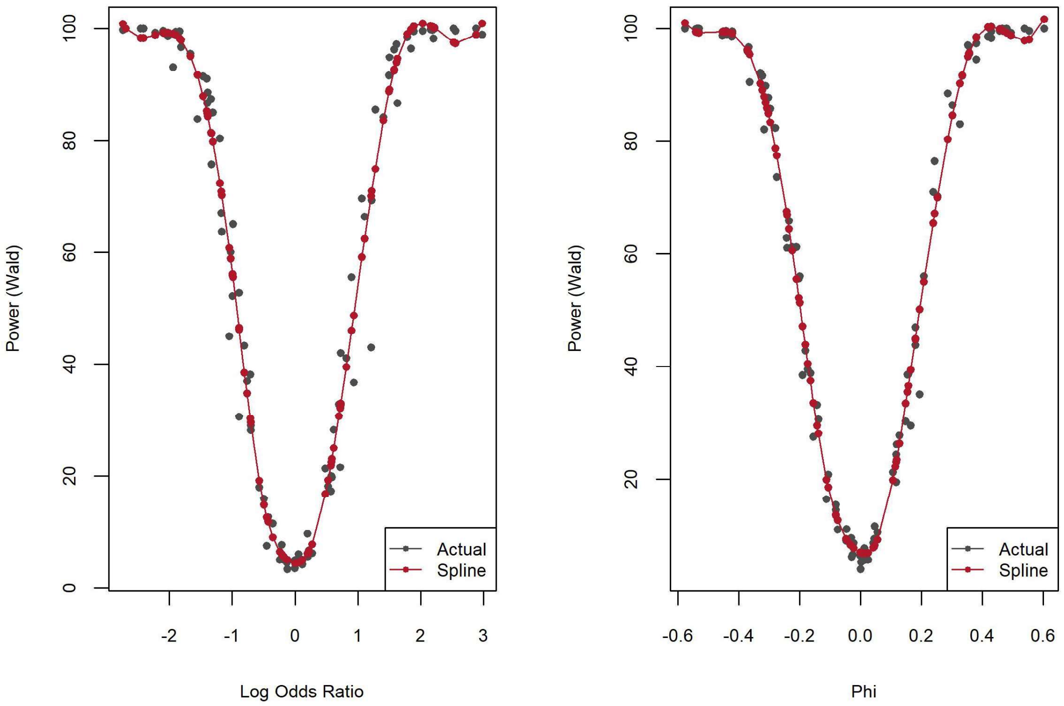


Figure 3.1: Wald tests: Odds ratio and canonical correlation

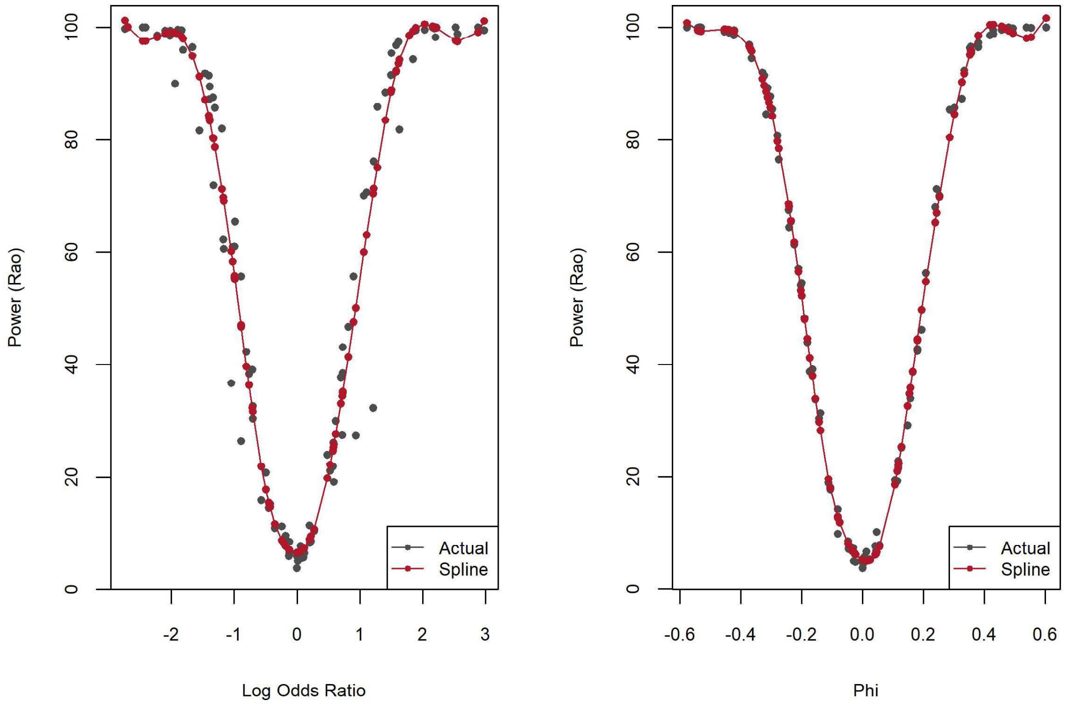


Figure 3.2: Rao tests: Odds ratio and canonical correlation

Comments on Figure 3.1 and Figure 3.2 : Structurally, the graphs are similar, even though the true values of ln(odds ratio) and Person ϕ are on a different scale. As $ln(\theta)$ and ρ moves away from the null value, the powers rise steeply towards 100 percent as expected. The spline model provides information of the underlying smoothness of power as a function of the measures of association. Similar comments do apply to the tests of Rao. A comprehensive comparison of the 4 tests is provided and in Figures 3.1 and 3.2

Comments on Figure 3.3: Each diagonal graph is a density histogram describing the distribution of power associated with one test. Structurally, the histograms are similar meaning that the distributions are similar. Every graph below the diagonals gives the scatter plot of a pair of powers coming from two different tests with a regression line drawn on the scatter

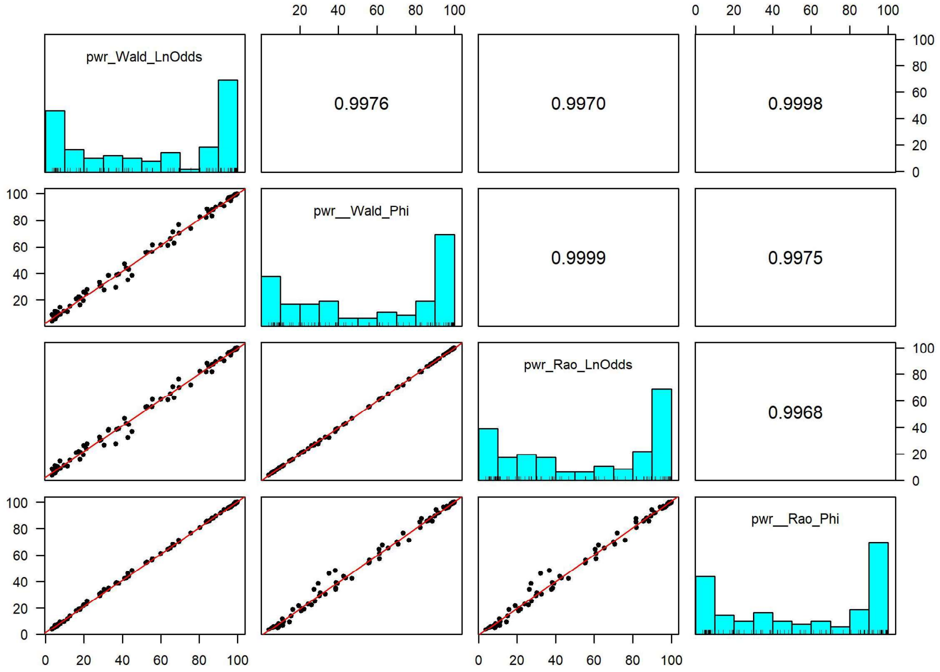


Figure 3.3: Correlation plots and histograms of powers

plot. Power pairs do lie more or less on the line. The graphs above the diagonal line give a Pearson correlation coefficient of the two power series. For Figure 3.3, we have generated 100 bivariate distributions of X and Y from the Uniform Dirichlet distribution on the simplex. For each distribution generated, $ln(Oddsratio)$ and Pearson ϕ was calculated. The scatter plot presented in Figure 3.3.

Comments on Figure 3.3: The Figure clearly indicates the similarity between the measures.

Discussion:

For testing independence of two binary variables, we examined the power of tests built upon ln(Odds ratio) and Pearson ϕ (Canonical correlation ρ) due to Wald and Rao. These tests use asymptotic variance formulas. Our comparisons are based on a random selection of

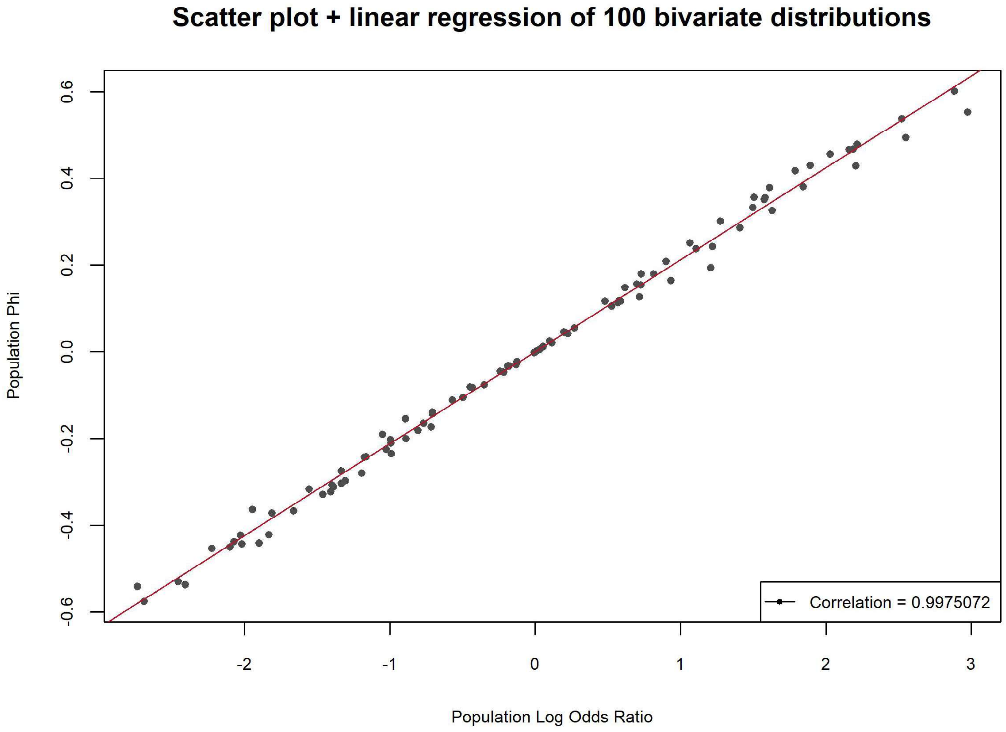


Figure 3.4: Correlation plot of $(ln(\theta), \rho)$ for 100 bivariate distributions of X and Y.

bivariate distributions from the uniform Dirichlet distribution on the simplex of bivariate distributions. We suggest that any of the four tests use in large samples.

A challenging task would be the determination of sample size for given level, power, and alternative values of the measure of association choice. There are pros and cons in using any measure of association for testing independence. The ln(Odds ratio) has an infinite range and confidence intervals based on Odds ratio could be very wide to interpret meaningfully. Pearson ϕ does not have this problem. The Odds ratio does not extend beyond the 2x2 case, where Pearson Φ is extendable to higher dimensional contingency tables. In case-control studies, the primary focus is testing equality of proportions of subjects achieving a cure. The odds ratio is used in this scenario, but Pearson ϕ or canonical correlation ρ are inappropriate to use in such a context.

We have shown that Rao scores statistic based on Pearson ϕ is related to the traditional χ^2 statistic of independence. Thus the χ^2 statistic is in the ambit of the main theme of the paper.

Appendix

Asymptotic variance of the likelihood estimator of Pearson ϕ

Asymptotic variance of the maximum likelihood estimator of Pearson correlation ϕ Steps:

1. Joint distribution of X and Y

$$Q = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

2. Pearson correlation

$$\begin{aligned} \rho &= \frac{ad - bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}} \\ &= \phi \\ &= UV^{-0.5}, \text{ where, U = ad - bc and V = (a+b)(a+c)(c+d)(b+d)} \end{aligned}$$

3. Generate data

$$D = \begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix}$$

4. Estimator of Q ,

$$\widehat{Q} = \begin{pmatrix} \frac{n_{11}}{n} & \frac{n_{12}}{n} \\ \frac{n_{21}}{n} & \frac{n_{22}}{n} \end{pmatrix}$$

For ease, in the description of the asymptotic formula, use a simple notation for the entries of $\widehat{Q}$

$$\widehat{Q} = \begin{pmatrix} j & k \\ l & m \end{pmatrix}$$

5. Estimate of ρ,

$$\widehat{\rho} = \frac{jm - lk}{\sqrt{(j+k)(j+l)(l+m)(k+m)}}$$

$$= f(j,k,l,m)$$

$$= x.y^{-0.5}, \text{ where, x = jm - lk and V = (j+k)(j+l)(l+m)(k+m)}$$

6. Asymptotic variance of $\widehat{\rho}$ using the delta method evaluated at their expectations, $j = E(j), k = E(k), l = E(l), m = E(m)$

$$AsymptoticVariance = \left(\frac{df}{dj}\right)^2 * var(j) + \left(\frac{df}{dk}\right)^2 * var(k) + \left(\frac{df}{dl}\right)^2 * var(l) +$$
$$\left(\frac{df}{dm}\right)^2 * var(m) + 2\left(\frac{df}{dj}\right) * \left(\frac{df}{dk}\right) * cov(j,k) +$$
$$2\left(\frac{df}{dj}\right) * \left(\frac{df}{dl}\right) * cov(j,l) + 2\left(\frac{df}{dj}\right) * \left(\frac{df}{dm}\right) * cov(j,m) +$$
$$2\left(\frac{df}{dk}\right) * \left(\frac{df}{dl}\right) * cov(k,l) + 2\left(\frac{df}{dk}\right) * \left(\frac{df}{dm}\right) * cov(k,m) +$$
$$2\left(\frac{df}{dl}\right) * \left(\frac{df}{dm}\right) * cov(l,m)$$

7. Calculate the variances and covariances,

$\mathrm{var}\,(j) = \frac{a(1-a)}{n}; \mathrm{var}\,(k) = \frac{b(1-b)}{n}$

$\mathrm{var}\,(l) = \frac{c(1-c)}{n}; \mathrm{var}\,(m) = \frac{d(1-d)}{n} \mathrm{cov}\,(j,k) = -\frac{ab}{n}; \mathrm{cov}\,(j,l) = -\frac{ac}{n}$

$\mathrm{cov}\,(j,m) = -\frac{ad}{n}; \mathrm{cov}\,(k,l) = -\frac{bc}{n}$

$\mathrm{cov}\,(k,m) = -\frac{bd}{n}; \mathrm{cov}\,(l,m) = -\frac{cd}{n}$

8.

$$\begin{aligned}
\frac{df}{dj} &= x\left(\frac{dy^{-0.5}}{dj}\right) + y^{-0.5}\left(\frac{dx}{dj}\right) \\
&= x(-0.5)y^{-\frac{3}{2}}\frac{dy}{dj} + y^{-0.5}\left(\frac{dx}{dj}\right) \\
&= -(0.5)xy^{-0.5}y^{-1}(2j+k+l)(l+m)(k+m) + y^{-1}m
\end{aligned}$$

9.

$$\begin{aligned}
\left(\tfrac{\partial f}{\partial j}\right)_{j=\mathrm{E}(j),k=\mathrm{E}(k),l=\mathrm{E}(l),m=\mathrm{E}(m)} &= -\tfrac{1}{2}uv^{-1/2}v^{-1}\left(2a+b+c\right)\left(c+d\right)\left(b+d\right) + v^{-1/2}d \\
&= -\tfrac{1}{2}\rho v^{-1}\left(2a+b+c\right)\left(c+d\right)\left(b+d\right) + v^{-1/2}d
\end{aligned}$$

10.

$$\left(\frac{\partial f}{\partial k}\right)_{j=\mathrm{E}(j),k=\mathrm{E}(k),l=\mathrm{E}(l),m=\mathrm{E}(m)} = -\tfrac{1}{2}\rho v^{-1}\left(2b+a+d\right)\left(a+c\right)\left(c+d\right) - v^{-1/2}c$$

11.

$$\left(\frac{\partial f}{\partial l}\right)_{j=\mathrm{E}(j),k=\mathrm{E}(k),l=\mathrm{E}(l),m=\mathrm{E}(m)} = -\tfrac{1}{2}\rho v^{-1}\left(2c+a+d\right)\left(a+b\right)\left(b+d\right) - v^{-1/2}b$$

12.

$$\left(\frac{\partial f}{\partial m}\right)_{j=\mathrm{E}(j),k=\mathrm{E}(k),l=\mathrm{E}(l),m=\mathrm{E}(m)} = -\tfrac{1}{2}\rho v^{-1}\left(2b+b+c\right)\left(a+b\right)\left(a+c\right) + v^{-1/2}a$$

13. The expression derived in steps 1 through 12 are plugged into the asymptotic variance formula in Step 6.

14. if $\rho = 0$ then Asymptotic variance $(\widehat{\rho}) = \frac{1}{n}$

Chapter 4

Differential impact of acute PM2.5 exposure on risk of stroke by stroke subtype,age,sex and race: A case-crossover study

Differential impact of acute $PM_{2.5}$ exposure on risk of stroke by stroke subtype, age, sex and race: A case-crossover study

Mohammad Alfrad Nobel Bhuiyan[1,2], Cole Brokamp[2,3], Tracy E. Madsen[2,3], Jane Khoury[2,3], Heidi Sucharew[2,3], Kathleen Alwell[2,3], Charles Moomaw[2,3], Md Monir Hossain[2,3], Brett Kissela[2,3], Dawn Kleindorfer[2,3]

[1]Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio, USA

[2]Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[3]Department of Pediatrics, University of Cincinnati, Cincinnati, Ohio, 45267, USA

Corresponding Author:

Mohammad Alfrad Nobel Bhuiyan
University of Cincinnati
Kettering Lab Building, 160 Panzeca Way, Cincinnati, OH 45267-0056
E-mail: bhuiyama@mail.uc.edu

Abstract:

Objective: To assess the relationship between acute ambient fine particulate matter ($PM_{2.5}$) and stroke onset and to determine whether this relationship is modified by stroke subtype, age, sex, and race.

Method: We used a case-crossover design, to examine the association of exposure to $PM_{2.5}$ and onset of incident stroke for the calendar year 2010. Data collected for the Greater Cincinnati Northern Kentucky Stroke Study (GCNKSS), for patients 20 years and older, initially ascertained using in-hospital ICD-9 discharge codes, were utilized.

Results: Of the 3267 incident strokes 2872 (88%) were infarct (INF) and Transient Ischemic Attack (TIA) and 395 (12%) hemorrhagic stroke; intracerebral hemorrhage (ICH) or subarachnoid hemorrhage (SAH). of these patients, 1855 (56%) were female, and 750 (23%) were Black. The overall mean daily $PM_{2.5}$ concentration for the year 2010 was 14.3 and standard deviation (SD) 6.0, the mean daily air temperature was 12.5C and SD 11.2 and relative humidity 76.0 $kg/m^2/s$ with a SD of 11.2. We found a significant association between $PM_{2.5}$ and stroke of any subtype three days prior to stroke onset, odds ratio (OR) 1.12 (95% CI: [1.03, 1.21]). For INF or TIA, $PM_{2.5}$ and stroke onset were associated at two days prior; OR 1.10 (95% CI: [1.01, 1.19]) and three days prior; OR 1.12 (95% CI: [1.03, 1.22]). Similarly, for patients $\geq$ 65 years, male patients and white patients $PM_{2.5}$ and infarct or TIA onset were associated at three days prior; OR 1.15 (95% CI: [1.04, 1.27]) OR 1.18 (95% CI: [1.05, 1.34]) and OR 1.12 (95% CI: [1.02, 1.23]), respectively. These associations were not seen for patients with hemorrhagic stroke.

Conclusion: Higher $PM_{2.5}$ exposure, particularly at three days before the event, was associated with stroke onset with differences evident by stroke subtype, age, sex, and race.

Introduction:

Globally, stroke is the second leading cause of death and long-term disability [67] and remains the fifth leading cause of death and fourth in long-term disability in the US [180]. Air pollution has a severe, negative impact on human health [116]. With rapid industrialization and economic growth, air pollution has become a serious concern all over the world [37]. According to the World Health Organization (WHO) report, 92% of the people in the world breathe unhealthy air and 6.5 million people die annually from exposure and identified air pollution as the most significant cause of premature death[153]. Additionally, arguments have been expressed leading to the expectation that exposure to elemental climate change impacts the risk for onset of stroke; in particular, acting as a trigger for the event [190],[156], [124]. Published studies have examined the association between the risk of cardiovascular events including stroke and exposure to air pollution [67], [61],[39],[37],[158],[187],[20],[172],[181],[195],[192],[130]. Some of these studies evaluated the effect of $PM_{2.5}$ on stroke risk; however overall results remain conflicting [190], [124], [138],[38]. Both changes from warm to cold and from cold to warm temperatures, and from higher to lower and from lower to higher humidity, have been associated with increased risk of stroke [136],[148], suggesting change rather than level of actual exposure act as the trigger. Also, increased exposure to particulate matter has been reported as both associated and not associated with an increased risk of stroke [172],[181],[195], [192],[130],[202]. Scientific statements from the American Heart Association (AHA) were released in 2004 and updated in 2010 showing evidence of causal relationships between $PM_{2.5}$ exposure and cardiovascular disease leading to morbidity and mortality [39], [37]. Moreover, several studies have shown that the impact of $PM_{2.5}$ on stroke is cumulative and that direct effects on stroke or other morbidity and mortality can be triggered occur through increases in heart rate and blood pressure [187], [20]. The majority of studies, however, have limitations such as “small” sample size and sub-optimal monitoring techniques associated with the specific particulate exposure under investigation. So, to clarify and fill the knowledge gap, we sought to examine the association between exposure to $PM_{2.5}$, using state of the art exposure estimation, and stroke occurrence in a time sensitive manner employing a novel statistical approach, the time-stratified case-crossover design. We achieved this by the integration of remote sensing satellite systems data with air monitoring network data to estimate the exposure. To address this aim we used data, for patients 20 years and older, from the 2010 Greater Cincinnati Northern Kentucky Stroke Study (GCNKSS). We included the first stroke in 2010 for each patient, ascertained by in-hospital methods.

Study population and design:

Stroke cases:

The GCNKSS is a population-based epidemiology study evaluating incidence and prevalence of stroke every 5-years; July 1993 to June 1994 and calendar years 1999, 2005 and 2010, data validation for 2015 is currently ongoing. The catchment area involves residents within the 5-counties encompassing Cincinnati, two within Ohio and three within Kentucky. This area includes about 1.3 million people, who are representative of the US population with respect to race (black/white) distribution and socioeconomic status. Ascertainment methods have remained the same over the study periods and have been described in detail elsewhere [34],[116]. Briefly, we captured events presenting at one of the area hospitals and discharged with an ICD9 stroke code. Also, other strokes ascertained in the outpatient settings; coroner's offices, public health, and hospital-based clinics and family practice centers, plus a weighted sample of events identified in nursing homes and doctor's offices, were collected. However, for the current study, we included only the first stroke within the calendar year 2010 for those who are greater or equal to 20 years of age at the onset of stroke and were ascertained in-hospital. The GCNKSS study was approved by the institutional review board at all participating hospitals. Study research nurses reviewed all medical records with discharge ICD9 codes (430 to 436) and decided as to whether a stroke event had occurred.

For all potential events the nurses abstracted demographics (including address at the time of stroke onset), presenting symptoms, co-morbidities, laboratory results, and diagnostic and neuroimaging data. All of the abstracted records were reviewed by study physicians and those determined to be cases were defined as one of five stroke categories adapted from the Classification for Cerebrovascular Diseases III and epidemiological studies of stroke as: cerebral ischemia (INF), intracerebral hemorrhage (ICH), subarachnoid hemorrhage (SAH), stroke of uncertain cause (UNK), or transient ischemic attack (TIA, symptoms lasting <24 hours). Residential locations of study participants at the time of each event were geocoded using our standalone and validated geocoder [36]. Only cases within the GCNKSS area that geocoded precisely enough to assign $PM_{2.5}$ concentrations were retained for analysis.

Table 4.1: Characteristics of the stroke patients (GCNKSS, 2010)

Characteristic	$Allstroke(n\%)$	$Hemorrhage(n\%)$	$INF\&TIA(n\%)$
N	3267	395 (12)	2872 (88)
Age ($\geq$ 65 years)	2045 (63)	211 (53)	1834 (63)
Age ($<$ 65 years)	1222 (37)	184 (47)	1038 (37)
Female	1855 (56)	223 (56)	1632 (56)
Male	1412 (44)	172 (44)	1240 (44)
White	2517 (77)	295 (75)	2222 (77)
Black	750 (23)	100 (25)	650 (23)

$PM_{2.5}$ Exposure:

Daily ambient concentrations of $PM_{2.5}$ were estimated using residential addresses included with the GCNKSS study extracted from the EHR using a previously developed and validated spatiotemporal model [35]. Briefly, our $PM_{2.5}$ model is based on satellite-derived measurements of aerosol optical depth (AOD), a measure of the scattering of electromagnetic radiation due to aerosols in the atmosphere. These measurements calibrated using ground-based

$PM_{2.5}$ monitoring and meteorological and land use data. Spatiotemporal datasets harmonized to a 1 x 1 km grid, and random forests were used to train a model to predict $PM_{2.5}$ concentrations. Our model demonstrated a leave-one-out cross-validated R^2 of 0.91. For analysis purposes, we considered the lagged effects of $PM_{2.5}$ by estimating concentrations at their residence for the days 0 (case day), and days 1, 2, 3, 4 and 5 before the case date (stroke onset date) and similarly for the days 0, 1, 2, 3, 4 and 5 the control dates (themselves beginning at 7 (day 0) and 14 (day 0) days before the date of stroke onset).

Statistical analysis:

We used a case-crossover design [136] to investigate the association between short term $PM_{2.5}$ exposure and incidence of different subtypes of stroke hemorrhagic (ICH, SAH), and non- hemorrhagic (INF, TIA). The advantage of a case-crossover study is that each case serves as its own control and, as such, is self-matched for fixed individual characteristics namely sex, age and socio-economic status, stroke risk factors such as smokng, underlying cardiovascular disease, obesity ,diabetes and controls for potential confounders that didn't vary during the month. For each person there is a 'case window' (the period of time during which a person was a case) and a 'control window' (The period of time when the person's time period was not associated with being a case). Using the case-crossover design the risk exposure during the case period is compared with risk exposure during the control period. Control periods are generally defined as fixed time intervals preceding and/or following the case period. In this analysis, the start of the case period was defined by the date of the stroke and the start of the control periods was defined as 7 days and 14 days before the case period [56]. The period of interest for exposure to $PM_{2.5}$ for both case and control dates was defined as the previous 5 days. The reason for just a 7 day "wash out" time between the case and control periods is to limit the potential for time-varying confounders including season, temperature and humidity [134],[180],[93],[79]. Daily average $PM_{2.5}$ concentration and weather conditions were linked by geocoding of residential location and day of stroke onset and all the other days of interest for each individual, as described above. Conditional logistic regression models were used to estimate the odds of a stroke at zero to five days given a 10 $\mu g/m^3$ increase in $PM_{2.5}$ exposure, adjusting for: if the day of stroke is a holiday or not, lag holiday (whether the day before the case day is a holiday or not) and daily average temperature and relative humidity. A natural cubic spline (NCS) function of calendar time was used to adjust for seasonality, daily average temperature and relative humidity on the same day and 5 previous days to allow for the potential nonlinear confounding effects of weather condition [4]. The knots for each spline were selected separately based on Akaike information Criterion (AIC) values. To control for the difference in the baseline hospital admission rates, day of the year, holiday and lag holiday were also incorporated in the model. A holiday is defined as a federal holiday in USA. Lag day, is defined as the time between the exposure and outcome [44]. The results are presented as adjusted odds ratio and 95% confidence interval (CI) for the stroke onset day per 10 $\mu g/m^3$ increase in $PM_{2.5}$ concentration. Temporal associations of $PM_{2.5}$ were examined with different lag structures from the case onset date (lag0) up to 5 lag days (lag5). Considering purely the single day lag estimate may underestimate the effect of $PM_{2.5}$, so cumulative effects were also evaluated using 2-days combined (lag 0-1), 3-days combined (lag 0-2), 4-days combined (lag 0-3), 5-days combined (lag 0-4) and 6-days combined (lag 0-5) to create a moving average of $PM_{2.5}$ concentrations [45]. We also conducted a sensitivity analysis to examine the robustness of the results in terms of the lag exposure; the degrees of freedom in the smoothing function of time trend (described above as NCS), daily mean temperature and daily relative humidity. Effect modification occurs when exposure has a different effect on subgroups of the population. Effect modification was examined for stroke subtype (hemorrhagic (ICH or SAH) versus non-hemorrhagic (INF or TIA)), age (<65, $\geq$ 65 years)[116], sex (Male, Female) and race (White, Black) in order to define any high-risk subgroup. To assess the effect modification, a Chi-squared test was used to test if there was a significant reduction in the log-likelihood after the addition of an interaction term between $PM_{2.5}$ and stroke subtype, age, sex and race separately, interactions with a resulting p-value of less than 0.1 were considered to be statistically significant. Including interaction terms in the model increases the precision of estimates by taking subgroup heterogeneity into account. Data management was conducted using SAS, version 9.4 (SAS, Cary, NC). All statistical and geospatial computing was done in R, version 3.4.3, using the survival and spline packages [184].

Results:

Of 3267 first stroke events per patient between January 1st 2010 to December 31st 2010, 2872 (88%) had INF and TIA, and 395 (12%) had hemorrhage (ICH or SAH). Of these patients, 1855 (56%) were female, and 750 (23%) were Black. The overall mean daily $PM_{2.5}$ concentration was 14.3 mg/m3 and standard deviation (SD) 6.0, with an interquartile range (IQR) from 4.2 to 34.8. The mean air temperature and relative humidities were 12.5(C) and 76.0 ($kg/m^2/s$), respectively. Characteristics of the stroke patients (GCNKSS, 2010) are presented in Table 4.2.

Association of $PM_{2.5}$ and stroke:

We found a significant association between $PM_{2.5}$ and stroke of any type at lag day 3 with adjusted odds ratio (OR) 1.12 (95% CI: [1.03, 1.21]) and an adjusted OR of 1.10 (95% CI: [1.01, 1.19]) and 1.12 (95% CI: [1.03, 1.22]) for infarct or TIA at lag days 2 and 3, respectively. On examination of cumulative lag days we found a significant association between $PM_{2.5}$ and stroke of any type for cumulative 0-2 days with adjusted OR 1.15 (95% CI: [1.04, 1.28]), increasing to adjusted OR 1.17 (95% CI: [1.03, 1.32]) for cumulative lag days 0-5. We found the same association for the same lag days for infarct or TIA with adjusted OR 1.18 (95% CI: [1.06, 1.31]), increasing to adjusted OR 1.23 (95%CI: [1.08, 1.40]) for cumulative lag days 0-5. There were no significant associations detected between $PM_{2.5}$ exposure and onset of hemorrhagic stroke. Adjusted odds ratios and 95% CIs for the odds of stroke of any type, and by stroke subtype, and $PM_{2.5}$ exposure for day of stroke and lag days 1-5 and associated cumulative lag days are shown in Table 4.2.

Table 4.2: Model estimates for the odds of stroke given PM2.5 exposure at 0 (case day) to lag day 5 and cumulative lag days 0 (case day) through lag day 5, overall and by stroke subtype

$Strokesubtype$	$LagDay$	$Oddsratio(95\%CI)$	$CumulativeLagday$	$Oddsratio(95\%CI)$
All	0(Case day)	1.05 (0.97, 1.14)		
	1	1.05 (0.96, 1.13)	0-1	1.09 (0.99, 1.19)
	2	1.08 (1.00, 1.17)	0-2	1.15 (1.04, 1.28)
	3	1.12 (1.03, 1.21)	0-3	1.20(1.08, 1.34)
	4	1.04 (0.95, 1.12)	0-4	1.19(1.06, 1.33)
	5	1.04 (0.96, 1.12)	0-5	1.17(1.03, 1.32)
Infarct / TIA	0(Case day)	1.06 (0.97, 1.15)		
	1	1.06 (0.97, 1.15)	0-1	1.10(1.00, 1.21)
	2	1.10 (1.01, 1.19)	0-2	1.18(1.06, 1.31)
	3	1.12 (1.03, 1.22)	0-3	1.23(1.09, 1.38)
	4	1.06 (0.97, 1.16)	0-4	1.23(1.08, 1.39)
	5	1.07 (0.98, 1.16)	0-5	1.23(1.08, 1.4)
Hemorrhagic	0 (Case day)	1.03(0.84, 1.26)		
	1	0.99 (0.80, 1.20)	0-1	1.03(0.83, 1.28)
	2	1.02 (0.83, 1.25)	0-2	1.08(0.85, 1.37)
	3	1.06 (0.86, 1.31)	0-3	1.12(0.86, 1.45)
	4	0.88 (0.71, 1.08)	0-4	1.04(0.78, 1.39)
	5	0.87 (0.71, 1.08)	0-5	0.96(0.70, 1.31)

$PM_{2.5}$ and Infarct or TIA; association by age:

We found a significant association between $PM_{2.5}$ and Infarct or TIA stroke for patients of the age greater than 65 years, for lag day 2 with an adjusted OR of 1.16 (95% CI: [1.05, 1.28]) and lag day 3 with an adjusted OR of 1.15 (95% CI: [1.04, 1.27]). Cumulative lag days 0-1 for age greater than 65 had an adjusted OR of 1.12 (95% CI: [1.01, 1.26]), increasing to 1.24 (95% CI: [1.06, 1.45]) for cumulative lag days 0-5. Adjusted odds ratios and 95% CIs for the odds of stroke for age groups and $PM_{2.5}$ exposure with a day of stroke and 1-5 lag days and associated cumulative lag days are shown in Table 4.3 and Figure 4.1 and 4.2.

Table 4.3: Model results for the likelihood of INF or TIA associated with $PM_2.5$ exposure for lag day 0 (case day) to lag day 5 and cumulative exposure for lag day 0 (case day) through lag day 5, by Age group

Lag	*Subgroup*	*Oddsratio*	*Cumulative*	*Oddsratio*
Day		(95%*CI*)	*Lagday*	(95%*CI*)
Age				
0 (Case day)	Age $<$ 65	1.05 (0.92, 1.19)		
	Age $\geq$ 65	1.06 (0.96, 1.18)		
1	Age $<$ 65	1.01 (0.88, 1.15)	0-1	1.07 (0.93 ,1.24)
	Age $\geq$ 65	1.09 (0.98, 1.2)		1.12 (1.01,1.26)
2	Age $<$ 65	1.00 (0.88, 1.14)	0-2	1.10 (0.94, 1.29)
	Age $\geq$ 65	1.16 (1.05, 1.28)		1.22 (1.08, 1.39)
3	Age $<$ 65	1.08 (0.95, 1.24)	0-3	1.14 (0.96, 1.36)
	Age $\geq$ 65	1.15 (1.04, 1.27)		1.28 (1.12, 1.47)
4	Age $<$ 65	1.12 (0.98, 1.28)	0-4	1.17 (0.97, 1.41)
	Age $\geq$ 65	1.03 (0.94, 1.14)		1.26 (1.09, 1.46)
5	Age $<$ 65	1.15 (1.01, 1.3)	0-5	1.20 (0.98, 1.46)
	Age $\geq$ 65	1.03 (0.93, 1.14)		1.24 (1.06, 1.45)

$PM_{2.5}$ and Infarct or TIA; association by sex:

We found that at lag days 3 to 5, male patients are more susceptible to Infarct or TIA stroke with an adjusted OR of 1.18 (95% CI: [(1.05, 1.34]), 1.15 (95% CI: [(1.01, 1.30]), and 1.13 (95% CI: [(1.01, 1.27]), respectively. No single lag day was significant for females. For cumulative lag days, we found a significant association between $PM_{2.5}$ and stroke for males for cumulative lag days 0-2 with adjusted OR 1.17 (95% CI: [1.01, 1.36]), increasing to 1.30 (95% CI: [1.09, 1.56]) for cumulative lag days 0-5. For females significance was found for cumulative lag days 0-2 with adjusted OR 1.18 (95% CI: [1.04, 1.35]), and 1.17 (95% CI: [1.00, 1.38]) for cumulative lag days 0-5. Model results for the odds of stroke for $PM_{2.5}$ exposure with 0-5 day lag and cumulative lag days for infarct or TIA by sex are shown in Table 4.4 and Figures 4.1 and 4.2.

Table 4.4: Model results for the likelihood of INF or TIA associated with $PM_2.5$ exposure for lag day 0 (case day) to lag day 5 and cumulative exposure for lag day 0 (case day) through lag day 5, by Sex

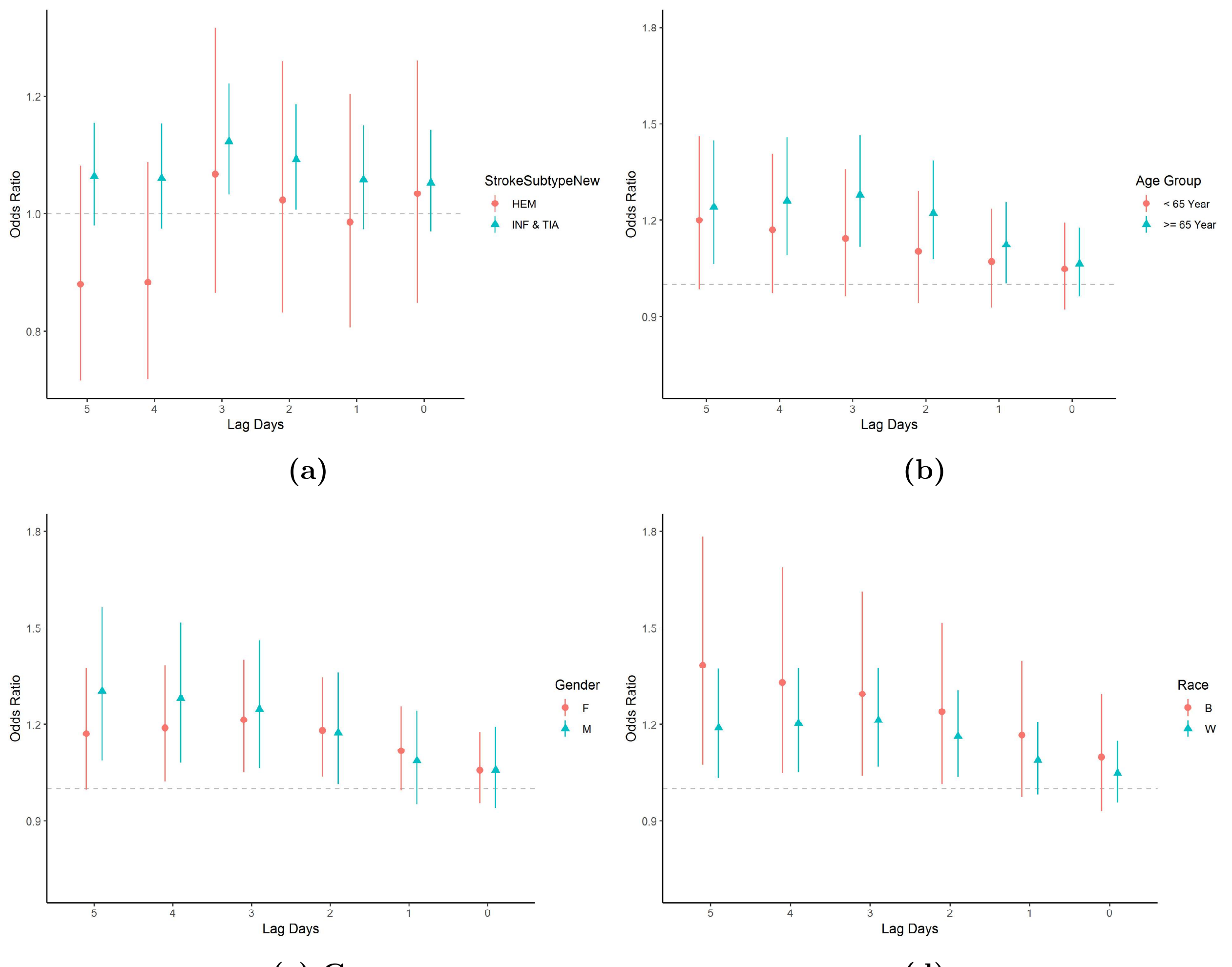
Lag Day	Sub $group$	$Oddsratio)$ $(95\%CI)$	$Cumulative$ $Lagday$	$Oddsratio$ $(95\%CI)$
Sex				
0 (Case day)	Female	1.06 (0.95, 1.18)		
	Male	1.06 (0.94, 1.19)		
1	Female	1.08 (0.97, 1.20)	0-1	1.12 (0.99, 1.26)
	Male	1.03 (0.91, 1.16)		1.09 (0.95, 1.24)
2	Female	1.08 (0.97, 1.20)	0-2	1.18 (1.04, 1.35)
	Male	1.12 (0.99 ,1.26)		1.17 (1.01 ,1.36)
3	Female	1.08 (0.97, 1.2)	0-3	1.21 (1.05, 1.4)
	Male	1.18 (1.05, 1.34)		1.25 (1.06, 1.46)
4	Female	1.01 (0.91, 1.12)	0-4	1.19 (1.02, 1.38)
	Male	1.15 (1.01, 1.3)		1.28 (1.08, 1.52)
5	Female	1.03 (0.93, 1.14)	0-5	1.17 (1.00, 1.38)
	Male	1.13 (1.01, 1.27)		1.30 (1.09, 1.56)

$PM_{2.5}$ and Infarct or TIA; association by race:

We found a significant association between $PM_{2.5}$ and Infarct or TIA stroke for patients of the white race, for lag day 3 with an adjusted OR of 1.12 (95% CI: [1.02, 1.23]). Cumulative lag days 0-2 for white patients had an adjusted OR of 1.16 (95% CI: [1.04, 1.31]), for 0-4 cumulative days, the adjusted OR was 1.20 (95% CI: [1.05, 1.38]). These results are shown in Table 4.5 and Figures 4.1 and 4.2.

Table 4.5: Model results for the likelihood of INF or TIA associated with $PM_2.5$ exposure for lag day 0 (case day) to lag day 5 and cumulative exposure for lag day 0 (case day) through lag day 5, by race

Lag	Sub	$Oddsratio)$	$Cumulative$	$Oddsratio$
Day	$group$	$(95\%CI)$	$Lagday$	$(95\%CI)$
Race				
0 (Case day)	Black	1.10 (0.93, 1.30)		
	White	1.05 (0.96, 1.15)		
1	Black	1.12 (0.95, 1.32)	0-1	1.17 (0.97, 1.40)
	White	1.04 (0.95, 1.14)		1.09 (0.98, 1.21)
2	Black	1.12 (0.94, 1.33)	0-2	1.24 (1.01, 1.52)
	White	1.09 (1.00, 1.20)		1.16 (1.04, 1.31)
3	Black	1.14 (0.96, 1.35)	0-3	1.29 (1.04, 1.61)
	White	1.12 (1.02, 1.23)		1.21 (1.07, 1.38)
4	Black	1.15 (0.96, 1.36)	0-4	1.33 (1.05, 1.69)
	White	1.04 (0.95, 1.15)		1.20 (1.05, 1.38)
5	Black	1.21 (1.02, 1.43)	0-5	1.38 (1.07, 1.78)
	White	1.04 (0.95, 1.14)		1.19 (1.03, 1.37)



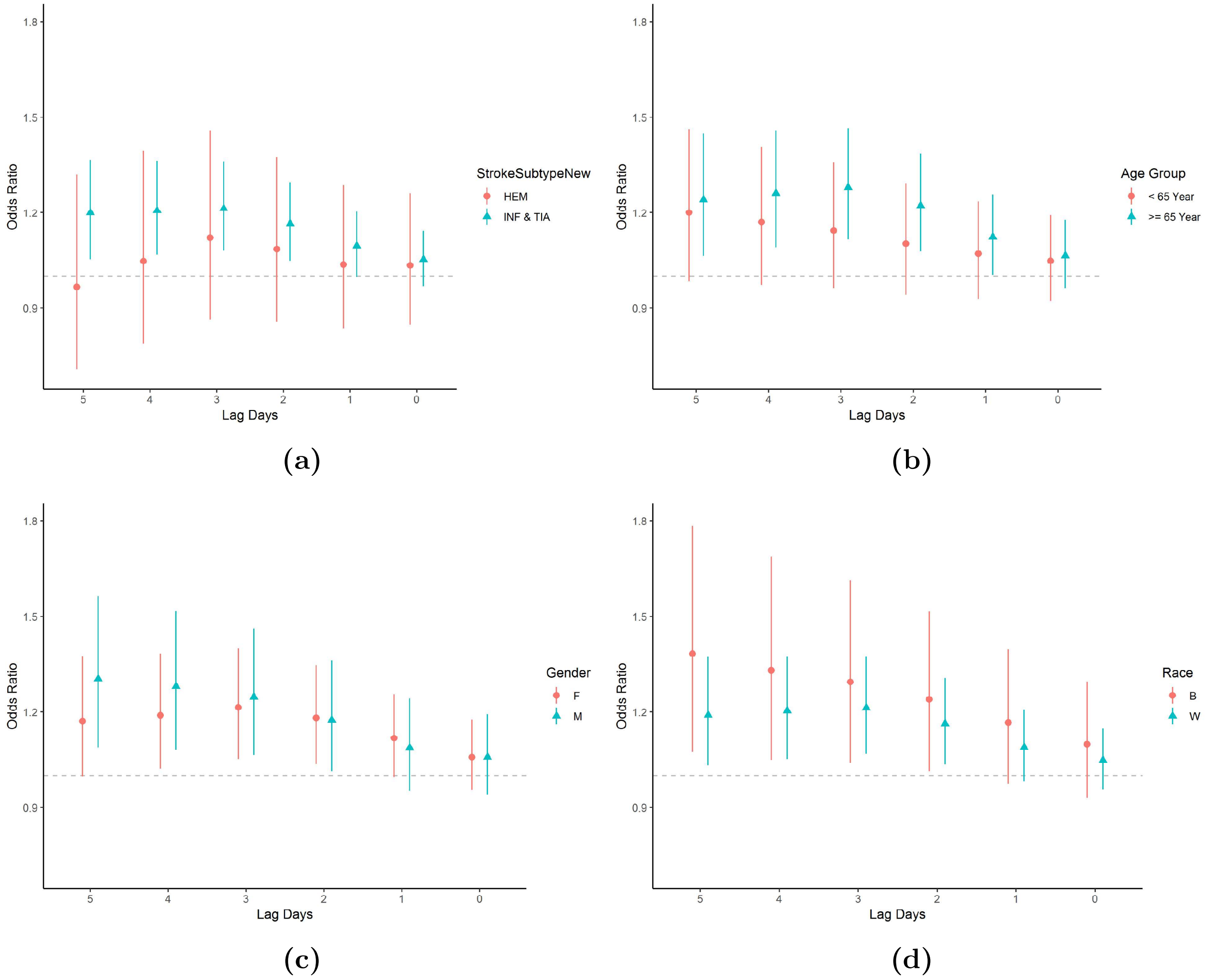
(a) (b)

(c) C (d)

Figure 4.1: Odds ratios and 95% CI for the likelihood of stroke given $PM_{2.5}$ exposure at 0 (case day) to 5 day lag by subgroup a) stroke subtype, and odds ratios and 95% CI for the likelihood of INF or TIA for subgroups b) age, c) sex and d) race.

Effect modification for stroke of any type and for Infarct or TIA by age, sex, and race:

The association of $PM_{2.5}$ and risk for stroke of any type was significantly modified by age greater than 65 (p = 0.03) and the association of $PM_{2.5}$ and risk for infarct or TIA was significantly modified by sex (p = 0.09), race (p = 0.09) and age group (p = 0.06).



(a)

(b)

(c)

(d)

Figure 4.2: Odds ratios and 95% CI for the likelihood of stroke given $PM_{2.5}$ exposure for cumulative lag days (where 5 represents an accumulation from lag day 5 to day of stroke) by subgroup a) stroke subtype, and odds ratios and 95% CI for the likelihood of INF or TIA for subgroups b) age, c) sex and d) race.

Discussion:

The Greater Cincinnati Northern Kentucky Stroke Study (GCNKSS) data for the year 2010 were utilized in this study to estimate the effect of $PM_{2.5}$ on the risk of stroke of any type and by stroke subtype defined as hemorrhagic or non-hemorrhagic (infarct or TIA). Additional stratification by age, sex, and race was also examined for patients with infarct or TIA. The potential effect modification, due to age, sex and race, was also examined by considering the inclusion of an interaction term in the models. Air pollutant and weather data from each study day were obtained from a previously developed and validated spatiotemporal model [36] were combined with the GCNKSS data [67]. This is a reproducible approach which may also be applied to other health surveillance systems to monitor different adverse health effects. We used the case-crossover design to analyze the association between $PM_{2.5}$ and incident of stroke [110],[134], [79]. In the case-crossover design, the study population consists of subjects who have experienced an episode of the health outcome of interest. Similar to a crossover study, each subject serves as his or her control. As in a matched case-control study, the inference is based on a comparison of exposure distribution rather than the risk of disease. The case-crossover study is most suitable for studying relations with the following characteristics: 1) the individual exposure varies within short time intervals; 2) the disease has an abrupt onset and short latency for detection and 3) the induction period is short. The case-crossover design allows the use of routinely monitored air pollution information and at the same time makes it possible to study individuals rather than days as the unit of observation. This design is also amenable for studying the effects of varying short-term air pollution exposure on health outcomes with an abrupt onset. In recent studies, the relationship between particulate matter and cardiovascular disease has been examined; Peng et al [156], Larrieu et al 3[124], Dominici et al [67], showed particulate matter air pollution is associated with hospital admission for cardiovascular disease, Want et al [192], Lisabeth et al [130], Yistiak et al [202] and Yaohua Tian et al [45] showed positive relationship between $PM_{2.5}$ and the first hospital admission for ischemic stroke. However, O'Donnell et all [148] didn't find any association between air pollution and the risk of acute ischemic stroke and McClure et al [138] didn't find any association between fine particulate matter ($PM_{2.5}$) and the risk of stroke in the REGARDS cohort. In our study, we found evidence of an association between $PM_{2.5}$ and stroke, in particular for non-hemorrhagic stroke of infarct or TIA. The short-time exposure to $PM_{2.5}$ was significantly associated with stroke, adjusted for temperature, relative humidity, the day of the week, long term trend, and seasonality of stroke events. The EPA is required to set a particulate matter National ambient air quality standard that is safe for the public health, and our findings indicate that $PM_{2.5}$ is an ongoing threat to public health and need further research to explore the source of pollutants and monitor the air quality for a healthy environment.

The source of particles for this air pollution needs to be identified for future strategic plans to control the health-related adverse effects. However, the source of air pollutants differs across location and time. Identifying the relationship between strokes about $PM_{2.5}$ concentration is of public health and regulatory interest. The EPA is required to set a particulate matter National ambient air quality standard that is safe for the public health. Our findings indicate that $PM_{2.5}$ is an ongoing threat to public health and need further research to explore the source of pollutants and monitor the air quality for a healthy environment.

Chapter 5

Source-specific contributions of particulate matter to asthma-related emergency department utilization

Source-specific contributions of particulate matter to asthma-related emergency department utilization

Mohammad Alfrad Nobel Bhuiyan[1,2],Patrick Ryan[2,3], Sivaraman Balachandran[4], Cole Brokamp[2,3]

[1]Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio, USA

[2]Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[3]Department of Pediatrics, University of Cincinnati, Cincinnati, Ohio, 45267, USA

[4] Department of Environmental Engineering, University of Cincinnati, Cincinnati, Ohio, USA

Corresponding Author:

Mohammad Alfrad Nobel Bhuiyan

University of Cincinnati

Kettering Lab Building, 160 Panzeca Way, Cincinnati, OH 45267-0056

E-mail: bhuiyama@mail.uc.edu

Abstract

Few studies have linked specific sources of ambient particulate matter smaller than 2.5 μm ($PM_{2.5}$) and asthma. In this study, we identified specific sources of $PM_{2.5}$ and examined their association with daily asthma hospital utilization in Cincinnati, Ohio, USA. We used Poisson regression models to estimate the daily number of asthma ED visits the day of and one, and two days following a 10 $\mu g/m^3$ increase in $PM_{2.5}$, adjusting for temporal trends, holidays, temperature, and humidity. In addition, we estimated the contributions of nine sources to daily concentrations of $PM_{2.5}$ using a chemical mass balance method and used a model-based clustering method to group days with similar source-specific contributions into six distinct clusters. Specifically, elevated $PM_{2.5}$ concentrations occurring on days characterized by low contributions of coal combustion showed a significantly reduced risk of hospital utilization for asthma (rate ratio: 0.86, 95% CI: [0.77, 0.95]) compared to other clusters. Reducing the contribution of coal combustion to $PM_{2.5}$ levels would be expected to reduce asthma-related hospital utilization.

Keywords: fine particulate matter, source apportionment, cluster, asthma, time-series

Introduction:

Air pollution is a global challenge (World Health Organization, 2018) and has a severe, negative impact on human health (Pascal et al., 2013, Abdalla et al., 2007, Pope III et al., 2006). Vehicles, households, and industries emit a complex mixture of air pollutants (Austin et al., 2012, Austin et al., 2013), including ambient particulate matter smaller than 2.5 μm ($PM_{2.5}$) (Anderson et al., 2012, Adams et al., 2015 Burnett et al., 1997, Deschamps et al., 2003). Despite reductions in $PM_{2.5}$ concentrations in the United States and other developed countries, industrialization and economic growth in Pakistan, Bangladesh, India, and other developing and highly populated countiries has led to an 18% increase in the global population-weighted $PM_{2.5}$ concentrations from 2010 to 2016 (HEI 2019). In total, an 92% of people worldwide breath unhealthy levels of air pollution, contributing to 6.5 million annual deaths (World Health Organization, 2018). In addition to chronic disease and mortality, epidemiologic evidence suggests short-term $PM_{2.5}$ exposure is associated with the development and exacerbations of asthma (Lam et al., 2016, Ding et al., 2017, Stevanović et al., 2006), triggering emergency department and hospital utilization in children (Iskander et al., 2012, Villeneuve et al., 2007). Asthma is a highly prevalent chronic respiratory disease (Sha et al., 2015) that contributes greatly to morbidity and hospital utilization worldwide. Children are particularly susceptible to $PM_{2.5}$ related health effects due to their immature immune system (Zhang et al. 2019) and ongoing development and growth. The composition of $PM_{2.5}$ varies according to its sources (Prieto-Parra et al., 2017), including fuel combustion from mobile sources like vehicles and stationary sources like power plants, industrial processes, and biomass burning (World Health Organization, 2017). Several studies have aimed to quantify the heterogeneous composition of $PM_{2.5}$, including identification of distinct multipollutant profiles (Austin et al., 2012), spatial clustering of air pollution monitoring sites (Austin et a., 2013), analyzing source specific contributions (Feng et al., 2018), and studying the effect of individual chemical constituents (Adams et al., 2015). However, much less attention has been focused on the source-specific contributions of particulate matter to health outcomes (Feng et al., 2018). Identifying sources contributing to $PM_{2.5}$ related health effects is critical to control the harmful sources of $PM_{2.5}$ (Heal et al., 2012) and to identify primary prevention strategies. In this study, we aimed to determine underlying $PM_{2.5}$ sources responsible for asthma-related pediatric hospital utilization. Daily estimates of the source-specific contributions of different $PM_{2.5}$ sources were estimated using a chemical mass balance source apportionment model, and a model-based clustering method (Fraley et al., 2002) was applied to group days having similar source profiles. Using daily counts of pediatric, asthma-related hospital utilization for one urban county in Cincinnati, Ohio, USA, we then examined whether the type $PM_{2.5}$, as determined by cluster membership, significantly modified the effect of $PM_{2.5}$ on hospital utilization.

Methods:

Source Apportionment:

A chemical mass balance (CMB) model (Pace et al., 1991) was used to estimate the contribution of the sources of fine particulate organic carbon at Ohio from 2011 to 2015. One in every three day $PM_{2.5}$ and elemental source measurements extracted from an AirData monitor (monitor ID: 39-061-0040) maintained by the Environmental Protection Agency (EPA). The measurements were taken as average from hourly measurements.

Health Outcome Data:

All emergency department (ED) and urgent care (UC) visits for asthma between 2011 and 2015 were identified within the Cincinnati Children's Hospital Medical Center's (CCHMC) electronic medical record (EHR) based on International Classification of Disease (ICD-9) codes 493.00–493.92) (World Health Organization, 2018). CCHMC is a pediatric academic health center that has a market share of 99% of all hospital admissions, and 81% of all hospital encounters among 0 to 14 year olds in Hamilton County (Beck et al., 2018). Hamilton County is located in Cincinnati, Ohio, USA and has 222 urban, suburban, and rural census tracts containing about 190,000 total children. The CCHMC Institutional Review Board approved this study and granted a waiver of informed consent.

Meteorological Data:

Average daily temperature and relative humidity were obtained from the North American Regional Reanalysis (NARR) dataset (Mesinger et al., 2006).

Statistical Analysis

We used Poisson regression models to investigate the association between daily counts of asthma-related hospital utilization and $PM_{2.5}$ concentrations on the same day as well as one and two days prior. Models were adjusted for day of the week, the day of the year, a federal holiday in the USA, temperature, and relative humidity. We used one dummy variable to indicate if one of the previous two days was a federal holiday or not and another dummy variable to indicate if the current day was a federal holiday. The continuous day of the year was used to adjust for seasonality and long-term trends by modeling it as a natural cubic spline with 8 degrees of freedom. We adjusted for daily average temperature and relative humidity using natural cubic splines with six and three degrees of freedom, respectively, to allow for the non-linear effect of weather conditions. The day of the week was also included in the model as a dummy variable. To emulate a "naive" approach, we also created similar models separately for each source component of $PM_{2.5}$. Lastly, we created models where the effect of average $PM_{2.5}$ for each day could be modified by cluster membership, described below. This was accomplished by adding an explicit interaction term between the $PM_{2.5}$ concentration and a dummy variable for cluster membership within each Poisson model to determine if significant effect modification existed. A Chi-squared test was used to determine if there was a significant reduction in the Akaike information criterion (AIC) after the addition of the interaction term. Model with a resulting p-value of less than 0.1 was considered to be significantly modified by the composition of $PM_{2.5}$. We utilized a model-based clustering method, specifically a Gaussian finite mixture model (Fraley et al., 1998), to group together days of the study period with similar source profiles. We chose model-based clustering over non-parametric clustering methods, like k-means or hierarchical clustering, to utilize a data-driven method for selecting the number of clusters. In the model-based clustering, we assumed sample observations arose from a finite normal mixture distribution, with a mixture probability or weight. Each component in the mixture model was called a cluster. The mixture model parameters were fitted using an expectation-maximization algorithm (Hastie et al., 2001). To select the number of clusters, we compared different models with different numbers of clusters and different parameterizations of the variance-covariance matrix and chose the model with the lowest Bayesian information criterion (BIC) (Fraley et al., 1998). A graph of the BIC for each combination of some clusters and variance-covariance matrix parameterizations included in the appendix (Figure A.16). To cluster based on the relative contributions of each source, rather than their absolute mass, we expressed each source as a modified Z-score (Austin et al., 2013). The modified Z-score allowed us to eliminate bias due to differences in scales and prevented outlier values from having too much influence on the cluster selection:

$$Z_{Modified} = \frac{(Sourcefraction - Median(Sourcefraction))}{Median(|Sourcefraction - Median(Sourcefraction)|)}$$

All statistical and geospatial computing was done in R (R Core Team, 2017), version 3.4.3, using the mclust package (Fraley et al., 2006).

Results:

Demographic Characteristics:

From 2011 – 2015, the daily total of asthma-related ED and/or UC visits ranged from 1 to 26 with a median of 8 total visits (25th percentile: 4, 75th percentile: 10). Figure 1 shows the dates of our study period arranged temporally and shaded by the magnitude of the daily visit totals. The number of total visits was fewer during the warm, summer season consisting of June, July, and August.

$PM_{2.5}$ Source Characteristics:

Throughout the study period, the median level of $PM_{2.5}$ was 9.9 μg/m3. Source apportionment analysis revealed nine distinct sources of $PM_{2.5}$ within our study region: gasoline

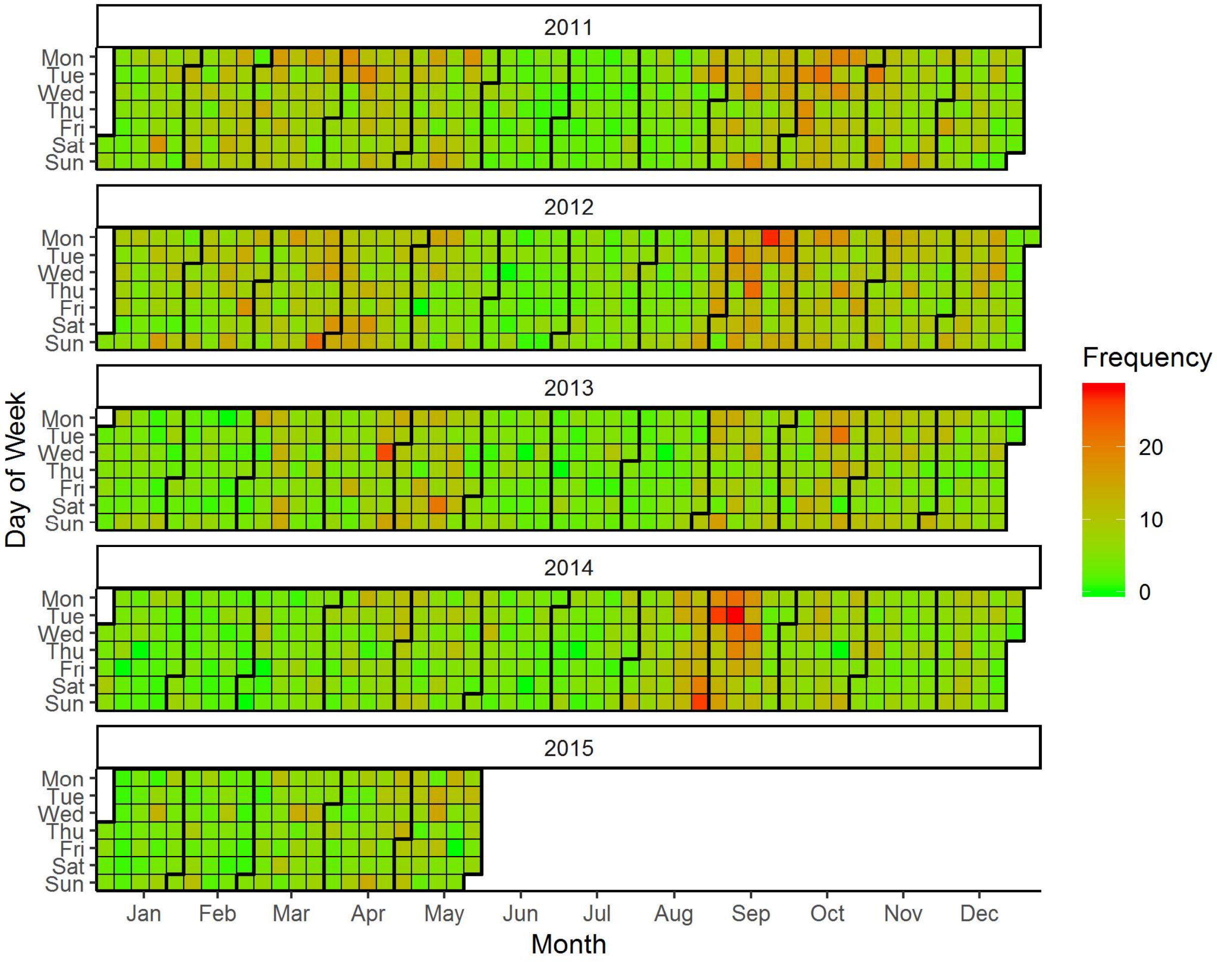


Figure 5.1: Calendar heat map of asthma-related hospital utilization from January 2011 to May 2015.

vehicles (GV), diesel vehicles (DV), dust (DUST), biomass burning (BURN), coal burning (COAL), organic carbon (SOC), sulfate (SO_4), ammonium (NH_4), and nitrate (NO_3). Observed $PM_{2.5}$ and each individual source were examined as time series plots from January 2011 to May 2015 (Figure A.1- A.10). From 2011 to 2015, the median concentration for GV was 0.21 $\mu g/$ m3, DV was 0.76 $\mu g/$ m3, BURN was 0.75 $\mu g/$m3, SOC was 0.76 $\mu g/$m3, NH_4 was 0.91$\mu g/$ m3, and NO_3 was 0.90 $\mu g/$ m3. SO_4 had the highest median concentration (2.08 $\mu g/$m3) whereas COAL and DUST concentrations had the lowest median concentrations (both 0.06 $\mu g/$m3). Summary statistics on the characteristics of $PM_{2.5}$, sources of air pollutants, and meteorological variables are provided in Table 1. The daily values of the concentration of $PM_{2.5}$ and the sources of $PM_{2.5}$ for the study period as a modified Z score are shown in Figure 2.

Table 5.1: Summary statistics of air pollution, meteorological, and ED utilization variables.

$Variable$	$Median$	Min	25^{th}	75^{th}	Max
				$Percentile$	$Percentile$
Daily Utilization (count)	7	1	4	10	26
$PM_2.5(\mu g/m^3)$	9.9	2	6.8	13.5	30.8
Temperature(K)	284.8	257.4	276.1	294.2	303.8
Humidity (kg/m2/s)	76.7	37.3	69.3	81.9	96.5
$PM_{2.5}Source$					
Concentration($\mu g/m^3$)					
DV	0.21	0.00	0.06	0.42	1.62
GV	0.76	0.02	0.58	0.93	2.24
BURN	0.75	0.00	0.54	1.09	6.36
COAL	0.06	0.00	0.02	0.10	0.50
DUST	0.06	0.00	0.00	0.16	1.31
NH4	0.91	0.03	0.57	1.41	4.99
NO3	0.90	0.06	0.49	2.00	12.50
SO4	2.08	0.00	1.31	3.11	12.97
SOC	0.74	0.00	0.44	1.15	3.46

Association between daily $PM_{2.5}$ concentrations and asthma-related hospital utilization:

We did not observe a significant association between total $PM_{2.5}$ and the rate of asthma-related hospital utilizations. The rate ratio (RR) and 95% confidence interval (CI) for hospital utilizations per 10 μg/m3 increase in $PM_{2.5}$ was 1.00 (95% CI: [0.93, 1.08]) on the same day, 0.97 (95% CI: [0.89, 1.05]) one day later, and 1.04 (95% CI: [0.96, 1.12]) two days

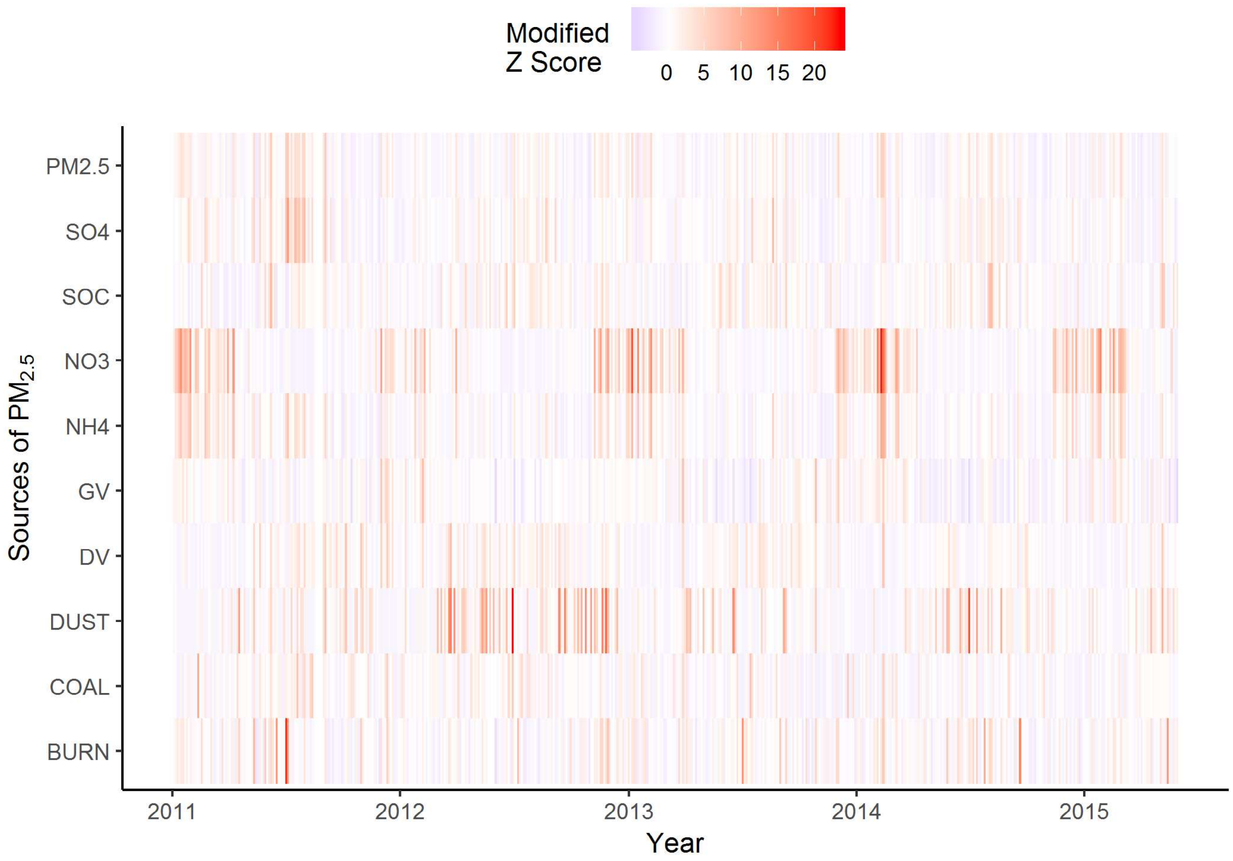


Figure 5.2: Heat map of the concentration of PM2.5 and different sources of $PM_{2.5}$ from January 2011 to May 2015 as modified Z scores.

later. Model results for $PM_{2.5}$ concentrations and asthma-related hospital utilization are depicted Figure A.12.

Association between source-specific $PM_{2.5}$ and asthma-related hospital utilization:

Model results for each individual $PM_{2.5}$ source are shown in Figure 3. When using individual models for each $PM_{2.5}$ source, $PM_{2.5}$ concentraitons due to DUST, DV, NH_4, BURN, and SOC were not associated with an increased risk for asthma-related hospital utilization. However, there was a significantly increased risk of asthma-related hospital utilization the same day of increased NO_3 (RR: 1.06, 95% CI: [1.01,1.13]), one day after increased GV (RR: 0.92, 95% CI: [0.86,0.99]), and both one and two days after increased COAL (RR: 1.01, 95% CI: [1.01, 1.06] and RR: 1.02, 95% CI: [1.00, 1.04], respectively).

Clustering sources of $PM_{2.5}$:

Our study period consisted of 522 total days, and we found that six clusters best fit the source apportionment fractional contributions as modified Z-scores. The heat map of the sources by cluster membership is shown in Figure 4. Cluster allocation of days and the mean and variance of the fractional contribution of each source for each cluster are presented in Table 2. Cluster 1 was allocated hundred and ninety-three days (36.97% of all days) and was characterized by high dust, SOC, and DV in conjunction with moderately high $PM_{2.5}$ and SO_4 and low NO_3. We nicknamed this cluster "high SO_4, SOC, and BURN with low COAL". This cluster mostly occurred in April (26 days), May (31 days), June (22 days), August (20 days), September (20 days) and October (21 days). Cluster 2 mostly occurred during the fall. Sixty days (11.49% of all days) allocated in this cluster and characterized by high dust, SOC, and DV in conjunction with moderately high nitrate, GV and ow DV. We nicknamed this cluster "high DUST and NO_3, with low SOC and DV". It occurred mostly

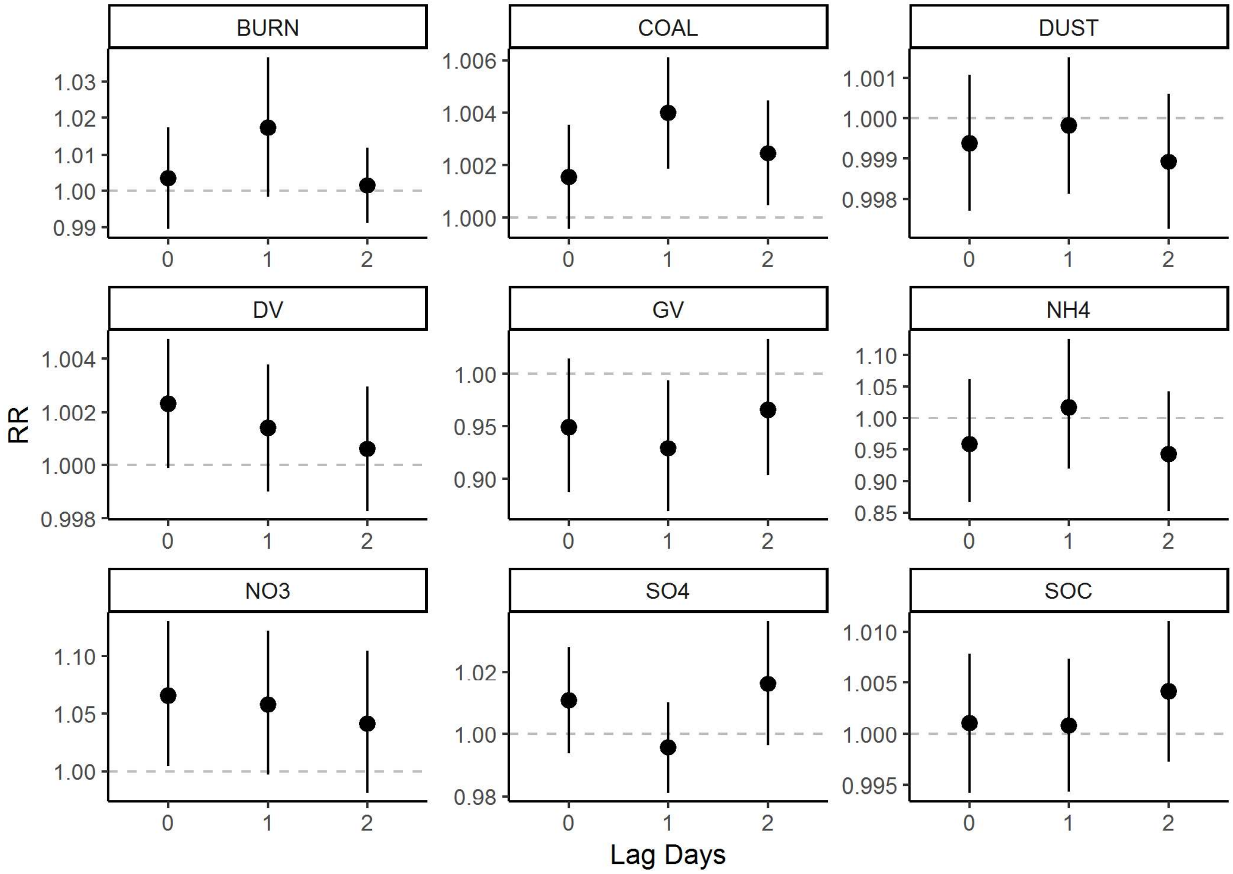


Figure 5.3: Risk ratio for the number of daily asthma-related hospital emergency department utilization for a 10 $\mu g/m^3$ increase in each source of $PM_{2.5}$ and for each lag day.

during January (11 days), February (12 days) and March (14 days). Cluster 3 characterized by high NO_3, NH_4, and GV. Hundred and five days (20.11% of all days) allocated in cluster 3.This cluster occurred more often in the earlier years. We nicknamed this cluster “high SO_4, NH_4, NO_3 with low DUST”. Cluster 3 mostly occurred in January (25 days) February (21 days), March (17 days) and December (16 days). Cluster 4 consisted of 81 days (15.51% of all days) with low coal, high NH_4, and NO_3. This cluster mostly occurred during the middle of the study period. We nicknamed this cluster “high NH_4, NO_3 with low COAL.” It occurred mostly during January (10 days), November (10 days) and December (11 days). Cluster 5 mostly occurred during spring and characterized by high DV, high SO_4, and low GV and fifty-nine days (11.30% of all days) allocated in cluster 5. We nicknamed this cluster “high SOC, SO_4 with low GV.” It occurred mostly during July (12 days), August (11 days) and September (11 days). Cluster 6 characterized by high burn, dust, low GV DV, and SOC. Fifteen days (2.87% of all days) allocated in cluster 6. We nicknamed this cluster “high BURN, SOC with low DUST, COAL, and DV.” It occurred mostly during May (2 days), June (2 days) and July (7 days). Calender heat map of the clusters by day shown in Figure 5.Cluster-specific risk ratios and 95% CIs for the number of daily asthma-related hospital utilizations for a 10 ug/m3 increase in PM2.5 and for each lag day. Bar plot of the clusters by week, month and year presented in Appendix A.11.

Table 5.2: The mean and standard deviation of the source fractions both overall and for each cluster.

character	*Total*	*Cluster*	*Cluster*	*Cluster*	*Cluster*	*Cluster*	*Cluster*
		1	2	3	4	5	6
Day(%)	522	193 (37)	60 (11)	105 (20)	81 (16)	59 (11)	15 (3)
$PM_{2.5}$							
Source							
BURN	0.11	0.11 (0.04)	0.12 (0.08)	0.09 (0.03)	0.10 (0.04)	0.10 (0.05)	0.28 (0.18)
COAL	0.01	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01(0.01)
DUST	0.02	0.03 (0.02)	0.02 (0.03)	0.01 (0.01)	0.01 (0.02)	0.01 (0.01)	0.02 (0.03)
DV	0.03	0.05 (0.03)	0.01 (0.00)	0.02(0.02)	0.03 (0.03)	0.05 (0.02)	0.02 (0.05)
GV	0.1	0.10 (0.04)	0.11 (0.05)	0.10 (0.04)	0.11 (0.05)	0.11 (0.07)	0.09 (0.09)
NH4	0.12	0.10 (0.02)	0.14 (0.04)	0.15 (0.03)	0.12 (0.04)	0.10 (0.04)	0.09 (0.05)
NO3	0.17	0.10 (0.05)	0.25 (0.12)	0.29 (0.08)	0.19 (0.12)	0.07 (0.04)	0.08 (0.05)
SO4	0.27	0.28 (0.08)	0.25 (0.08)	0.23 (0.07)	0.27 (0.09)	0.34 (0.14)	0.28 (0.19)
SOC	0.1	0.22 (0.07)	0.10 (0.04)	0.10 (0.05)	0.17 (0.08)	0.21 (0.09)	0.13 (0.19)

Effect modification of $PM_{2.5}$ by cluster-defined fractional composition:

Examining the effect modification of daily $PM_{2.5}$ by cluster membership allowed us to assess the health impact of the composition of $PM_{2.5}$ independently of its total mass. We found significant effect modification on lag day one (p = 0.004), but not lag zero or two-day effects. Within the lag day one model, we estimated the individual RR and 95% confidence intervals for each type of $PM_{2.5}$ composition presented in the appendix (Figure A.13 – A.15). Compared to the other clusters, the cluster that was identified as high NH_4, NO_3 with low COAL had a significantly lower RR of 0.86 (95% CI: [0.77,0.95]). This suggests

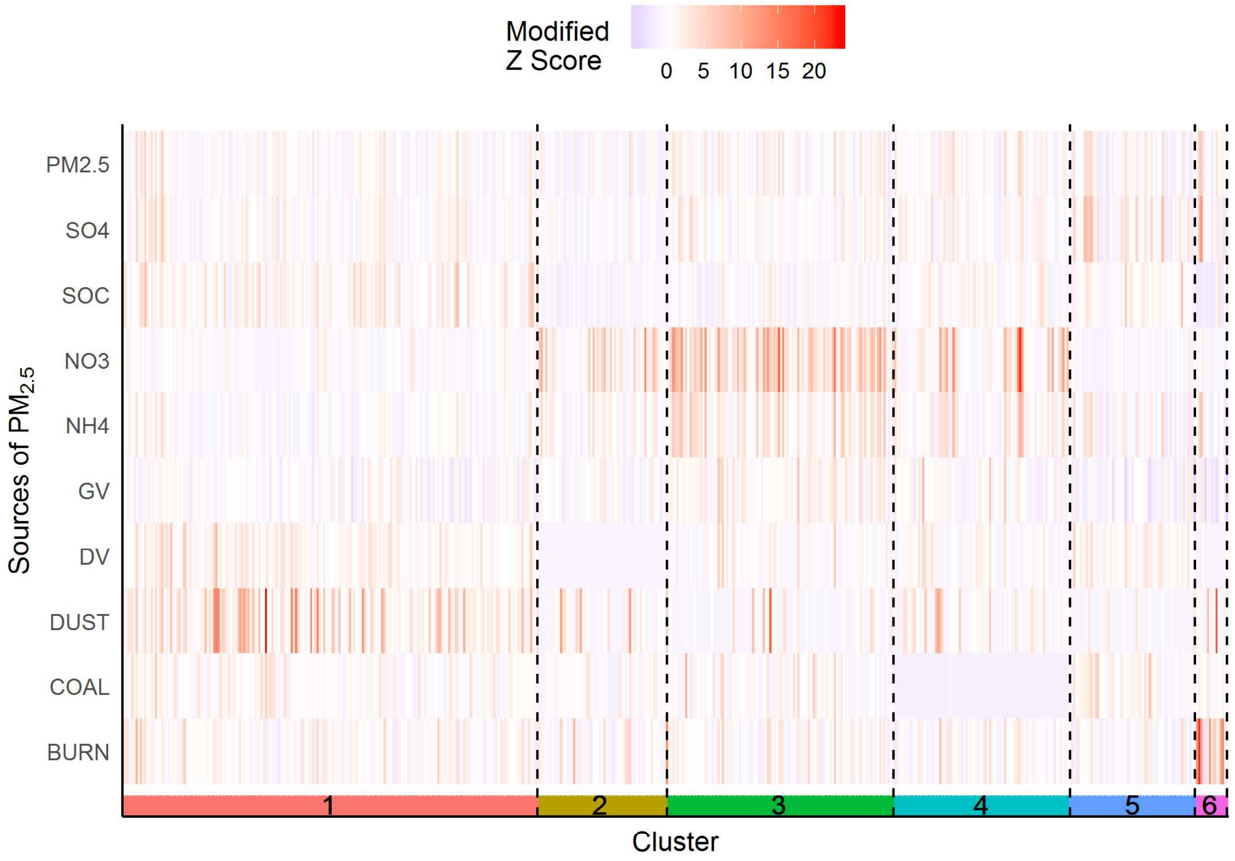


Figure 5.4

Figure 5.5: Heatmap of the sources of PM2.5 as modified Z scores grouped by cluster.

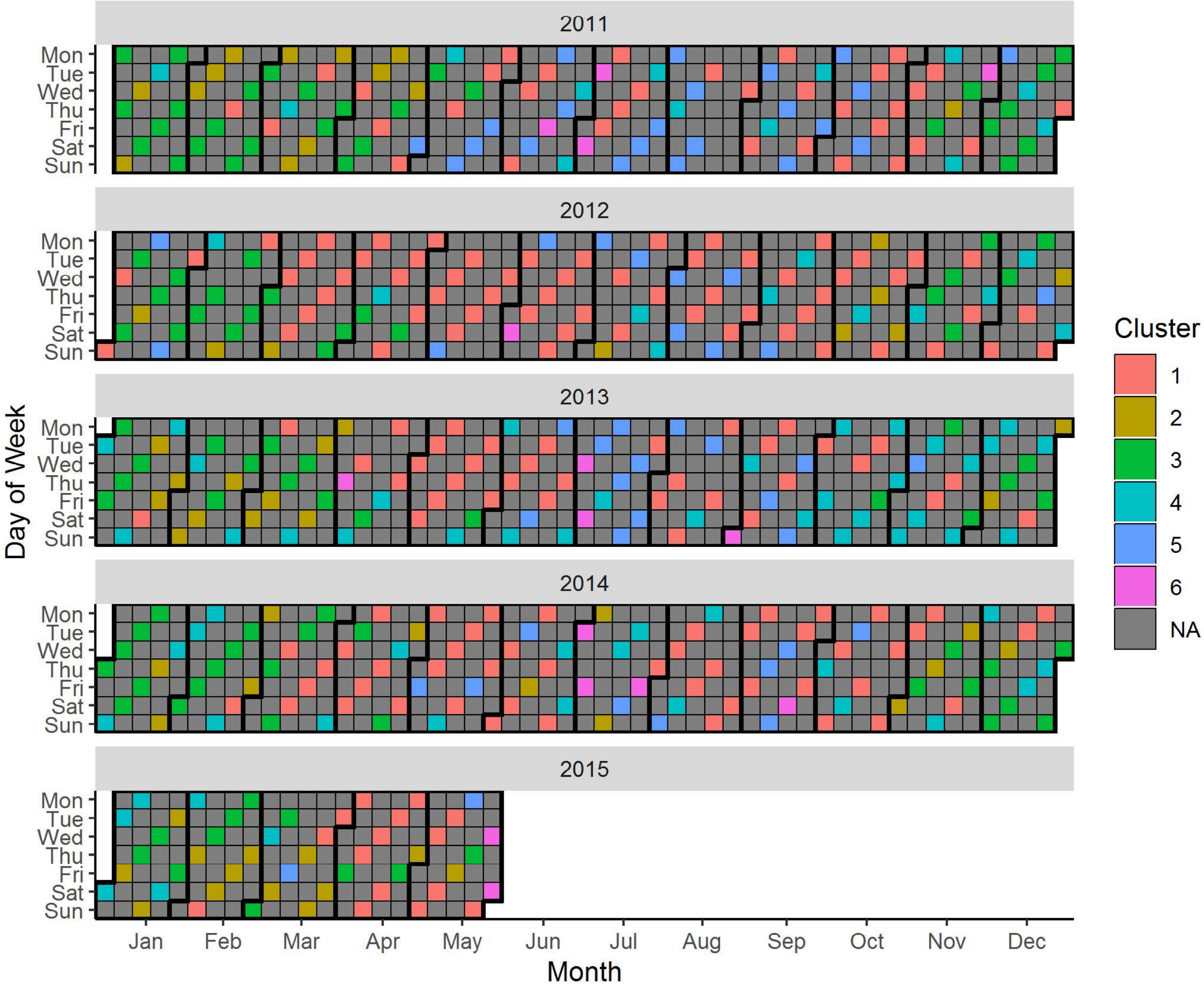


Figure 5.6

Figure 5.7: Calendar heat map of clusters. PM2.5 was observed only 1 in every three days and days with unobserved PM2.5 are grey.

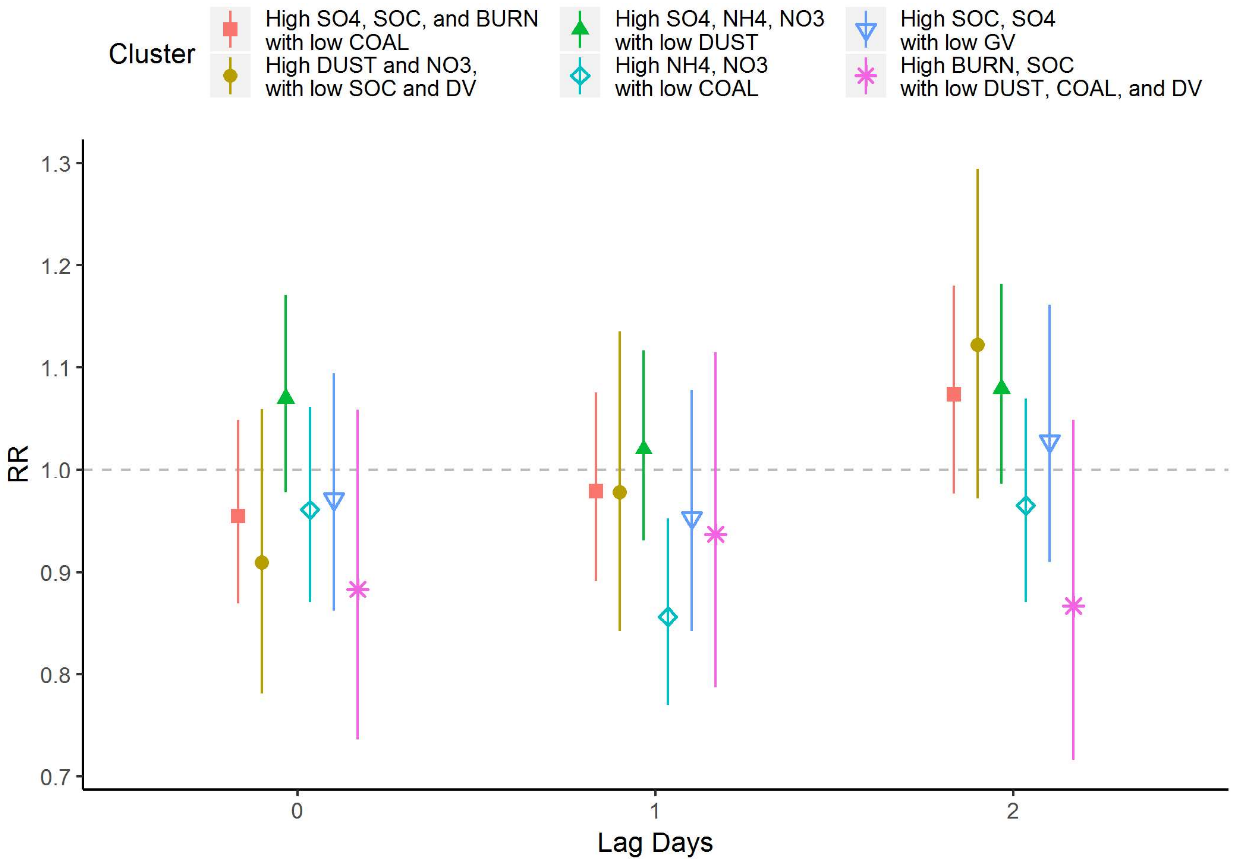


Figure 5.8

Figure 5.9: Cluster-specific risk ratios and 95% CIs for the number of daily asthma-related hospital utilizations for a 10 ug/m3 increase in PM2.5 and for each lag day

that an increase in $PM_{2.5}$ occurring on days characterized by high contributions of NH_4 and NO_3 and low contributions of COAL is associated with a smaller increased risk concerning asthma-related hospital utilization in comparison to other types of $PM_{2.5}$.

Discussion:

It has been of importance to figure out the most health-relevant sources of $PM_{2.5}$, both from scientific standpoints and regulatory perspectives (Bell et al., 2012, Hime et al., 2018, Heal et al., 2012). Association between source-specific exposure and health effect plays an important role to protect public health as well as develop policies (Brook et al., 2004, Brunekreef et al., 2002). Even though sources of air pollution were informative, these studies were often challenging to conduct because source- specific exposure was not directly observed but estimated (Zanobetti et al., 2009). Frequently, a two-stage approach was applied to estimate the source-specific health effect by adding source exposure assessment and health outcome regression (Krall et al., 2017). The previous study showed the ultrastructural study of the effect of air pollution by sulfate (SO2) on the respiratory system (Abdallah et al., 2007), the impact of climate and nitrate (NO_3) on prevalence of asthma (de Marco et al., 1999), effect of source-specific particulate matter on health (Adams et al., 2015), differential effect of source-specific particulate matter on emergency hospital utilization (Pun et al., 2014), health effect of short term exposure to source-specific particles (Samoli et al., 2016), source-specific fine particulate air pollution on heart disease (Siponen et al., 2015). Several studies highlighted the differential toxicities of fine particulate matter from various sources (Park et al., 2018) and clustered air pollution monitoring sites (Austin et al., 2013). In this study, we used model-based cluster analysis to classify days based on their source fraction profile. We observed a significant association of asthma with high coal. Previously, various research articles have also analyzed the adverse effect of $PM_{2.5}$ components in China (Deng et al., 2018, Norbäck et al., 2018, Zhao et al., 2007, Lee at al., 2006, Cai et al., 2014) and USA

(Nardone et al., 2018, Kravitz-Wirtz et al., 2018, Gharibi et al., 2018, Hansel et al., 2018). Limited research has been done on the impact of source-specific air pollution on asthma in the USA (Krall et al., 2017). We clustered each day based on the sources and assessed the seasonal variability in emission sources and source-specific health effects.

This article aimed to determine the source-specific contributions of $PM_{2.5}$ on daily hospital utilization rates related to asthma. We found that during summer its mostly high dust, organic carbon, and diesel vehicles in conjunction with moderately high $PM_{2.5}$ and sulfate. During the spring season high ammonium, nitrate, sulfate, and in the winter season organic carbon, sulfate, gasoline vehicle, and nitrate were the major sources of $PM_{2.5}$.Our results were in line with emerging evidence that supports source-specific emission regulation (Krall and Strickland, 2017).

The strength of this study is, we have used a novel data-driven clustering method to identify the sources of $PM_{2.5}$ and investigated the impacts of exposure to source-specific $PM_{2.5}$ on daily asthma ED utilization. Clustering on the fraction of sources allowed us to examine the effect of the composition of $PM_{2.5}$ independently of its mass. Model-based clustering has more advantages than other clustering methods. K-mean clustering and hierarchical clustering based on distance connectivity, and the number of clusters are preselected, and the cluster orientation changes with the scale. Another strength is that we have used EPA approved Chemical Mass Balance (CMB) model for $PM_{2.5}$ source apportionment (Al-Naiema et al., 2018, Ashrafi et al., 2018, Lu et al., 2018).CMB model has a powerful advantage to the source attribution process because the model interprets the actual measurement of the ambient data (Pace et al., 1991). However, several potential limitations should also be taken into consideration. First of all, the measurements taken in every three days of sampling. But due to varied lag days according to health data, we utilized all days of asthma counts. Another limitation was the lack of spatial resolution of sources. The spatial resolution of sources was not possible due to only one elemental $PM_{2.5}$ monitor in the study region. Coal burning identified as the major source of $PM_{2.5}$ associated with an increased risk of asthma

ED utilization in this study. By previous findings, Yu et al., 2018 observed that coal burning associated with increased health risk. Our big takeaway is that cluster 4 (low coal) did not show as much risk compared to other types of $PM_{2.5}$ exposures. Our study warrants further studies on identifying which types of $PM_{2.5}$ are most harmful to human health and also further explorations into how to direct primary prevention strategies towards the types of air pollution that show the most health effects.

Appendix

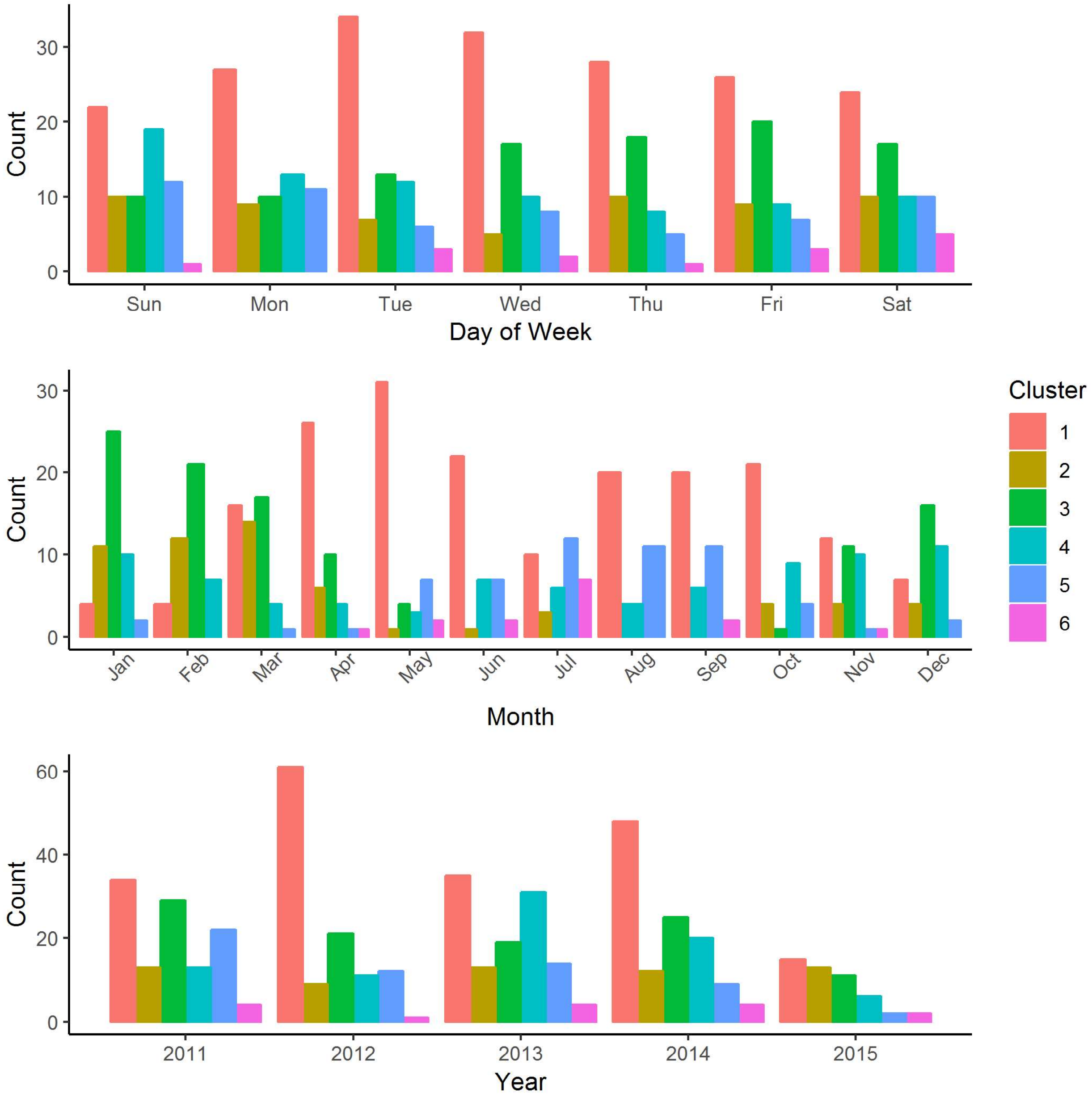
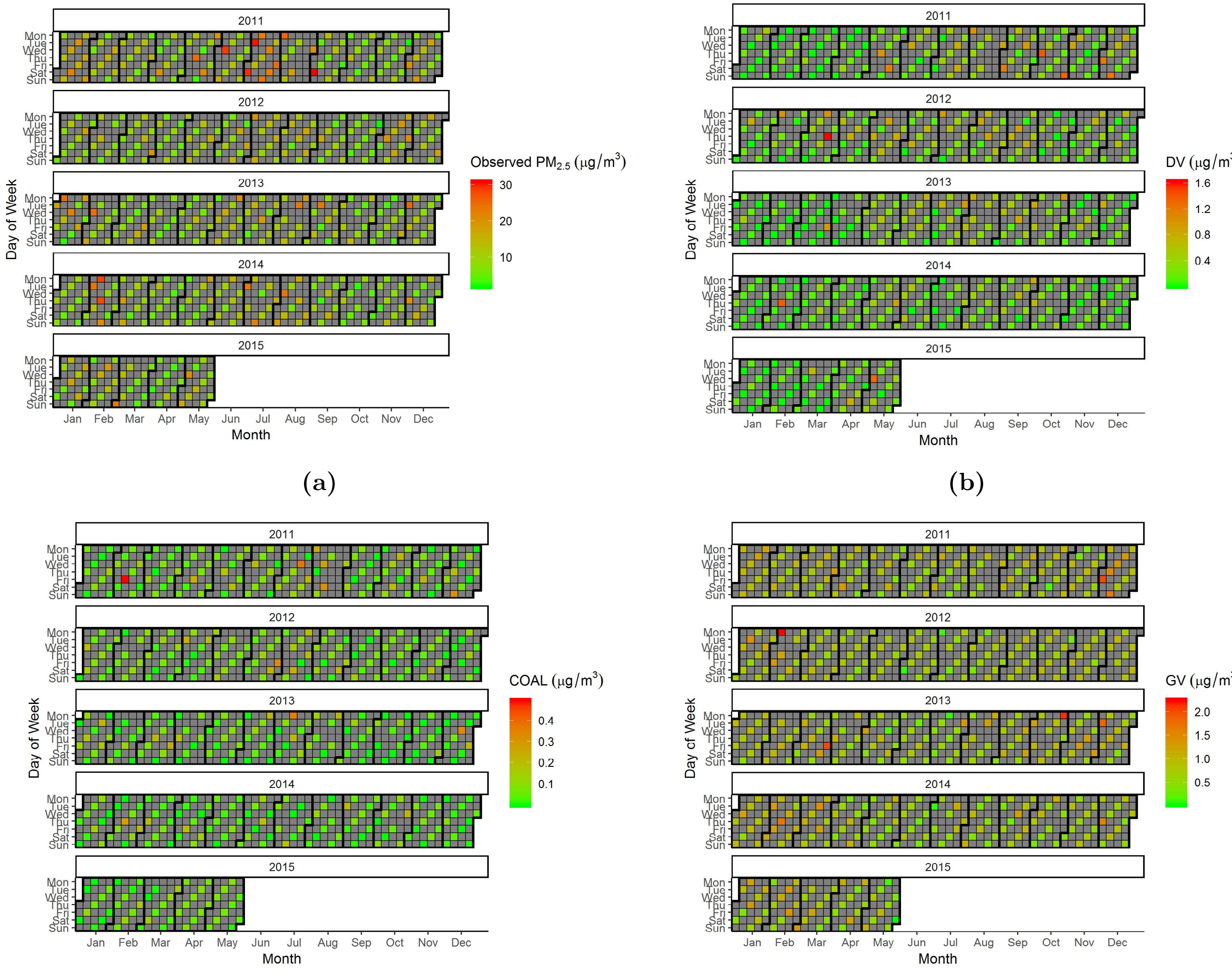


Figure 5.10

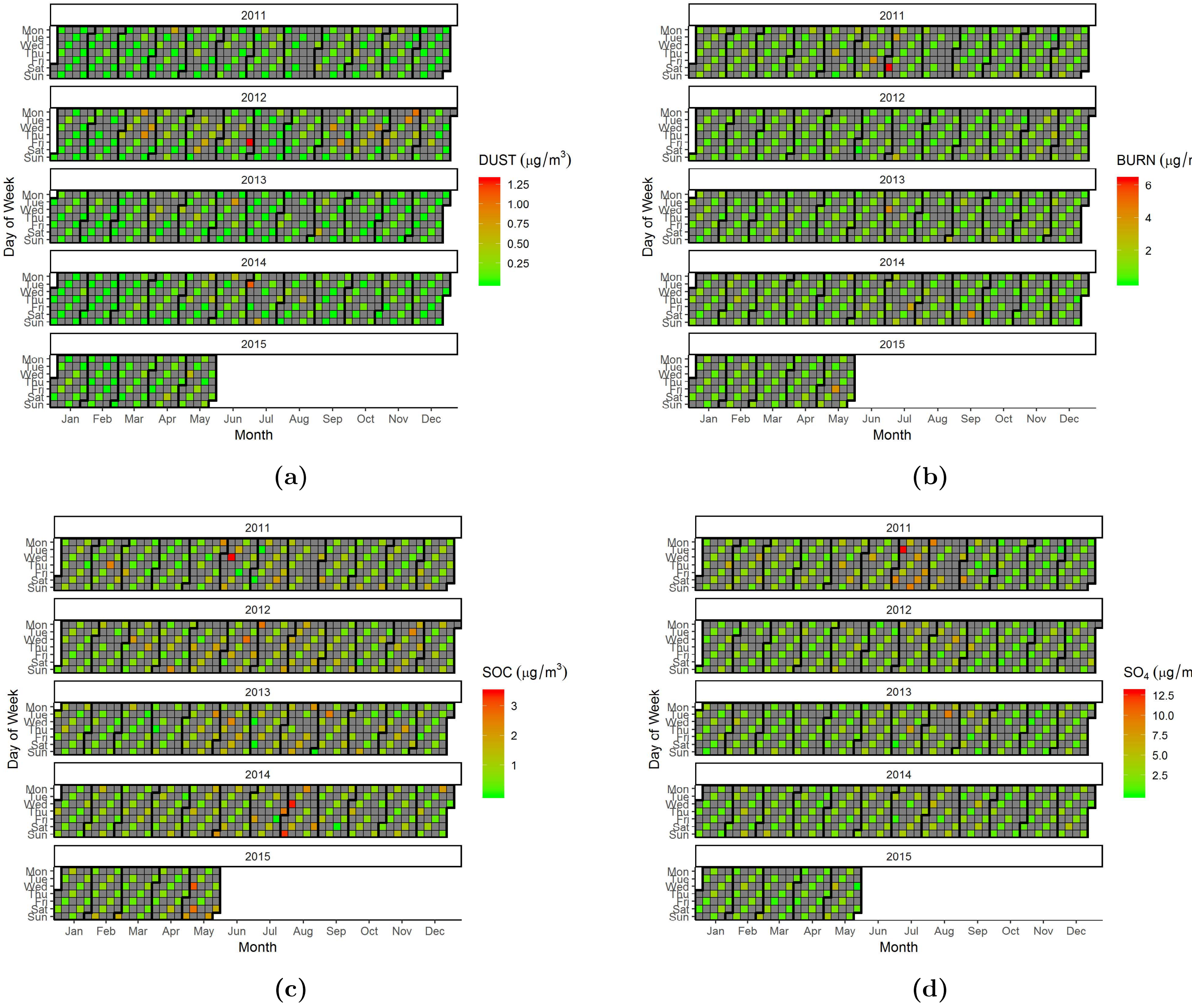
Figure 5.11: Bar plot of the clusters by week, month and year



(a) **(b)**

(c) **(d)**

Figure 5.12: Calendar heat map of (a)$PM_{2.5}$, (b) DV, (c) $COAL$, (d) GV concentration from January 2011 to May 2015



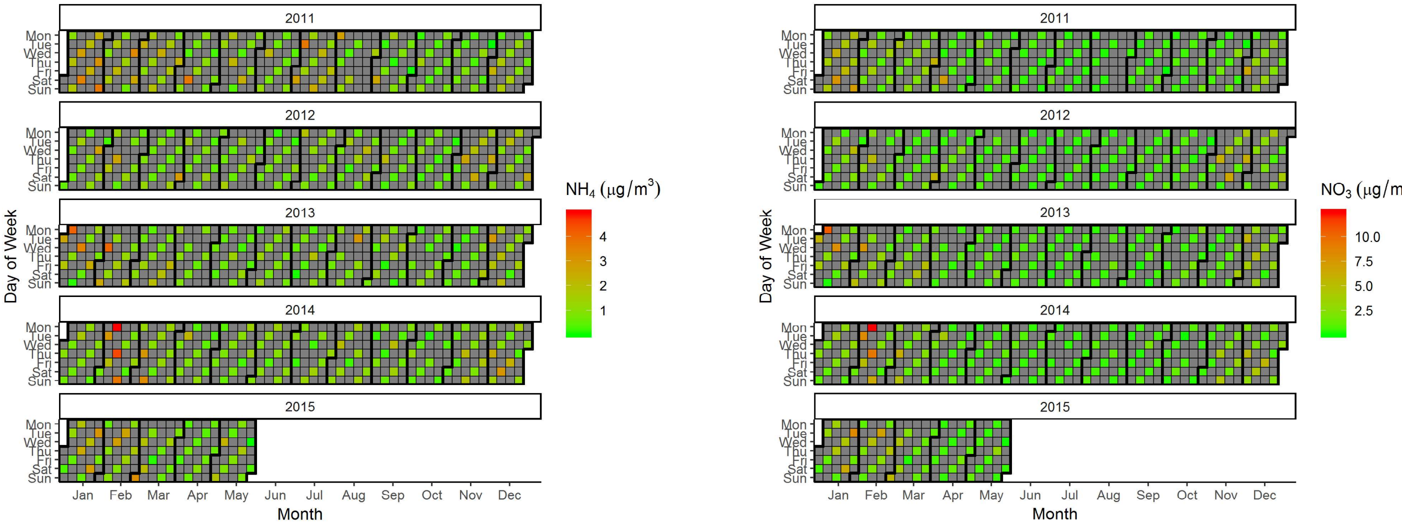
(a)

(b)

(c)

(d)

Figure 5.13: Calendar heat map of (a)$DUST$, (b) $BURN$, (c) SOC, (d) SO_4 concentration from January 2011 to May 2015



(a)

(b)

Figure 5.14: Calendar heat map of NH_3, NO_3 concentration from January 2011 to May 2015

Discussion

In this dissertation, I have proposed four statistical modeling approaches with methodologic developments and applications to real data. The common focus was to extend the existing methods to overcome the limitations that may hinder wider applications. The first approach, I focused on modeling growth curves. Growth curve models can be useful whenever there is a focus on the analysis of change over time, such as when examining developmental changes, evaluating treatment effects, or analyzing diary data. Although growth models go by a variety of different names, all of these approaches share a common focus on the estimation of individual differences in within-person change over time. Growth curve models estimate smoothed trajectories that are unique to each individual based on the set of observed repeated measures. One, of the limitation of current approaches is that they do not provide estimates of uncertainty of model parameters. The Bayesian paradigm produces uncertainty estimates of all model parameters inherently. It also handles the correlations among variables that are observed at various levels through hierarchical modeling framework. For example, the correlations between subject-specific covariates and within subject longitudinally observed data. I extended the widely used shape invariant model (SITAR) to Bayesian framework. Using the real data (heights of ADHD patients), I compared my Bayesian version to the original SITAR model and found similar parameter estimates (average age at peak velocity, size, tempo, and velocity) with slightly lower predictive errors. This development is amenable to other extensions, such as modeling the clustering patterns in growth curves and evaluating the impacts of various spline functions on parameter estimates. My future work will focus on methodologic development along these directions. One of the limitations of Bayesian SITAR is that it requires considerable time to converge depending on the data size, model complexity, and correlations among parameters. My future work will also focus on gaining efficiency in computation time by leveraging the recent advancement of Bayesian algorithms and parallel processing.

In the second approach, I focused on measures of association for categorical data. .Measures of association are used in various fields of research but are especially common in the areas of epidemiology and psychology, where they are frequently used to quantify relationships between exposures and diseases or behaviors. A measure of association may be determined by several analytic methods, including correlation analysis and regression analysis. The method used to determine the strength of association depending on the characteristics of the data for each variable. Data may be measured on an interval/ratio scale, an ordinal/rank scale, or a nominal/categorical scale. These three characteristics can be thought of as continuous, integer, and qualitative categories, respectively. A typical example for quantifying the association between two variables measured on an interval/ratio scale is the analysis of relationship between a person's height and weight. Each of these two characteristic variables is measured on a continuous scale. The appropriate measure of association for this situation is Pearson's correlation coefficient, r (rho), which measures the strength of the linear relationship between two variables on a continuous scale. The coefficient r takes on the values of -1 to $+1$. Values of -1 or $+1$ indicate a perfect linear relationship between the two variables, whereas a value of 0 indicates no linear relationship and negative values simply indicate the direction of the association, whereby as one variable increases, the other decreases. Correlation coefficients that differ from 0 but are not -1 or $+1$ indicate a linear relationship, although not a perfect linear relationship. In practice, ρ (the population correlation coefficient) is estimated by r, which is the correlation coefficient derived from sample data. Although Pearson's correlation coefficient is a measure of the strength of an association (specifically the linear relationship), it is not a measure of the significance of the association.

The significance of an association is a separate analysis of the sample correlation coefficient, r, using a student's t-test to measure the difference between the observed r and the expected r under the null hypothesis. Similarly, an odds ratio is an appropriate measure of strength of association for categorical data derived from a case-control study. The odds ratio is often interpreted the same way that relative risk is interpreted when measuring the strength of the association, although this is somewhat controversial when the risk factor being studied is common. Using simulated data, I compared the power and equality of three commonly used statistical method for testing association in clinical categorical data (the odds ratio, Pearson correlation, and canonical correlation). I showed the mathematical equality of the canonical correlation and Pearson correlation coefficients and found similar power for testing association between odds ratios and canonical correlation based on the Wald test and the Rao test. My future work will focus on expanding the 2*2 contingency table to k*k contingency table and compare the power analysis with real and simulated data.

In the third and fourth approaches, I focused on modeling the heath impact due to exposure to fine particles, $PM_{2.5}$. Health studies have shown a significant association between exposure to particle pollution and health risks. Health effects may include cardiovascular effects such as cardiac arrhythmia and heart attacks, and respiratory effects such as asthma attacks and bronchitis. Exposure to particle pollution can result in increased hospital admissions, emergency room visits, absences from school or work, and restricted activity days, especially for those with pre-existing heart or lung disease, older people, and children. The size of particles is directly linked to their potential for causing health problems. Fine particles ($PM_{2.5}$) pose the greatest health risk. The composition of $PM_{2.5}$ varies according to its sources, including fuel combustion from mobile sources like vehicles and stationary sources like power plants, industrial processes, and biomass burning. Several studies have aimed to quantify the heterogeneous composition of $PM_{2.5}$, including identification of distinct multi pollutant profiles, spatial clustering of air pollution monitoring sites, analyzing source specific contributions, and studying the effect of individual chemical constituents. However, much less attention has been focused on the source-specific contributions of particulate matter to health outcomes. Identifying sources contributing to $PM_{2.5}$ related health effects is critical to control the harmful sources of $PM_{2.5}$ and to identify primary prevention strategies. These fine particles can get deep into lungs and some may even get into the bloodstream. Exposure to these particles can affect a person's lungs and heart. In the third chapter I have shown the association between onset of Stroke and $PM_{2.5}$ using a novel case-crossover design and found differential impact by type of stroke and age at stroke. In the fourth and final chapter, I have shown the association of different sources of $PM_{2.5}$ with emergency hospital utilization due to asthma using a novel data-driven clustering technique and found seasonal variability in emission sources and source-specific impact on asthma admissions. My future work will focus on confirming these findings in upcoming studies with spatially observed data sets.

Bibliography

[1] Ibrahim L Abdalla. “Ultrastructural study of the effect of air pollution by SO2 on the respiratory air-ways”. In: *African Journal of Health Sciences* 14.3 (2007), pp. 129–136.

[2] Kate Adams et al. “Particulate matter components, sources, and health: Systematic approaches to testing effects”. In: *Journal of the Air & Waste Management Association* 65.5 (2015), pp. 544–558.

[3] Alan Agresti. *Analysis of ordinal categorical data*. Vol. 656. John Wiley & Sons, 2010.

[4] Alan Agresti. *Categorical data analysis*. Vol. 482. John Wiley & Sons, 2003.

[5] Y Akbas, C Takma, et al. “Canonical correlation analysis for studying the relationship between egg production traits and body weight, egg weight and age at sexual maturity in layers”. In: *Czech Journal of Animal Science* 50.4 (2005), pp. 163–168.

[6] Naomi S Altman and Julio Villarreal. “Self-Modeling Regression for Longitudinal Data with Time-Invariant Covariates”. In: ().

[7] Zorana J Andersen et al. “Ambient particle source apportionment and daily hospital admissions among children and elderly in Copenhagen”. In: *Journal of Exposure Science and Environmental Epidemiology* 17.7 (2007), p. 625.

[8] Jonathan O Anderson, Josef G Thundiyil, and Andrew Stolbach. “Clearing the air: a review of the effects of particulate matter air pollution on human health”. In: *Journal of Medical Toxicology* 8.2 (2012), pp. 166–175.

[9] Bianca Patrizia Andreini. “The first WHO global Conference on Air Pollution and Health”. In: *Italian Journal of Occupational and Environmental Hygiene* 9.4 (2019), p. 157.

[10] Elja Arjas, Liping Liu, and Niko Maglaperidze. “Prediction of growth: a hierarchical Bayesian approach”. In: 39.6 (1997), pp. 741–759.

[11] Khosro Ashrafi et al. “Source Apportionment of Total Suspended Particles (TSP) by Positive Matrix Factorization (PMF) and Chemical Mass Balance (CMB) Modeling in Ahvaz, Iran”. In: *Archives of environmental contamination and toxicology* 75.2 (2018), pp. 278–294.

[12] Richard W Atkinson et al. “Acute effects of particulate air pollution on respiratory admissions: results from APHEA 2 project”. In: *American journal of respiratory and critical care medicine* 164.10 (2001), pp. 1860–1866.

[13] Elena Austin et al. “A framework for identifying distinct multipollutant profiles in air pollution data”. In: *Environment international* 45 (2012), pp. 112–121.

[14] Elena Austin et al. “A framework to spatially cluster air pollution monitoring sites in US based on the PM2. 5 composition”. In: *Environment international* 59 (2013), pp. 244–254.

[15] Francis R Bach and Michael I Jordan. “A probabilistic interpretation of canonical correlation analysis”. In: (2005).

[16] Nour Baız and Isabella Annesi-Maesano. “Is the asthma epidemic still ascending?” In: *Clinics in chest medicine* 33.3 (2012), pp. 419–429.

[17] Emmanuel S Baja et al. “Traffic-related air pollution and QT interval: modification by diabetes, obesity, and oxidative stress gene polymorphisms in the normative aging study”. In: *Environmental health perspectives* 118.6 (2010), pp. 840–846.

[18] József Baranyi and Terry A Roberts. “A dynamic approach to predicting bacterial growth in food”. In: *International journal of food microbiology* 23.3-4 (1994), pp. 277–294.

[19] Daniel Barry. “A Bayesian model for growth curve analysis”. In: *Biometrics* (1995), pp. 639–655.

[20] Carlo R Bartoli et al. “Mechanisms of inhaled fine particulate air pollution–induced arterial blood pressure changes”. In: *Environmental health perspectives* 117.3 (2008), pp. 361–366.

[21] Thomas F Bateson and Joel Schwartz. “Children’s response to air pollutants”. In: *Journal of Toxicology and Environmental Health, Part A* 71.3 (2007), pp. 238–243.

[22] Heather J Beacon, Simon G Thompson, and Peter D England. “The analysis of complex patterns of longitudinal binary response: an example of transient dysphagia following radiotherapy”. In: *Statistics in medicine* 17.22 (1998), pp. 2551–2561.

[23] Ken J Beath. “Infant growth modelling using a shape invariant model with random effects”. In: *Statistics in medicine* 26.12 (2007), pp. 2547–2564.

[24] Andrew F Beck et al. “Pervasive income-based disparities in inpatient bed-day rates across conditions and subspecialties”. In: *Health Affairs* 37.4 (2018), pp. 551–559.

[25] Mark P Becker and Clifford C Clogg. “Analysis of sets of two-way contingency tables using association models”. In: *Journal of the American statistical association* 84.405 (1989), pp. 142–151.

[26] Sam Behseta and Robert E Kass. “Testing equality of two functions using BARS”. In: *Statistics in medicine* 24.22 (2005), pp. 3523–3534.

[27] Sam Behseta, Robert E Kass, and Garrick L Wallstrom. “Hierarchical models for assessing variability among functions”. In: *Biometrika* 92.2 (2005), pp. 419–434.

[28] Michelle L Bell. “Assessment of the health impacts of particulate matter characteristics.” In: *Research Report (Health Effects Institute)* 161 (2012), pp. 5–38.

[29] Michelle L Bell, Keita Ebisu, and Kathleen Belanger. “Ambient air pollution and low birth weight in Connecticut and Massachusetts”. In: *Environmental health perspectives* 115.7 (2007), pp. 1118–1124.

[30] Norbert Berend. “Contribution of air pollution to COPD and small airway dysfunction”. In: *Respirology* 21.2 (2016), pp. 237–244.

[31] Catherine S Berkey. “Comparison of two longitudinal growth models for preschool children”. In: *Biometrics* (1982), pp. 221–234.

[32] Catherine S Berkey and Nan M Laird. “Nonlinear growth curve analysis: estimating the population parameters”. In: *Annals of human biology* 13.2 (1986), pp. 111–128.

[33] Norman E Breslow and David G Clayton. “Approximate inference in generalized linear mixed models”. In: *Journal of the American statistical Association* 88.421 (1993), pp. 9–25.

[34] Joseph Broderick et al. “The Greater Cincinnati/Northern Kentucky Stroke Study: preliminary first-ever and total incidence rates of stroke among blacks”. In: *Stroke* 29.2 (1998), pp. 415–421.

[35] Cole Brokamp et al. “Decentralized and reproducible geocoding and characterization of community and environmental exposures for multisite studies”. In: *Journal of the American Medical Informatics Association* 25.3 (2017), pp. 309–314.

[36] Cole Brokamp et al. “Predicting daily urban fine particulate matter concentrations using a random forest model”. In: *Environmental science & technology* 52.7 (2018), pp. 4173–4179.

[37] Robert D Brook. “You are what you breathe: evidence linking air pollution and blood pressure”. In: *Current hypertension reports* 7.6 (2005), pp. 427–434.

[38] Robert D Brook et al. "Air pollution and cardiovascular disease: a statement for healthcare professionals from the Expert Panel on Population and Prevention Science of the American Heart Association". In: *Circulation* 109.21 (2004), pp. 2655–2671.

[39] Robert D Brook et al. "Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association". In: *Circulation* 121.21 (2010), pp. 2331–2378.

[40] Lyndia C Brumback and Mary J Lindstrom. "Self modeling with flexible, random time transformations". In: *Biometrics* 60.2 (2004), pp. 461–470.

[41] Bert Brunekreef and Stephen T Holgate. "Air pollution and health". In: *The lancet* 360.9341 (2002), pp. 1233–1242.

[42] Jennifer Brunet et al. "The association between past and current physical activity and depressive symptoms in young adults: a 10-year prospective study". In: *Annals of epidemiology* 23.1 (2013), pp. 25–30.

[43] Richard T Burnett et al. "The role of particulate size and chemistry in the association between summertime ambient air pollution and hospitalization for cardiorespiratory diseases." In: *Environmental health perspectives* 105.6 (1997), pp. 614–620.

[44] Jing Cai et al. "Acute effects of air pollution on asthma hospitalization in Shanghai, China". In: *Environmental pollution* 191 (2014), pp. 139–144.

[45] Eduardo Carracedo-Martınez et al. "Case-crossover analysis of air pollution health effects: a systematic review of methodology and application". In: *Environmental health perspectives* 118.8 (2010), pp. 1173–1182.

[46] K Chen et al. "The effects of air pollution on asthma hospital admissions in Adelaide, South Australia, 2003–2013: time-series and case–crossover analyses". In: *Clinical & Experimental Allergy* 46.11 (2016), pp. 1416–1430.

[47] D Allen Chu et al. “Global monitoring of air pollution over land from the Earth Observing System-Terra Moderate Resolution Imaging Spectroradiometer (MODIS)”. In: *Journal of Geophysical Research: Atmospheres* 108.D21 (2003).

[48] Nina Annika Clark et al. “Effect of early life exposure to air pollution on development of childhood asthma”. In: *Environmental health perspectives* 118.2 (2009), pp. 284–290.

[49] William G Cochran. “Some methods for strengthening the common χ 2 tests”. In: *Biometrics* 10.4 (1954), pp. 417–451.

[50] William G Cochran. “The χ2 test of goodness of fit”. In: *The Annals of Mathematical Statistics* (1952), pp. 315–345.

[51] Tim J Cole, Malcolm DC Donaldson, and Yoav Ben-Shlomo. “SITAR?a useful instrument for growth curve analysis”. In: *International journal of epidemiology* 39.6 (2010), pp. 1558–1566.

[52] Tim J Cole, Jenny V Freeman, and Michael A Preece. “British 1990 growth reference centiles for weight, height, body mass index and head circumference fitted by maximum penalized likelihood”. In: *Statistics in medicine* 17.4 (1998), pp. 407–429.

[53] Timothy J Cole and Pamela J Green. “Smoothing reference centile curves: the LMS method and penalized likelihood”. In: *Statistics in medicine* 11.10 (1992), pp. 1305–1319.

[54] David Collett. *Modelling binary data. Texts in statistical science.* 2003.

[55] Brian A Cosgrove et al. “Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project”. In: *Journal of Geophysical Research: Atmospheres* 108.D22 (2003).

[56] Earl W Count. “growth patterns of the human physique: an approach to kinetic anthropometry: part I”. In: *Human Biology* 15.1 (1943), pp. 1–32.

[57] C Cox. “Delta method”. In: *Encyclopedia of biostatistics* 2 (2005).

[58] Qihong Deng et al. “Parental stress and air pollution increase childhood asthma in China”. In: *Environmental research* 165 (2018), pp. 23–31.

[59] Kim Marie Deschamps. “Associations between daily asthma hospital admissions and ambient air pollutants in Montreal, 1992 to 1999”. PhD thesis. Concordia University, 2003.

[60] Laurent Devien et al. “Sources of household air pollution: The association with lung function and respiratory symptoms in middle-aged adult”. In: *Environmental research* 164 (2018), pp. 140–148.

[61] Ana V Diez Roux et al. “Long-term exposure to ambient particulate matter and prevalence of subclinical atherosclerosis in the Multi-Ethnic Study of Atherosclerosis”. In: *American journal of epidemiology* 167.6 (2008), pp. 667–675.

[62] Peter Diggle et al. *Analysis of longitudinal data*. Oxford university press, 2002.

[63] Ilaria DiMatteo, Christopher R Genovese, and Robert E Kass. “Bayesian curve-fitting with free-knot splines”. In: *Biometrika* 88.4 (2001), pp. 1055–1071.

[64] Ling Ding et al. “Air pollution and asthma attacks in children: A case–crossover analysis in the city of Chongqing, China”. In: *Environmental pollution* 220 (2017), pp. 348–353.

[65] Douglas W Dockery and C Arden Pope. “Acute respiratory effects of particulate air pollution”. In: *Annual review of public health* 15.1 (1994), pp. 107–132.

[66] Douglas W Dockery et al. “An association between air pollution and mortality in six US cities”. In: *New England journal of medicine* 329.24 (1993), pp. 1753–1759.

[67] Francesca Dominici et al. “Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases”. In: *Jama* 295.10 (2006), pp. 1127–1134.

[68] William P Dunlap, Charles J Brody, and Tammy Greer. “Canonical correlation and chi-square: relationships and interpretation”. In: *The Journal of general psychology* 127.4 (2000), pp. 341–353.

[69] David B Dunson. “Commentary: practical advantages of Bayesian analysis of epidemiologic data”. In: *American journal of Epidemiology* 153.12 (2001), pp. 1222–1226.

[70] Sylvain Durrleman and Richard Simon. “Flexible regression models with cubic splines”. In: *Statistics in medicine* 8.5 (1989), pp. 551–561.

[71] Jill A Engel-Cox et al. “Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality”. In: *Atmospheric environment* 38.16 (2004), pp. 2495–2509.

[72] Morten W Fagerland, Stian Lydersen, and Petter Laake. “Recommended tests and confidence intervals for paired binomial proportions”. In: *Statistics in medicine* 33.16 (2014), pp. 2850–2875.

[73] Valery L Feigin et al. “Global burden of stroke and risk factors in 188 countries, during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013”. In: *The Lancet Neurology* 15.9 (2016), pp. 913–924.

[74] Baihuan Feng et al. “High level of source-specific particulate matter air pollution associated with cardiac arrhythmias”. In: *Science of The Total Environment* 657 (2019), pp. 1285–1293.

[75] WHO Task Force. “Stroke-1989. Recommendations on stroke prevention, diagnosis, and therapy. Report of the WHO Task Force on Stroke and other Cerebrovascular Disorders”. In: *Stroke* 20.10 (1989), pp. 1407–1431.

[76] Chris Fraley and Adrian E Raftery. “How many clusters? Which clustering method? Answers via model-based cluster analysis”. In: *The computer journal* 41.8 (1998), pp. 578–588.

[77] Chris Fraley and Adrian E Raftery. *MCLUST version 3: an R package for normal mixture modeling and model-based clustering*. Tech. rep. WASHINGTON UNIV SEATTLE DEPT OF STATISTICS, 2006.

[78] Chris Fraley and Adrian E Raftery. “Model-based clustering, discriminant analysis, and density estimation”. In: *Journal of the American statistical Association* 97.458 (2002), pp. 611–631.

[79] Mitchell H Gail, Jay H Lubin, and Lawrence V Rubinstein. “Likelihood calculations for matched case-control studies and survival studies with tied death times”. In: *Biometrika* 68.3 (1981), pp. 703–707.

[80] Th Gasser et al. “A method for determining the dynamics and intensity of average growth”. In: *Annals of Human Biology* 17.6 (1990), pp. 459–474.

[81] Ulrike Gehring et al. “Traffic-related air pollution and the development of asthma and allergies during the first 8 years of life”. In: *American journal of respiratory and critical care medicine* 181.6 (2010), pp. 596–603.

[82] Andrew Gelman. *Prior distributions for variance parameters in hierarchical models*. Tech. rep. EERI Research Paper Series, 2004.

[83] Andrew Gelman and Donald B Rubin. “Inference from iterative simulation using multiple sequences”. In: *Statistical science* (1992), pp. 457–472.

[84] Diklah Geva et al. “A longitudinal analysis of the effect of prenatal alcohol exposure on growth”. In: *Alcoholism: Clinical and Experimental Research* 17.6 (1993), pp. 1124–1129.

[85] Hamed Gharibi et al. “Pesticide and Acute Asthma Attacks in California, Usa in 2005 to 2011: A Bidirectional Symmetric Case-Crossover Study”. In: *ISEE Conference Abstracts*. Vol. 2018. 1. 2018.

[86] Zvi Gilula. “Grouping and association in contingency tables: an exploratory canonical correlation approach”. In: *Journal of the American Statistical Association* 81.395 (1986), pp. 773–779.

[87] Zvi Gilula. “On some similarities between canonical correlation models and latent class models for two-way contingency tables”. In: *Biometrika* 71.3 (1984), pp. 523–529.

[88] Zvi Gilula and Shelby J Haberman. “The analysis of multivariate contingency tables by restricted canonical and restricted association models”. In: *Journal of the American Statistical Association* 83.403 (1988), pp. 760–771.

[89] Robert Gittins. *Canonical analysis: a review with applications in ecology.* Vol. 12. Springer Science & Business Media, 2012.

[90] William J Glynn and Robb J Muirhead. “Inference in canonical correlation analysis”. In: (1978).

[91] Harvey Goldstein et al. “Multilevel growth curve models that incorporate a random coefficient model for the level 1 variance function”. In: *Statistical methods in medical research* (2017), p. 0962280217706728.

[92] Melissa J Griffin, Jeffrey D Wardell, and Jennifer P Read. “Recent sexual victimization and drinking behavior in newly matriculated college students: A latent growth analysis.” In: *Psychology of Addictive behaviors* 27.4 (2013), p. 966.

[93] Pi Guo et al. “Ambient air pollution and risk for ischemic stroke: a short-term exposure assessment in South China”. In: *International journal of environmental research and public health* 14.9 (2017), p. 1091.

[94] Pawan Gupta et al. “Multi year satellite remote sensing of particulate matter air quality over Sydney, Australia”. In: *International Journal of Remote Sensing* 28.20 (2007), pp. 4483–4498.

[95] Pawan Gupta et al. “Satellite remote sensing of particulate matter and air quality assessment over global cities”. In: *Atmospheric Environment* 40.30 (2006), pp. 5880–5892.

[96] Mohammad Z Al-Hamdan et al. “Environmental public health applications using remotely sensed data”. In: *Geocarto international* 29.1 (2014), pp. 85–98.

[97] Mohammad Z Al-Hamdan et al. “Methods for characterizing fine particulate matter using ground observations and remotely sensed data: potential use for environmental public health surveillance”. In: *Journal of the Air & Waste Management Association* 59.7 (2009), pp. 865–881.

[98] T Hastie, R Tibshirani, and JJH Friedman. “The elements of statistical learning (Vol. 1): Springer New York”. In: (2001).

[99] William L Hayes. *Statistics for psychologists*. 1963.

[100] Mathew R Heal, Prashant Kumar, and Roy M Harrison. “Particles, air quality, policy and health”. In: *Chemical Society Reviews* 41.19 (2012), pp. 6606–6630.

[101] Christopher Hertzog and John R Nesselroade. “Assessing psychological change in adulthood: an overview of methodological issues.” In: *Psychology and aging* 18.4 (2003), p. 639.

[102] Neil Hime, Guy Marks, and Christine Cowie. “A comparison of the health effects of ambient particulate matter air pollution from five emission sources”. In: *International journal of environmental research and public health* 15.6 (2018), p. 1206.

[103] Philip K Hopke. *Receptor modeling for air quality management*. Vol. 7. Elsevier, 1991.

[104] John Hornstein. “Atmospheric Thermodynamics”. In: *Eos, Transactions American Geophysical Union* 79.45 (1998), pp. 548–548.

[105] George Howard et al. “Racial and geographic differences in awareness, treatment, and control of hypertension: the REasons for Geographic And Racial Differences in Stroke study”. In: *Stroke* 37.5 (2006), pp. 1171–1178.

[106] Alan Huang, Matthew P Wand, et al. “Simple marginally noninformative prior distributions for covariance matrices”. In: *Bayesian Analysis* 8.2 (2013), pp. 439–452.

[107] Paul D Isaac and Glenn W Milligan. “A comment on the use of canonical correlation in the analysis of contingency tables.” In: (1983).

[108] Joshua J Jackson et al. “Can an old dog learn (and want to experience) new tricks? Cognitive training increases openness to experience in older adults.” In: *Psychology and aging* 27.2 (2012), p. 286.

[109] Gareth James et al. *An introduction to statistical learning.* Vol. 112. Springer, 2013.

[110] Holly Janes, Lianne Sheppard, and Thomas Lumley. “Case-crossover analyses of air pollution exposure data: referent selection strategies and their implications for bias”. In: *Epidemiology* (2005), pp. 717–726.

[111] Rachel M Jenss and Nancy Bayley. “A mathematical method for studying the growth of a child”. In: *Human Biology* 9.4 (1937), p. 556.

[112] Johan Karlberg. “On the modelling of human growth”. In: *Statistics in medicine* 6.2 (1987), pp. 185–192.

[113] Robbert E Kass, Luke Tierney, and Joseph B Kadane. “Approximate methods for assessing influence and sensitivity in Bayesian analysis”. In: *Biometrika* 76.4 (1989), pp. 663–674.

[114] Chunlei Ke and Yuedong Wang. “Semiparametric nonlinear mixed-effects models and their applications”. In: *Journal of the American Statistical Association* 96.456 (2001), pp. 1272–1298.

[115] Marianthi-Anna Kioumourtzoglou et al. “PM2. 5 and survival among older adults: effect modification by particulate composition”. In: *Epidemiology (Cambridge, Mass.)* 26.3 (2015), p. 321.

[116] Dawn Kleindorfer et al. “The unchanging incidence and case-fatality of stroke in the 1990s: a population-based study”. In: *Stroke* 37.10 (2006), pp. 2473–2478.

[117] Thomas R Knapp. “The analysis of the data for two-way contingency tables”. In: *Research in nursing & health* 22.3 (1999), pp. 263–268.

[118] Jenna R Krall and Matthew J Strickland. “Recent approaches to estimate associations between source-specific air pollution and health”. In: *Current environmental health reports* 4.1 (2017), pp. 68–78.

[119] Nicole Kravitz-Wirtz et al. “Early-Life Air Pollution Exposure, Neighborhood Poverty, and Childhood Asthma in the United States, 1990–2014”. In: *International journal of environmental research and public health* 15.6 (2018), p. 1114.

[120] Robert J Kuczmarski. “2000 CDC growth charts for the United States; methods and development”. In: (2002).

[121] Holly Ching-yu Lam et al. “The short-term association between asthma hospitalisations, ambient temperature, other meteorological factors and air pollutants in Hong Kong: a time-series study”. In: *Thorax* 71.12 (2016), pp. 1097–1109.

[122] Paul C Lambert et al. “Analysis of ambulatory blood pressure monitor data using a hierarchical model incorporating restricted cubic splines and heterogeneous within-subject variances”. In: *Statistics in medicine* 20.24 (2001), pp. 3789–3805.

[123] H Lancaster. “0.(1969). The chi-squared distribution”. In: *Wiley, New York. MR* 40 (2002), p. 6667.

[124] Sophie Larrieu et al. “Short term effects of air pollution on hospitalizations for cardiovascular diseases in eight French cities: the PSAS program”. In: *Science of the Total Environment* 387.1-3 (2007), pp. 105–112.

[125] Pablo M Lavados, Verónica V Olavarrıa, and Lorena Hoffmeister. “Ambient temperature and stroke risk: evidence supporting a short-term effect at a population level from acute environmental exposures”. In: *Stroke* 49.1 (2018), pp. 255–261.

[126] WH Lawton, EA Sylvestre, and MS Maggio. “Self modeling nonlinear regression”. In: *Technometrics* 14.3 (1972), pp. 513–532.

[127] SL Lee, WHS Wong, and YL Lau. “Association between air pollution and asthma admission among children in Hong Kong”. In: *Clinical & Experimental Allergy* 36.9 (2006), pp. 1138–1146.

[128] Angelika van der Linde. “Dimension reduction with linear discriminant functions based on an odds ratio parameterization”. In: *International statistical review* 71.3 (2003), pp. 629–666.

[129] Mary J Lindstrom. “Self-modelling with random shift and scale parameters and a free-knot spline shape function”. In: *Statistics in medicine* 14.18 (1995), pp. 2009–2021.

[130] Lynda D Lisabeth et al. “Ambient air pollution and risk for ischemic stroke and transient ischemic attack”. In: *Annals of neurology: official journal of the American neurological association and the Child neurology society* 64.1 (2008), pp. 53–59.

[131] Maria M Llabre et al. “Applying latent growth curve modeling to the investigation of individual differences in cardiovascular recovery from stress”. In: *Psychosomatic Medicine* 66.1 (2004), pp. 29–41.

[132] S López-Pintado and IW McKeague. “Growthrate: Bayesian reconstruction of growth velocity”. In: *R package version* 1 (2011).

[133] Zhaojie Lu et al. “A hybrid source apportionment strategy using positive matrix factorization (PMF) and molecular marker chemical mass balance (MM-CMB) models”. In: *Environmental pollution* 238 (2018), pp. 39–51.

[134] Thomas Lumley and Drew Levy. “Bias in the case–crossover design: implications for studies of air pollution”. In: *Environmetrics: The official journal of the International Environmetrics Society* 11.6 (2000), pp. 689–704.

[135] Ejnar Lyttkens. “Regression aspects of canonical correlation”. In: *Journal of Multivariate Analysis* 2.4 (1972), pp. 418–439.

[136] Malcolm Maclure. “The case-crossover design: a method for studying transient effects on the risk of acute events”. In: *American journal of epidemiology* 133.2 (1991), pp. 144–153.

[137] R de Marco et al. “The impact of climate and NO 2 outdoor pollution on prevalence of asthma and allergic rhinitis in Italy.” In: ().

[138] Leslie A McClure et al. “Fine Particulate Matter (PM2. 5) and the Risk of Stroke in the REGARDS Cohort”. In: *Journal of Stroke and Cerebrovascular Diseases* 26.8 (2017), pp. 1739–1744.

[139] Ian W McKeague et al. “Analyzing growth trajectories”. In: *Journal of developmental origins of health and disease* 2.6 (2011), pp. 322–329.

[140] Linda O Mearns et al. “The North American regional climate change assessment program: overview of phase I results”. In: *Bulletin of the American Meteorological Society* 93.9 (2012), pp. 1337–1362.

[141] Seema Mihrshahi et al. “The childhood asthma prevention study (CAPS): design and research protocol of a randomized trial for the primary prevention of asthma”. In: *Controlled clinical trials* 22.3 (2001), pp. 333–354.

[142] S Milani, A Bossi, and E Marubini. “Individual growth curves and longitudinal growth charts between 0 and 3 years”. In: *Acta Paediatrica* 78.s350 (1989), pp. 95–104.

[143] Elizabeth Mostofsky, Brent A Coull, and Murray A Mittleman. “Analysis of Observational Self-matched Data to Examine Acute Triggers of Outcome Events with Abrupt Onset”. In: *Epidemiology* 29.6 (2018), pp. 804–816.

[144] Ibrahim M Al-Naiema et al. “Source apportionment of fine particulate matter organic carbon in Shenzhen, China by chemical mass balance and radiocarbon methods”. In: *Environmental pollution* 240 (2018), pp. 34–43.

[145] Anthony Nardone et al. “Ambient air pollution and asthma-related outcomes in children of color of the USA: a scoping review of literature published between 2013 and 2017”. In: *Current allergy and asthma reports* 18.5 (2018), p. 29.

[146] Marie-Louise Newell, Mario-Cortina Borja, and Catherine Peckham. “Height, weight, and growth in children born to mothers with HIV-1 infection in Europe.” In: *Pediatrics* 111.1 (2003), e52–60.

[147] Dan Norbäck et al. “Asthma and rhinitis among Chinese children—indoor and outdoor air pollution and indicators of socioeconomic status (SES)”. In: *Environment international* 115 (2018), pp. 1–8.

[148] Martin J O’Donnell et al. “Fine particulate air pollution (PM2. 5) and the risk of acute ischemic stroke”. In: *Epidemiology (Cambridge, Mass.)* 22.3 (2011), p. 422.

[149] ME O’Neill. “A note on the canonical correlations from contingency tables”. In: *Australian Journal of Statistics* 23.1 (1981), pp. 58–66.

[150] ME O’Neill. “Asymptotic distributions of the canonical correlations from contingency tables”. In: *Australian Journal of Statistics* 20.1 (1978), pp. 75–82.

[151] ME O’neill. “Distributional expansions for canonical correlations from contingency tables”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1978), pp. 303–312.

[152] Ken KL Ong et al. “Size at birth and early childhood growth in relation to maternal smoking, parity and infant breast-feeding: longitudinal birth cohort study and analysis”. In: *Pediatric research* 52.6 (2002), p. 863.

[153] World Health Organization et al. *WHO child growth standards: length/height for age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age, methods and development.* World Health Organization, 2006.

[154] Minhan Park et al. “Differential toxicities of fine particulate matters from various sources”. In: *Scientific reports* 8.1 (2018), p. 17007.

[155] Mathilde Pascal et al. “Assessing the public health impacts of urban air pollution in 25 European cities: results of the Aphekom project”. In: *Science of the Total Environment* 449 (2013), pp. 390–400.

[156] Roger D Peng et al. “Coarse particulate matter air pollution and hospital admissions for cardiovascular and respiratory diseases among Medicare patients”. In: *Jama* 299.18 (2008), pp. 2172–2179.

[157] M Hashem Pesaran and Allan Timmermann. “Testing dependence among serially correlated multicategory variables”. In: *Journal of the American Statistical Association* 104.485 (2009), pp. 325–337.

[158] A Peters et al. “Particulate air pollution is associated with an acute phase response in men. Results from the Monica–Augsburg Study”. In: *European heart journal* 22.14 (2001), pp. 1198–1204.

[159] José C Pinheiro and Douglas M Bates. “Linear mixed-effects models: basic concepts and examples”. In: *Mixed-effects models in S and S-Plus* (2000), pp. 3–56.

[160] C Arden Pope III and Douglas W Dockery. “Health effects of fine particulate air pollution: lines that connect”. In: *Journal of the air & waste management association* 56.6 (2006), pp. 709–742.

[161] C Arden Pope III et al. “Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution”. In: *Jama* 287.9 (2002), pp. 1132–1141.

[162] Melinda C Power et al. "Exposure to air pollution as a potential contributor to cognitive function, cognitive decline, brain imaging, and dementia: a systematic review of epidemiologic research". In: *Neurotoxicology* 56 (2016), pp. 235–253.

[163] Laura Prieto-Parra et al. "Air pollution, PM2. 5 composition, source factors, and respiratory symptoms in asthmatic and nonasthmatic children in Santiago, Chile". In: *Environment international* 101 (2017), pp. 190–200.

[164] Vivian Chit Pun et al. "Differential effects of source-specific particulate matter on emergency hospitalizations for ischemic heart disease in Hong Kong". In: *Environmental health perspectives* 122.4 (2014), pp. 391–396.

[165] Liping Qiao et al. "PM2. 5 constituents and hospital emergency-room visits in Shanghai, China". In: *Environmental science & technology* 48.17 (2014), pp. 10406–10414.

[166] C Radhakrishna Rao. "A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance". In: *Qüestiió: quaderns d'estadıstica i investigació operativa* 19.1 (1995).

[167] C Radhakrishna Rao. "Some statistical methods for comparison of growth curves". In: *Biometrics* 14.1 (1958), pp. 1–17.

[168] Peter M Robinson. "Generalized canonical analysis for time series". In: *Journal of Multivariate Analysis* 3.2 (1973), pp. 141–160.

[169] Alan C Rush, Joseph J Dougherty, and Jill A Engel-Cox. "Correlating seasonal averaged in situ monitoring of fine PM with satellite remote sensing data using geographic information system (GIS)". In: *Remote Sensing in Atmospheric Pollution Monitoring and Control*. Vol. 5547. International Society for Optics and Photonics. 2004, pp. 91–103.

[170] Jassim Al-Saadi et al. "Improving national air quality forecasts with satellite aerosol observations". In: *Bulletin of the American Meteorological Society* 86.9 (2005), pp. 1249–1262.

[171] Evangelia Samoli et al. “Differential health effects of short-term exposure to source-specific particles in London, UK”. In: *Environment international* 97 (2016), pp. 246–253.

[172] Hans Scheers et al. “Long-term exposure to particulate matter air pollution is a risk factor for stroke: meta-analytical evidence”. In: *Stroke* 46.11 (2015), pp. 3058–3066.

[173] Tamara Schikowski et al. “Ambient air pollution: a cause of COPD?” In: *European Respiratory Journal* 43.1 (2014), pp. 250–263.

[174] James J Schlesselman. *Case-control studies: design, conduct, analysis*. Oxford University Press, 1982.

[175] Steven K Schmidt, Stephen Simkins, and Martin Alexander. “Models for the kinetics of biodegradation of organic compounds not supporting growth.” In: *Applied and Environmental Microbiology* 50.2 (1985), pp. 323–331.

[176] L Sha et al. “The prevalence of asthma in children: a comparison between the year of 2010 and 2000 in urban China”. In: *Zhonghua jie he he hu xi za zhi= Zhonghua jiehe he huxi zazhi= Chinese journal of tuberculosis and respiratory diseases* 38.9 (2015), pp. 664–668.

[177] Takao Shohoji et al. “A prediction of individual growth of height according to an empirical Bayesian approach”. In: *Annals of the Institute of Statistical Mathematics* 43.4 (1991), pp. 607–619.

[178] Andrew J Simpkin et al. “Modelling height in adolescence: a comparison of methods for estimating the age at peak height velocity”. In: *Annals of human biology* 44.8 (2017), pp. 715–722.

[179] Taina Siponen et al. “Source-specific fine particulate air pollution and systemic inflammation in ischaemic heart disease patients”. In: *Occup Environ Med* 72.4 (2015), pp. 277–283.

[180] Richard L Smith et al. “Regression models for air pollution and daily mortality: analysis of data from Birmingham, Alabama”. In: *Environmetrics: The official journal of the International Environmetrics Society* 11.6 (2000), pp. 719–743.

[181] Massimo Stafoggia et al. “Long-term exposure to ambient air pollution and incidence of cerebrovascular events: results from 11 European cohorts within the ESCAPE project”. In: *Environmental health perspectives* 122.9 (2014), pp. 919–925.

[182] Slavica Stevanović and Dragana Nikić. “Exposure to air pollution and development of allergic rhinitis and asthma”. In: *Facta universitatis-series: Medicine and Biology* 13.2 (2006), pp. 114–118.

[183] K Subramanyam and M Bhaskara Rao. “Analysis of odds ratios in 2× n ordinal contingency tables”. In: *Journal of multivariate analysis* 27.2 (1988), pp. 478–493.

[184] R Core Team et al. “R: A language and environment for statistical computing”. In: (2013).

[185] R Core Team. *R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2014.*

[186] Yaohua Tian et al. “Fine particulate air pollution and first hospital admissions for ischemic stroke in Beijing, China”. In: *Scientific reports* 7.1 (2017), p. 3897.

[187] Bruce Urch et al. “Acute blood pressure responses in healthy adults during controlled air pollution exposures”. In: *Environmental health perspectives* 113.8 (2005), pp. 1052–1055.

[188] Kert Viele, Mark Lancaster, and Robin L Cooper. “Self-modeling structure of evoked postsynaptic potentials”. In: *Synapse* 60.1 (2006), pp. 32–44.

[189] Paul J Villeneuve et al. “Outdoor air pollution and emergency department visits for asthma among children and adults: a case-crossover study in northern Alberta, Canada”. In: *Environmental Health* 6.1 (2007), p. 40.

[190] Bailuis Walker and Charles P Mouton. “Environmental influences on cardiovascular health”. In: *Journal of the National Medical Association* 100.1 (2008), pp. 98–103.

[191] Jun Wang and Sundar A Christopher. “Intercomparison between satellite-derived aerosol optical thickness and PM2. 5 mass: Implications for air quality studies”. In: *Geophysical research letters* 30.21 (2003).

[192] Yi Wang, Melissa N Eliot, and Gregory A Wellenius. “Short-term changes in ambient particulate matter and risk of stroke: a systematic review and meta-analysis”. In: *Journal of the American heart association* 3.4 (2014), e000983.

[193] Yuedong Wang, Chunlei Ke, and Morton B Brown. “Shape-Invariant Modeling of Circadian Rhythms with Random Effects and Smoothing Spline ANOVA Decompositions”. In: *Biometrics* 59.4 (2003), pp. 804–812.

[194] Scott Weichenthal, Marianne Hatzopoulou, and Mark S Goldberg. “Exposure to traffic-related air pollution during physical activity and acute changes in blood pressure, autonomic and micro-vascular function in women: a cross-over study”. In: *Particle and fibre toxicology* 11.1 (2014), p. 70.

[195] Gregory A Wellenius et al. “Ambient air pollution and the risk of acute ischemic stroke”. In: *Archives of internal medicine* 172.3 (2012), pp. 229–234.

[196] Sten P Willemsen et al. “A multivariate Bayesian model for embryonic growth”. In: *Statistics in medicine* 34.8 (2015), pp. 1351–1365.

[197] John Wingerd. “The relation of growth from birth to 2 years to sex, parental size and other factors, using Rao’s method of the transformed time scale”. In: *Human biology* (1970), pp. 105–131.

[198] John Wishart. “Growth-rate determinations in nutrition studies with the bacon pig, and their analysis”. In: *Biometrika* 30.1/2 (1938), pp. 16–28.

[199] Wuxiang Xie et al. “Relationship between fine particulate air pollution and ischaemic heart disease morbidity and mortality”. In: *Heart* 101.4 (2015), pp. 257–263.

[200] Ayşe Canan Yazici et al. “An application of nonlinear canonical correlation analysis on medical data”. In: *Turkish Journal of Medical Sciences* 40.3 (2010), pp. 503–510.

[201] Xiaofang Ye et al. “Ambient temperature and morbidity: a review of epidemiological evidence”. In: *Environmental health perspectives* 120.1 (2011), pp. 19–28.

[202] Maayan Yitshak Sade et al. “Air pollution and ischemic stroke among young adults”. In: *Stroke* 46.12 (2015), pp. 3348–3353.

[203] Kuai Yu et al. “Association of solid fuel use with risk of cardiovascular and all-cause mortality in rural China”. In: *Jama* 319.13 (2018), pp. 1351–1361.

[204] Laura B Zahodne, Devangere P Devanand, and Yaakov Stern. “Coupled cognitive and functional change in Alzheimer’s disease and the influence of depressive symptoms”. In: *Journal of Alzheimer’s Disease* 34.4 (2013), pp. 851–860.

[205] Antonella Zanobetti et al. “Fine particulate air pollution and its components in association with cause-specific emergency admissions”. In: *Environmental Health* 8.1 (2009), p. 58.

[206] M Zelen. “The analysis of several 2× 2 contingency tables”. In: *Biometrika* 58.1 (1971), pp. 129–137.

[207] Hai Zhang, Raymond M Hoff, and Jill A Engel-Cox. “The relation between Moderate Resolution Imaging Spectroradiometer (MODIS) aerosol optical depth and PM2. 5 over the United States: a geographical comparison by US Environmental Protection Agency regions”. In: *Journal of the Air & Waste Management Association* 59.11 (2009), pp. 1358–1369.

[208] Yanwu Zhang et al. “The short-term association between air pollution and childhood asthma hospital admissions in urban areas of Hefei City in China: A time-series study”. In: *Environmental research* 169 (2019), pp. 510–516.

[209] Zhuohui Zhao et al. “Asthmatic symptoms among pupils in relation to winter indoor and outdoor air pollution in schools in Taiyuan, China”. In: *Environmental health perspectives* 116.1 (2007), pp. 90–97.

[210] Ke Zu et al. “Concentration-response of short-term ozone exposure and hospital admissions for asthma in Texas”. In: *Environment international* 104 (2017), pp. 139–145.

[211] MH Zwietering et al. “Modeling of the bacterial growth curve”. In: *Applied and environmental microbiology* 56.6 (1990), pp. 1875–1881.