Statistical Methods for Panel Studies with Applications in Environmental Epidemiology

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Biostatistics

have examined a dissertation entitled

"Statistical Methods for Panel Studies with Applications in Environmental Epidemiology"

presented by Alfa Ibrahim Mouke Yansane

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature Brent A Coull

Typed name: Prof. Brent Coull

Signature Paul Catalano

Typed name: Prof. Paul Catalano

Signature Diane Gold

Typed name: Prof. Diane Gold

Signature .....

Typed name:

Date: November 30, 2011

Statistical Methods for Panel Studies with Applications in Environmental Epidemiology

A thesis presented

by

Alfa Ibrahim Mouké Yansané

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

November, 2011

Thesis Advisor: Professor Brent A. Coull Alfa Ibrahim Mouké Yansané

Statistical Methods for Panel Studies with Applications in Environmental Epidemiology

Abstract

Pollution studies have sought to understand the relationships between adverse health effects and harmful exposures. Many environmental health studies are predicated on the idea that each exposure has both acute and long term health effects that need to be accurately mapped. Considerable work has been done linking air pollution to deleterious health outcomes but the underlying biological pathways and contributing sources remain difficult to identify. There are many statistical issues that arise in the exploration of these longitudinal study designs such as understanding pathways of effects, addressing missing data, and assessing the health effects of multipollutant mixtures. To this end this dissertation aims to address the afore mentioned statistical issues.

Our first contribution investigates the mechanistic pathways between air pollutants and measures of cardiac electrical instability. The methods from chapter 1 propose a path analysis that would allow for the estimation of health effects according to multiple paths using structural equation models. Our second contribution recognizes that panel studies suffer from attrition over time and the loss of data can affect the analysis. Methods from Chapter 2 extend current regression calibration approaches by imputing missing data through the use of moving averages and assumed correlation structures. Our last contribution explores the use of factor analysis and two-stage hierarchical regression which are two commonly used approaches in the analysis of multipollutant mixtures. The methods from Chap-

ter 3 attempt to compare the performance of these two existing methodologies for estimating health effects from multipollutant sources.

Contents

List of Figures

List of Tables

Acknowledgments

To begin, I would like to thank God for guiding my steps during the pursuit and completion of the doctoral program. At each celebration and each adversity, I have felt God's presence by always placing me in the right environment and among the best people.

There have been a number of key figures that have been instrumental in my development as a researcher whom I would like to acknowledge. I truly had a "dream team" of advisors. Through their concerted efforts and expert tutelage, I was able to learn and progress. I would like to recognize and thank my advisor, mentor, and professor, Brent Coull. Of his many genuine qualities I truly appreciated his warmth and ease. I always felt challenged but equally supported through course work , research, and the job search. He was always welcoming, available and understanding which were irreplaceable qualities. I would like to thank Dr. Diane Gold for all of her encouragement throughout my time in Boston. Her expertise helped me to realize the importance of the work and how it can be used to aid real people. To Dr. Paul Catalano, thank you for your ideas, endless energy, and optimism which helped me to refine my statistical thinking and lift my spirits. In addition to my committee, I was lucky enough to have two academic advisors Professor Michael Hughes and Professor Louise Ryan each of whom always showed the utmost confidence and faith in me.

Thank you to the invaluable staff in the Biostatistics department who kept me moving in the right direction. They were able to act as a surrogate family for me since I was far from home. Thank you to Aunt Jelena Follwieller, Aunt Vickie Beaulieu, Aunt Phoebe Hackett, and my sister Rachel Boschetto. Lastly, I would like to thank Sabrina Toomer for her generosity and love. She always treated me like family and took care of me when I needed it-thank you so much.

Thank you to my many friends who shared this educational endeavor and toast to our successes and continued life long friendship. Thank you Binta Beard for holding me down, your sacrifices and love will be valued forever. To Loni Phillip, Matt

Austin, Alane Izu, Shannon Stock, Roland Matsouaka, Alisa Stephens, Christina McIntosh, and Linda Valeri thank you so much for your friendship. Thank you to Ellen "Kittie" Richardson and Ronald "Kuda" Mills for your love and encouragement. I needed all of you in order to be successful so I am happy to share this with you.

Most importantly, I would like to thank my family. To my loving parents Dr. Aguibou Mouke Yansane and Maryam Cire Fofana, thank you for your love and support throughout my entire life. Each of you worked tirelessly to ensure that I would become a man who is loving and of good moral character. I hope that I have made you proud and been the blessing in your lives as you have been in mine. My sister, Kadidja Didi Mouke Yansane, I have always admired your courage, strength, and willingness to love. Thank you for always being there with your humor, kind words, advice, and love. Lastly, I dedicate this dissertation to the memory of my grandparents, Sekou Fofana and Aisha Cisse Fofana who passed away before the completion of my work. I was always in their thoughts and prayers, save a place in heaven.

Distributed Lag Path Analysis: Cardiovascular Effects of Ambient Air Pollution

[1]Alfa I. Yansané, [2,3]Diane R. Gold, [1,4]Paul J. Catalano,[1] Brent A. Coull

[1]Department of Biostatistics, Harvard School of Public Health
[2]Department of Environmental Health, Harvard School of Public Health
[3]Department of Medicine, Brigham and Women's Hospital/Harvard Medical School and
[4]Department of Biostatistics, Dana-Farber Cancer Institute

1.1 ABSTRACT

Epidemiological studies have consistently demonstrated that elevated levels of particulate matter (PM) are associated with increased mortality and morbidity. Further studies have demonstrated a consistent increased risk for cardiovascular events such as myocardial infarction, stroke, cardiac arrhythmia, atherosclerosis, and angina (Mittleman et al. 2000; Rich Q, 2005; Dockery et al., 2005; Berger et al., 2006). In spite of prior evidence linking air pollution to these adverse health outcomes, the underlying causal, physiological, and biological pathways are less understood. The purpose of this article is to model and identifying the mechanistic pathways of effects by conducting a path analysis within a structural equation framework. This approach corresponds to jointly fitting two generalized additive distributed lag health outcome models, such that inferences on the health effects can be determined through direct and indirect pathways. We compare the performance of our approach in estimating the health effects (changes in cardiovascular outcomes) to that of an existing approach of modeling the outcomes separately. Simulation results and subsequent data analysis suggest that the proposed distributed lag path analysis are effective in simultaneously estimating the health effects from direct and and indirect path while conventional methods can not. We employ the proposed methods in the analysis of an Exposure, Epidemiology, and Risk Program study that investigates the effects of particulate air pollution ($PM_{2.5}$) on ST-Segment depression, T-wave alternans (TWA), and heart rate variability (HRV).

1.2 INTRODUCTION

Epidemiological studies have consistently demonstrated that elevated levels of particulate matter (PM) are associated with increased mortality and morbidity. Further studies have demonstrated a consistent increased risk for cardiovascular events such as myocardial infarction, stroke, cardiac arrhythmia, atherosclerosis, and angina (Mittleman et al. 2000; Rich Q, 2005; Dockery et al., 2005; Berger et al., 2006). In spite of prior evidence linking air pollution to these adverse health outcomes, the underlying causal, physiological, and biological pathways are less understood. Identifying these mechanistic pathways will allow scientists, researchers, and medical professionals to become more informed and thus effectively focus medical interventions and treatments.

One of the primary objectives of PM research is the assessment of the health effects related to specific types of air pollution. Particulate matter, sulfur dioxide, oxides of nitrogen, carbon oxides, and ozone have each been shown to be both chronic and acute contributors to adverse effects on human health (Brook et al. 2004). The scientific interest of this paper is to explore the relationship between particulate air pollution and a measure of cardiac electrical instability, T-wave alternans (TWA). Further, this study seeks to explore whether pollution leads to TWA through causing autonomic dysfunction, measured as a reduction in heart rate variability (HRV). We hope to jointly model these phenomena to understand interrelationships between the separate cardiac outcomes so that the effects of exposure can be decomposed into direct and indirect effects (via other outcomes).

Investigations looking at the health effects of air pollution recognize that a health outcome can be affected by exposures experienced either at the time the outcome is measured or during some time previous to the health assessment. Accounting

for both contemporaneous and lagged effects would give a more well rounded assessment of pollution and help to avoid exposure misclassification biases. Some studies have shown that pollutant exposures measured at different lengths of time have will have a varied impact on the outcome (Chuang et al. 2008). Further, the relevant time windows may change depending on the outcome. Therefore, appropriate models must account for both immediate and lagged exposure effects, repeated measures, smoothed terms, and missing values. In this paper, we propose to develop methods that allow one to examine pathways of effects, when the lagged effects of exposure are potentially of interest. We plan to use distributed lag models merged within a structural equation framework to examine the relationship between air pollution and different electrical cardiac outcomes that are known precursors to cardiovascular events.

At present, existing analyses attempt to consider temporal resolution through the use of moving averages. The "moving average" method of analysis calculates exposure concentrations over various pre-specified intervals of time. Hence, each model produces one effect estimate for the respective moving average. In this modeling scheme the pollutant could be modeled as a linear or smoothed term depending the assumed relationship. It has been recognized in the literature that the effects of pollution are sensitive to the length of the moving averages used for exposure measures so effects may not be fully captured.

Another issue arises when attempting to consider multiple cardiac endpoints simultaneously, because different lags of exposure may be most relevant for the different outcomes. This means that each endpoint has its own pivotal time interval where the adverse health effects may be the highest in magnitude. If all of the models used the same time interval and resolution, it is possible that effects may

be seen in one outcome but not others. In this paper, we propose a path analysis that jointly fits two or more distributed lag models using the pollutants as exposures and the measures of cardiac electrical instability as outcomes. Modeling these outcomes jointly will allow for both direct and indirect effects to be estimated at varying time lags.

The data that motivates the proposed research comes from three analyses conducted through the Exposure, Epidemiology, and Risk Program in Boston on the effects of particulate air pollution (particulate matter, black carbon, carbon monoxide, ozone, nitrogen dioxide, and sulfur dioxide) on T-wave alternans and heart rate variability. Harvard researchers have conducted a number of regression analyses using moving averages of exposure and these outcomes. Pollutants were measured from a central site while the heart outcomes were calculated by a personal monitor at half hour intervals. There has been some exploration of potential biological pathways for this relationship such as; Direct paths through the cardiovascular system, blood, and lung receptors, or indirect paths through pulmonary oxidative stress and inammatory response (Brook et al. 2004). In order to explore the intermediate effects and their inter-relationship with other outcomes, a pathway model can be implemented and we will introduce and derive approaches for the implementation of such a model. Our proposed work seeks to help elucidate the electro-physiological mechanism to complement the existing research.

This paper is organized as follows: Section 1.3 describes in detail the design and data from a study evaluating the effects of particulate air pollution on electrical cardiac instability. Section 1.4 presents the distributed lag model and subsequent pathway model, while Section 1.5 discusses the direct and indirect effects of exposure. Section 1.6 gives a short treatment of the Bayesian approach to estimation

and Section 1.7 presents a simulation study to examine the effectiveness pathway analytic model, compared to the moving average approach. Section 1.8 demonstrates an application of the distributed lag pathway model (DLPWM) to analyze the afore mentioned study from Exposure, Epidemiology, and Risk Program and finally in Section 8 we discuss our findings along with implications for future path analyses.

1.3 DATA DESIGN

1.3.1 Data Collection

The study population consisted of a recruited panel of patients with documented coronary artery disease from the greater Boston area. Specifically, subjects were recruited within route 495 (the outer most boundary of the greater Boston metropolitan region) and a 40 km radius from the central pollution monitoring site. Each subject had experienced a percutaneous coronary intervention for an acute coronary syndrome or for worsening stable coronary artery disease. In each study, patients were excluded with atrial fibrillation and left bundle branch block (LBBB) because of the intent to evaluate heart rate variability and ST-Segment as outcomes. Further exclusions included patients who had bypass graft surgery within the last 3 months because accurate interpretations of the T-wave and ST-Segment would have been compromised. Other exclusions were active smokers, drug or alcohol abuse problems, and those with psychiatric illness. Subjects received a home visit within 2 to 4 weeks after the hospital discharge, followed by 3 additional visits at approximately 3 month intervals. There were 48 subjects yielding 129 person-visits with 6135 observations. Each patient had approximately 48 half-hour ST-segment,

T-wave alternans, and heart rate variability (HRV) measurements taken, which were linked with air pollution measurements at corresponding times.

The outcomes were measured using 24 hour 3 lead Holter ECG monitoring and the electrodes were placed in modified V5 and VF positions. In the subsequent visits, patients were given a follow-up questionnaire regarding cardiac and respiratory symptoms, and medication use. They later received 24-hour Holter monitoring. Ambient concentrations of particulate air matter with aerodynamic diameter less than $2.5\mu m$ $(PM_{2.5})$ and black carbon (BC) were measured at the central monitoring site located on the roof of Countway Library, Harvard Medical School, in downtown Boston, MA. $PM_{2.5}$ concentrations were measured using Tapered Element Oscillation Microbalance (TEOM, Model 1400A, Rupprecht and Pataschnick, Albany, NY). Ambient BC was measured using an aethalometer. $PM_{2.5}$ and BC concentrations were summarized in half hour intervals with analysis based on half hour, 12 hour lagged, and cumulative exposures. Indoor $PM_{2.5}$ and BC measurements were also taken. O_3, SO_2, and CO measurements were obtained using state monitoring sites in Boston, MA.

1.3.2 Single Outcome Analyses

A first analysis assessed the relationship between heart rate variability (HRV) and ambient air pollution among the post coronary event patients (Zanobetti et al. 2009). Authors explored this relationship because reduced HRV has been linked to increased risk of myocardial infarction, increased mortality in patients with heart failure, and is a marker for fatal ventricular arrhythmia (Gold et al. 2000; Task Force of the European Society of Cardiology the North American Society of Pacing Electro-physiology, 1996). HRV was measured using four different metrics;

standard deviation of normal-to-normal heart beat intervals $(SDNN)$ and square root of the mean of the squared differences between adjacent normal RR intervals (r-MSSD), high frequency (HF), and total power (TP). The smaller the standard deviation in the RR intervals corresponded with lower HRV measures. The authors used generalized additive models to control for confounding, which allowed for the covariates to have non-linear effects on outcome. For both r-MSSD and HF, the authors found significant negative associations with $PM_{2.5}$ and BC. There was a tendency for the stronger r-MSSD associations to occur at longer averaging times.

The second analysis of this study was to explore the relationship between particulate pollution and T-wave alternans (Zanobetti et al. 2009). T-wave alternans (TWA) are periodic beat to beat variations in the amplitude of the T-wave in an electrocardiogram (ECG). It is most often measured in patients who have had myocardial infarctions or other heart damage to see if they are at high risk of developing a potentially lethal cardiac arrhythmia. The shape of the T-wave could be a key indicator of cardiac health and mortality (Nieminen et al. 2007; Stein et al. 2008). For example, inverted or negative T-waves can be a sign of coronary ischemia, whereas tall or tented symmetrical T-waves may indicate hyperkalemia. TWA is also a marker of cardiac electrical instability measured as differences in the magnitudes between adjacent waves. Increases in the previous 1 to 12 hour averaged ambient $PM_{2.5}$ and BC were associated with increases in TWA, with peak cumulative effects in between 6 and 12 hours. The authors' estimated that for a 1 unit increase in 6 hour averaged $PM_{2.5}$ there was an increase of $1.7\%(0.6, 2.7)$in TWA.

1.4 MODEL AND NOTATION

1.4.1 Modeling Framework

An alternative to separate "moving average" models is the distributed lag model (DLM). Distributed lag models generalize the single time point or moving average models because they estimate differential air pollution effects for all lagged time points simultaneously, rather than from separate models. Our data have been collected so that measurements for each pollutant have been collected for 48 separate half-hour time lags along with the corresponding electrical cardiac instability outcomes at those times. Therefore the data are suited for distributed lag modeling framework.

We begin with a generalized additive distributed lag model that adjust for lagged exposures, linear and non-linear effects of confounders, and random subject effects,

$$Y_{it} = \eta_0 + \sum_{l=0}^{q} \beta_l x_{i,t-l} + \sum_{j=1}^{d} f_j(s_{itj}) + \boldsymbol{\gamma}^T x_{it,linear} + U_i + \epsilon_{it} \tag{1.1}$$

where q is the the number of lagged time points. X_{it} is the half-hour pollution measure of subject i at time t ($PM_{2.5}$) (Zanobetti et al., 2000). Y_{it} is the outcome measure of subject i at time t. The $\boldsymbol{\epsilon}_{it}$ is the error of subject i at time t and is normally distributed with zero mean and constant variance $\boldsymbol{\sigma}_{\epsilon}^2$. The U_i is the random coefficient due to subject i also with mean zero and variance $\boldsymbol{\sigma}_{u}^2$. The vector $\mathbf{x_{it,linear}}$ is a vector of variables modeled linearly and $\boldsymbol{\gamma}$ represent the effect estimates. The overall impact of a unit change in in exposure over q days is given by $\Sigma_{t=0}^{q}\beta_l$ (Schwartz et al., 2000). Due to collinearity, it is necessary to constrain the β_l to be a polynomial

or spline function of l. For each β_l there were three different options utilized. They will be represented by the following equations.

Option 1(Parametric):

$$\beta_l = \sum_{r=1}^{p} \tau_r l^r$$

where $0 \leq l \leq q$

Option 2 (Thin-Plate Spline)(Crainiceanu et al., 2005):

$$\beta_l = \sum_{r=1}^{2} \tau_r l^r + \sum_{k=1}^{K} \nu_k |l - \kappa_k|^3$$

where $0 \leq l \leq q$ and

Option 3 (Truncated Spline):

$$\beta_l = \sum_{r=1}^{p} \tau_r l^r + \sum_{k=1}^{K} \nu_k (l - \kappa_k)_+^p$$

where $0 \leq l \leq q$ and

$$(l - \kappa_k)_+^p = \begin{cases} (l - \kappa_k)^p & if \quad l \geq \kappa_k \\ 0 & if \quad l < \kappa_k \end{cases},$$

where $\kappa_1, \ldots, \kappa_K$ is a set of K distinct numbers between 0 and q. β_l is a piecewise pth degree polynomial in l, with join points (knots) at the κ_k. The ν_k are coefficients associated with the basis function $(l - \kappa_k)_+^p$.

Each smooth function f_j can be estimated using a penalized spline of degree p

(Carroll et al., 2003):

$$f_j(s_{itj}) = \sum_{c=1}^{p} \alpha_{j,c} s_{itj}^c + \sum_{k=1}^{K_j} \omega_{j,k} (s_{itj} - \kappa_{j,k})_+^p$$

The s_{ijt} is the confounder variable for the i^{th} subject, the j^{th} variable modeled as a smooth function at time t. The $\alpha_{j,c}$ is the coefficient for j^{th} smoothed term. While the $\omega_{j,k}$ are the coefficients corresponding to the basis function $(s_{itj} - \kappa_{j,k})_+^p$ for the j^{th} smoothed variable. Each smoothed term can be expressed in the form of a linear mixed model with both fixed and random terms.

1.4.2 Distributional Assumptions

Model (1.1) can be simplified through matrix representations below:

$$\mathbf{Y} = \mathbf{X_{Lag}}\boldsymbol{\tau} + \mathbf{Z_{Lag}}\mathbf{u} + \textstyle\sum_{j=1}^{d} \mathbf{X_{smooth,j}}\boldsymbol{\alpha}_j + \sum_{j=1}^{d} \mathbf{Z_{smooth}}\mathbf{w_j} + \mathbf{X_{Linear}}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

$$\boldsymbol{\tau} = [\eta_0, \tau_0, \tau_1, \ldots, \tau_p]^T$$
$$\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \ldots, \alpha_p]^T$$
$$\mathbf{u} = [U_1, U_2, \ldots, U_m]^T$$
$$\boldsymbol{\nu} = [\nu_1, \nu_2, \ldots, \nu_K]^T$$
$$\mathbf{w}_j = \left[\omega_1, \omega_2, \ldots, \omega_{K_j}\right]^T$$
$$\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_b]^T$$

Let p represent the the polynomial degree, let m represent the total number

of distinct subjects such that $1 \leq i \leq m.$, and let n_i represent the number of observations at time t. By concatenating the model further and using the fact that the spline coefficients can be modeled as random effects, the above equation can be reduced to the simple mixed model (1.2) in the following form:

$$\mathbf{Y_1} = \left[\begin{array}{c|c|c} \mathbf{X_{Lag}} & \mathbf{X_{Smooth}} & \mathbf{X_{Linear}} \end{array} \right] \left[\begin{array}{c} \boldsymbol{\tau} \\ \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \end{array} \right] + \left[\begin{array}{c|c} \mathbf{Z_{Lag}} & \mathbf{Z_{Smooth}} \end{array} \right] \left[\begin{array}{c} \mathbf{u} \\ \\ \mathbf{w} \end{array} \right] + \boldsymbol{\epsilon} \quad (1.2)$$

$$\mathbf{Y_1} = \mathbf{X}\boldsymbol{\beta}_1 + \mathbf{Z}\mathbf{b_1} + \boldsymbol{\epsilon} \quad (1.3)$$

$$\mathbf{Cov} \left[\begin{array}{c} \mathbf{b} \\ \boldsymbol{\epsilon} \end{array} \right] = \left[\begin{array}{cccc} \sigma^2_{\mathbf{u}}\mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{\nu}\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma^2_{\omega,\mathbf{j}}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \sigma^2_{\epsilon}\mathbf{I} \end{array} \right]$$

The form of this model could be applied to other outcomes, for example ST segment. The model would be analogous to the one above including the same confounders but with a different outcome.

$$\mathbf{Y_2} = \mathbf{X}\boldsymbol{\beta_2} + \mathbf{Z}\mathbf{b_2} + \boldsymbol{\epsilon}, \quad (1.4)$$

1.4.3 The Pathway Analytic Model

Path analysis can be used to test theoretical models that specify causal relationships between a number of observed variables (Hatcher, 1994). Structural equa-

tion models (SEM's) are a set of flexible models that enable the modeling of multivariate data for path analyses. SEM's can handle both simple and hierarchical modeling structures (Sanchez et al., 2005). An essential tool for SEM's is the path diagram or directed acyclical graph (DAG) that details causal relationships graphically. Each variable is represented by its own box. Single-headed arrows represent causal relationships between two different variables. In Figure 1.1, the x variable represents the independent variable or antecedent variable, predicted to precede and have a causal effect on y. The y box represents the consequent variable or the dependent variable. The straight, single-headed arrow is generally used to represent a directional causal path in a path diagram while also detailing the statistical model that describes the relationship. The z variable can be considered an intermediate (mediator) variable because it is on the causal pathway from x to y and it is caused by x.

Through normal likelihood theory, estimation of parameters, confidence intervals, and p-values can be calculated. Standard approaches to pathway analysis usually make the assumption that the variables of interest are normally distributed. Subsequently, direct paths have point and interval estimates while indirect paths are the product of the estimate of the independent variable to the intermediate variable and that for the association between estimate of the intermediate variable to the dependent variable. In the normal theory case, these parameters could be estimated using to least squares equations. The following example is a simple illustration of the modeling scheme.

Let y be the outcome variable, x be the independent variable, and z be the intermediate variable whereby θ_2 denotes the linear association between x and y, θ_1 denotes the linear association between x and z, and θ_3 the linear association

between z and y. The DAG for the model is as follows:

$$z = \theta_0 + x\theta_1 + e_z \tag{1.5}$$

$$y = \theta_{00} + x\theta_2 + z\theta_3 + e_y \tag{1.6}$$

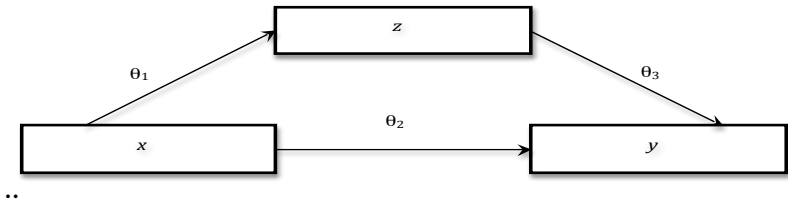


Figure 1.1: General Form

By substituting the value z from equation (1.5) into equation (1.6) we have the resulting equation that allows one to estimate $\theta_1\theta_3$ and θ_2.

$$y = \theta_{00} + \theta_0\theta_3 + x\theta_2 + x\theta_1\theta_3 + e_z\theta_3 + e_y \tag{1.7}$$

This method is effective when the variables are normally distributed (Gajewski et al., 2006). There is also the question of calculating the appropriate standard errors using this method because the standard errors for θ_1 and θ_3 are correlated. Therefore, there is a natural congruence between the directed acyclical graph (DAG) and its corresponding model.

Extending this basic conceptual structure using generalized distributed lag models we present Figure 1.2 below as a potential directed acyclical graph (DAG). We

propose that the GADLMs below are an accurate reflection of the DAG and will be able to estimate the direct effects of pollutant on T-wave alternans as well as the indirect effects through the intermediary HRV. Equation (1.8) represents a model that estimates the direct effects of the exposure (x_{it}) on outcome 1 $(Y_{1,it})$ where the $\beta_{1,l}$ are the parameters of interest. While model (1.9) estimates the effect of exposure (x_{it}) on the second outcome $(Y_{2,it})$ through the intermediary $(Y_{1,it})$. The $\beta_{2,l}$ are the distributed lag function for the direct effects between T-wave alternans and the pollutant. The φ_l represents the coefficient of the intermediate outcome $(Y_{1,it})$. Hence, through the SEM framework model (1.9) accounts for multiple endpoints on the causal pathway and yields interpretable direct and indirect effects.

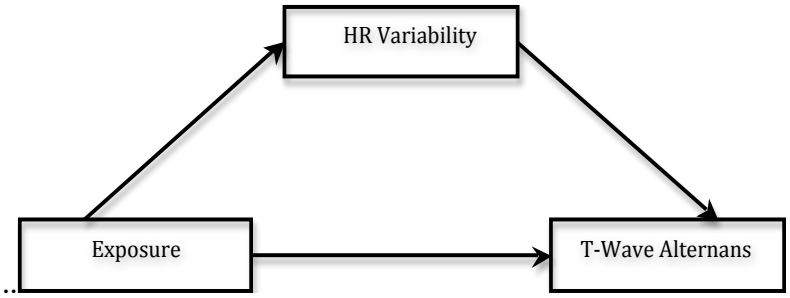


Figure 1.2: DAG for Air pollutant exposure

$$Y_{1,it} = \eta_{1,0} + \sum_{l=0}^{q} \beta_{1,l} x_{i,t-l} + \sum_{j=1}^{d} f_j(s_{itj}) + \boldsymbol{\gamma}_1^T w_{it} + U_{1,i} + \epsilon_{1,it}, \tag{1.8}$$

$$Y_{2,it} = \eta_{2,0} + \sum_{l=0}^{q} \beta_{2,l} x_{i,t-l} + \sum_{l=0}^{q} \varphi_l Y_{1,i,t-l} + \sum_{j=1}^{d} g_j(s_{itj}) + \boldsymbol{\gamma}_2^T w_{1,it} + U_{2,i} + \epsilon_{2,it}, \tag{1.9}$$

1.5 DIRECT AND INDIRECT EFFECTS

1.5.1 Interpretation

In a broad sense, the relationship between an exposure of interest and an outcome can be singular or multifactorial. We are interested in quantifying the relationship through detailing the magnitude, direction, and causal pathway. A direct effect is defined as a link between an exposure and outcome. Given the previous DAGs, the direct effects are represented by a single arrow with no intermediaries. On the other hand an indirect effect is the link between an exposure and outcome that consist of intermediaries on the pathway. Therefore, more than one arrow is needed to describe the relationship. This is significant because researchers will be able to explore whether the effects of a pollutant can be seen directly or indirectly through an intermediary. This will illuminate many questions regarding the electrophysiological pathway between pollutants and electrical cardiac outcomes as well as test for interrelationships. For example, in our current data set, HRV precedes TWA on the electro-physiological pathway and researchers would like to investigate the scientific trail where the pollutants are the most influential.

Using figure 1.2 as the model DAG, we estimate the direct effect between outcome (TWA) and the exposure (pollutant) with a set of parameters $\beta_{2,l}$ represented by a smoothed curve, and the indirect effects between the outcome variable and the exposure through the mediating variable is represented by another curve proven below:

$$Y_{1,it} = \eta_{1,0} + \sum_{l'=0}^{q'} \beta_{1,l'} x_{i,t-l'} + \sum_{j=1}^{d} f_j(s_{itj}) + \boldsymbol{\gamma}_1^T w_{it} + U_{1,i} + \epsilon_{1,it}, \tag{1.10}$$

$$Y_{2,it} = \eta_{2,0} + \sum_{l=0}^{q} \beta_{2,l} x_{i,t-l} + \sum_{l=0}^{q} \varphi_l Y_{1,i,t-l} + \sum_{j=1}^{d} g_j(s_{itj}) + \boldsymbol{\gamma}_2^T w_{1,it} + U_{2,i} + \epsilon_{2,it} \quad (1.11)$$

Now we substitute equation (1.10) into equation (1.11) and rearrange the terms.

$$\begin{aligned} Y_{2,it} &= \eta_{2,0} + \sum_{l=0}^{q} \beta_{2,l} x_{i,t-l} + \sum_{l=0}^{q} \varphi_l \left[\eta_{1,0} + \sum_{l'=0}^{q'} \beta_{1,l'} x_{i,t-l'-l} + \cdots + \epsilon_{1,i,t-l}, \right] \\ &+ \sum_{j=1}^{d} g_j(s_{itj}) + \boldsymbol{\gamma}_2^T w_{1,it} + U_{2,i} + \epsilon_{2,it} \end{aligned}$$

Since we are only interested in the direct and indirect effects of exposure and their interpretations we will use the following as our model of interest where $[**]$ represents the other terms/confounders in the model after some.

$$\begin{aligned} Y_{2,it} &= \left[\sum_{l=0}^{q} \beta_{2,l} x_{i,t-l} + \sum_{l=0}^{q} \sum_{l'=0}^{q'} \varphi_l \beta_{1,l'} x_{i,t-l'-l} \right] + [**]. \\ &= \left[\sum_{l=0}^{q} \beta_{2,l} x_{i,t-l} + \sum_{k=0}^{q+q'} \beta_k^* x_{i,t-k} \right] + [**] \end{aligned}$$

The $\beta_{2,l}$ parameters represent the set of direct effects of lagged exposure on our outcome and

where $\beta_k^* = \sum_{l+l'=k} [\varphi_l \beta_{1l'}]$ represents the indirect effect of the lagged exposure on outcome.

1.6 ESTIMATION

Standard distributed lag models (DLM's) could be fit using maximum likelihood methods by including all covariates in a generalized linear-mixed model. ML methods require large sample sizes for asymptotic optimality of the resulting ML estimators. Since we have 48 lagged exposures to be included in the model, these methods may not be optimal. Further, the number of parameter estimates increases when conducting the pathway analyses to account for the new effects.

We propose a non-informative Bayesian approach to modeling these distributed lag data. We wish to estimate the effect of $PM_{2.5}$ on two separate cardiovascular outcomes simultaneously. The estimates will be represented by $\boldsymbol{\tau}_r$ for $r = 1, \ldots, p$. Non-informative priors are proposed for each parameter so that the estimates are primarily data driven. Each τ is distributed as follows:

$$\boldsymbol{\tau} \sim N(0, \boldsymbol{\Psi})$$

Where $\boldsymbol{\Psi} = \mathbf{10,000}$. An informative approach could have been done as well but more data from previous studies would have been needed.

1.7 SIMULATION STUDY

We conducted a simulation study to examine the effectiveness of DLM pathway model to estimate the changes in TWA and HRV. We would also like to perform a direct comparison of estimates between a moving average, parametric, and semi parametric approaches. The intended outcomes, the lagged exposure, and the co-

efficients must be simulated under varying assumptions in order to get a complete picture of the model effectiveness while decomposing the exposure-outcome relationship. We began with 50 subjects yielding 100 person-visits with 5000 total observations. There were also 50 measurements taken for each of the 100 subjects and 50 lagged time-points created to mimic those in the real data.

1.7.1 Simulating Lagged Data

To simulate exposure, we generated 50 $PM_{2.5}$ exposure variables lagged by half hour intervals from $\mathbf{X} \sim N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_\mathbf{x})$. We assumed,

$$\boldsymbol{\Sigma}_\mathbf{x} = \begin{bmatrix} \boldsymbol{\sigma}_\mathbf{x}^\mathbf{2} & \rho & \rho^\mathbf{2} & \dots & \rho^\mathbf{49} \\ \rho & \boldsymbol{\sigma}_\mathbf{x}^\mathbf{2} & \rho & \dots & \rho^\mathbf{48} \\ \rho^\mathbf{2} & \rho & \boldsymbol{\sigma}_\mathbf{x}^\mathbf{2} & \dots & \rho^\mathbf{47} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^\mathbf{49} & \rho^\mathbf{48} & \rho^\mathbf{47} & \dots & \boldsymbol{\sigma}_\mathbf{x}^\mathbf{2} \end{bmatrix},$$

where $\rho = 0.4$ to reflect the fact that $PM_{2.5}$ measurements taken at closer intervals are more highly correlated. $\boldsymbol{\sigma}_{\boldsymbol{x}}^\mathbf{2} = 1$ and $\boldsymbol{\mu}_x = 5$ were taken from averages in the real data.

1.7.2 Simulating Outcome 1

Next, we picked initial values under varied conditions for τ that were also chosen based estimates from the models of the real data. Each β_l was then calculated using the afore mentioned formula: $\beta_l = \sum_{r=1}^{p} \tau_r l^r$ where p= 3 representing a cubic

polynomial spline. These values were needed to simulate the direct effect between Y_1(HRV) and $\mathbf{X}(PM_{2.5})$. Y_1 was simulated from the following distribution:

$$Y_1|X, \beta_1 \sim N(\sum_{l'=0}^{q'} \beta_{1,l'} x_{i,t-l'}, \sigma^2_{y_1})$$

1.7.3 Simulating Outcome 2

Our next task was to simulate Y_2 which reflects the indirect effect $PM_{2.5}$ and TWA. In addition to choosing initial values for τ_2, initial values for φ were designated as they represent the intermediary effects between HRV and TWA. Since HRV only acts on TWA for a short period of time, the lagged relationship between HRV and TWA only spanned 2 time points. Although for thoroughness in understanding we conducted simulation with 6 intermediary time points. The intermediary time point values denoted by the variable φ_l and given by the function, $\varphi[i] = 0.1 - .001i^2$ for $i = 1, \ldots, 6$. Y_2 was then simulated from the following distribution:

$$Y_2|Y_1, \beta_2 \sim N(\sum_{l=0}^{q} \beta_{2,l} x_{i,t-l} + \sum_{l=0}^{q} \varphi_l Y_{1,i,t-l}, \sigma^2_{y_2})$$

1.7.4 Simulation Path Model

Table 1.1 gives the assumed true values used for each simulated scenario. They represent distinct, biologically plausible combinations of distributed lag functions for HRV and TWA. Using the values of $\boldsymbol{\tau}$ in table 1.1, the corresponding health effect estimates β_l are calculated. Next, the X, Y_1, and Y_2) values are generated for 100 data sets on which the distributed lag pathway model was conducted. Each

path model only included lagged exposure variables and lagged Y_2 values. The form of the simulated path model is as follows:

$$Y_{1,it} = \eta_{1,0} + \sum_{l'=0}^{q'} \beta_{1,l'} x_{i,t-l'} + U_{1,i} + \epsilon_{1,it}, \tag{1.12}$$

$$Y_{2,it} = \eta_{2,0} + \sum_{l=0}^{q} \beta_{2,l} x_{i,t-l} + U_{2,i} + \epsilon_{2,it} \tag{1.13}$$

where $\beta_l = \sum_{r=1}^{p} \tau_r l^r$. Both the parameters and hyper parameters were given the following non-informative priors.

$$\tau \sim N(0, 100000)$$

$$U_i \sim N(0, \sigma_u^2)$$

$$\sigma_u^2 \sim IG(0.01, 0.01)$$

$$\sigma_y^2 \sim IG(0.01, 0.01)$$

For each data set, the posterior distribution is estimated using MCMC methods using 10,000 iterations and subsequently keeping 1,000 posterior values. The mixing of each model is checked visually by the trace plots to see if convergence was achieved. Next, the median and point-wise 95% credible interval of each lagged time point are calculated for each of the 100 sets of posterior effect estimates.

1.7.5 Simulation Results

Each simulated pathway model will be designated by 4 output graphs where the X-axis represents the lagged time points and the Y-axis is the magnitude of the posterior effect estimate. Each plot includes a true (red curve) and estimated (black curve) distributed lag function and 95% confidence bands (black) for the relationship between: 1) PM and HRV $(\beta_{1,l})$, 2) PM and TWA $(\beta_{2,l})$, 3) HRV and TWA (φ_l), and 4) Indirect Effect of PM on TWA through HRV(β_l^*). The blue curve in the 3rd position plot represents a reduced setting on the magnitude of the relationship between HRV and TWA. The blue curve in the 4th position gives the indirect association of $PM_{2.5}$ on TWA through HRV for the afore mentioned reduced setting.

Figure 1.3 represents the first simulation setting model due to its biological pattern. In the literature we see that HRV has a negative relationship with $PM_{2.5}$ and TWA has a positive relationship with $PM_{2.5}$. For this simulation scenario we note that the estimated distributed lag function quite accurately estimates the true DLF and is also within 95% credible limits in all 4 graphs.

The pathway model for the first simulation shows that the relationship between $PM_{2.5}$ and HRV is negative. Early lags reflect the highest effects while later lags move towards 0. The relationship between $PM_{2.5}$ and TWA is positive with most of the effect occurring at earlier lags as well. The distributed lag function for the indirect effects (β_l^*) are almost identical to the distributed lag function between $PM_{2.5}$ and HRV except for the first 12 time points. As we would expect, the indirect effect depends on the magnitude of the distributed lag function for the relationship between HRV and TWA given by φ_l, and our procedure is able to appropriately separate out these direct and indirect effects. Other simulation scenarios were completed and the conclusions were similar.

Model	Initial-τ_0	Initial-τ_1	Initial-τ_1	Initial-τ_3
Model 1:				
HRV(τ_1)	-0.01	0.0011	-0.000041	0.00000043
TWA(τ_2)	0.01	-0.0011	0.00004	-0.00000043
Model 2:				
HRV(τ_1)	0.00	0.00	0.00	0.00
TWA(τ_2)	0.01	-0.0011	0.00004	-0.00000043
Model 3:				
HRV(τ_1)	-0.01	0.0011	-0.000041	0.00000043
TWA(τ_2)	0.0	0.00	0.00	0.00
Model 4:				
HRV(τ_1)	-0.01	0.0011	-0.000041	0.00000043
TWA(τ_2)	0.001	-0.0005	0.00006	-0.000001
Model 5:				
HRV(τ_1)	0.01	-0.0005	0.00006	-0.000001
TWA(τ_2)	0.01	-0.0011	0.00004	-0.00000043
Model 6:				
HRV(τ_1)	0.01	-0.00055	0.000041	0.000001
TWA(τ_2)	0.01	-0.0011	0.00004	-0.00000043
Model 7:				
HRV(τ_1)	-0.001	0.00015	-0.000005	0.0000002
TWA(τ_2)	0.01	-0.0011	0.00004	-0.00000043
Model 8:				
HRV(τ_1)	0.01	0.0011	0.000041	0.00000043
TWA(τ_2)	0.01	-0.0011	0.00004	-0.00000043

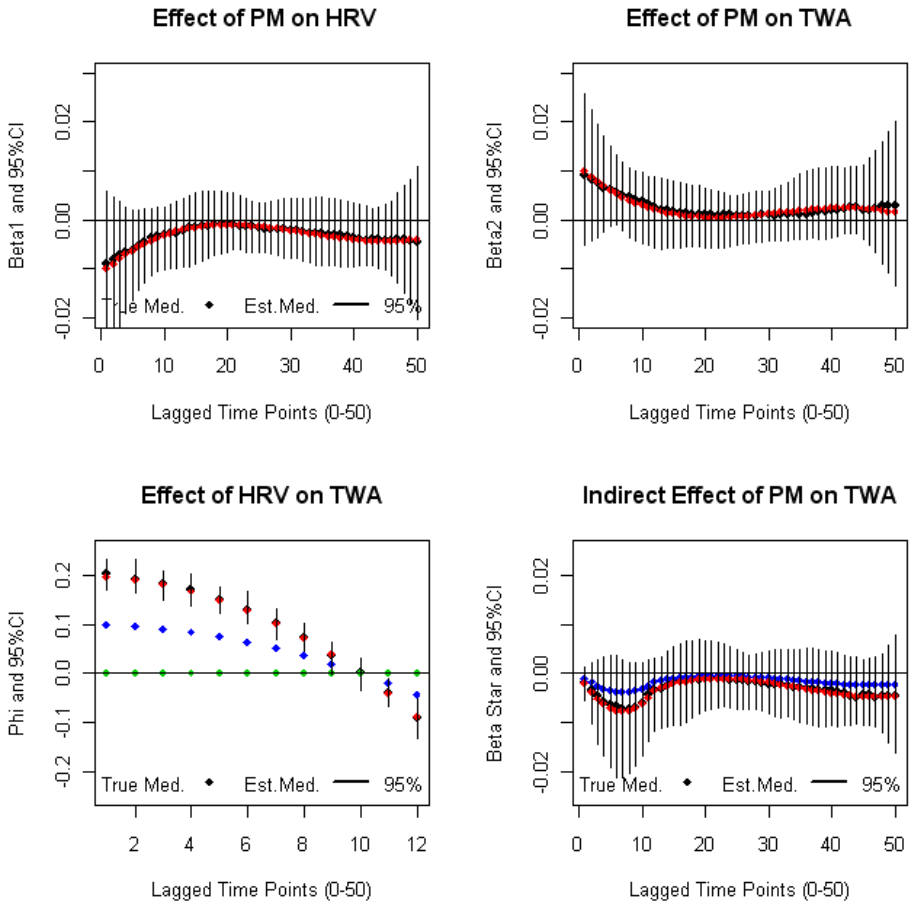
Table 1.1: This table represents the initial values chosen for the simulation study. For each model, 4 initial values were chosen for both HRV and TWA. Each model corresponds to a distinct plausible distributed lag function.

1.8 DATA ANALYSIS

1.8.1 Prior Elicitation

Given the motivating heart data, DLMM models were fit using the same confounders as the initial analysis described in Section 1.2. The following model was used for this analysis:

$$Y_{1,it} = \eta_{1,0} + \sum_{l'=0}^{q'} \beta_{1,l'} x_{i,t-l'} + \boldsymbol{\gamma}_1^T w_{it} + U_{1,i} + \epsilon_{1,it}, \tag{1.14}$$



..

Figure 1.3: Model 1-Regular HRV; Regular TWA

where:

$$\beta_{1,l} = \sum_{r=1}^{p} \tau_r l^r$$

where $0 \leq l \leq q$

We used the above parametric parameterization of $\beta_{1,l}$ for p=2, 3, and 4 which corresponds to quadratic, cubic, and quartic single pollutant models. We also attempted to use the "Thin-plate spline" and "Truncated spline" parameterizations for $\beta_{1,l}$ but the results were less stable. Each model adjusted for subject and day

of the week as indicator variables while average heart rate, mean temperature, hour of the day using quadratic effects. Quadratic effects were chosen for these variables because univariate generalized additive models were run and quadratic effects seemed to be a plausible approximation. Lastly, numerical date was controlled for using the following parameterization:

date1 = $\frac{sin(2*\pi*date)}{T}$, date2 = $\frac{cos(2*\pi*date)}{T}$

To conduct this Bayesian analysis we utilized Markov Chain Monte Carlo methods MCMC to estimate the parameters for the DLM using the R statistical package (The Comprehensive R Archive Network: http://cran.r-roject.org/). The "R2Winbugs" function was used so that Winbugs could be accessed within the R platform (Crainiceanu et al., 2005). Since our model was fit in the Bayesian setting, we assigned non-informative priors for the model parameters.

$$\tau \sim N(0, 100000)$$

$$u_i \sim N(0, \sigma_u^2)$$

$$\omega_j \sim N(0, \sigma_{\omega_j}^2)$$

$$\sigma_u^2 \sim IG(0.01, 0.01)$$

$$\sigma_{\omega_j}^2 \sim IG(0.01, 0.01)$$

$$\sigma_y^2 \sim IG(0.01, 0.01)$$

Each model was run using a burn-in period of 20,000 iterations. The convergence of each estimated parameter was checked by visual inspection of the trace plots. We kept 5,000 posterior samples thinned by 5 for each of the 49 lagged estimates.

The total effect of a particular pollutant over q hours was calculated by summing over the lagged coefficients given by the posterior samples. This yielded 5,000 posterior samples of the total pollutant effect over 24 hours so that medians and confidence intervals could be produced.

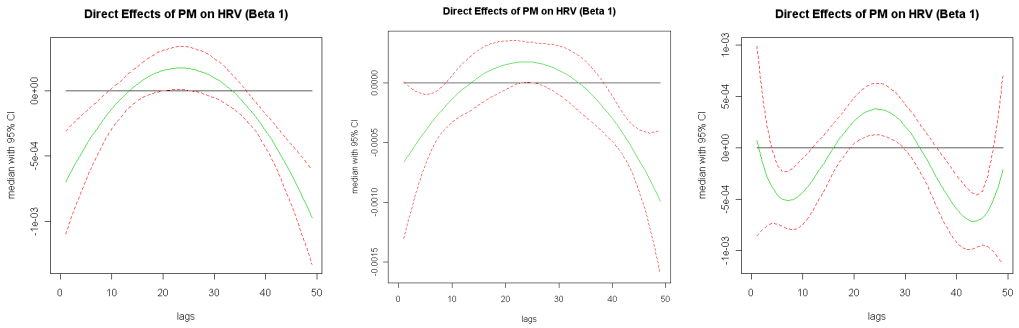
1.8.2 Health effects analysis

Heart Rate Variability

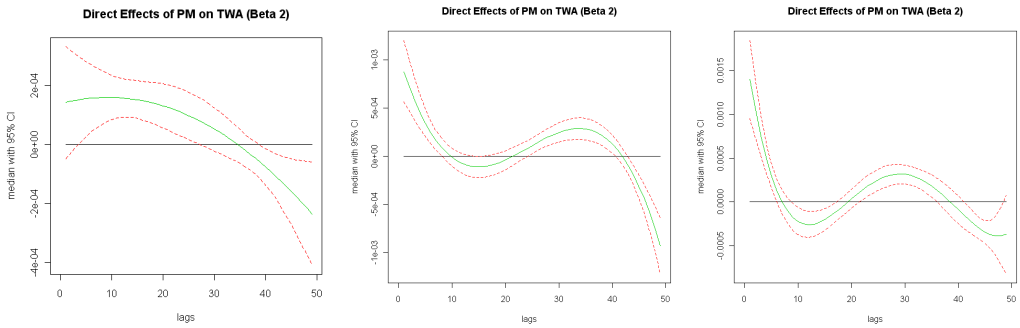
Figure 1.4 is a plot of the distributed lag function for the relationship between HRV (measured as r-MSSD) and $PM_{2.5}$. Each graph represents a quadratic, cubic, and quartic parameterization of this relationship. Subject, day of the week, average heart rate, mean temperature, hour of the day, and were the included confounders. Figure 4 reveals that the effect of $PM_{2.5}$ on heart rate variability has a curvilinear shape that is concave and the effect is mostly negative across the three model versions. The overall impact of a unit change in $PM_{2.5}$ over 48 lags (24 hours) was associated with a -0.00872 (-0.013, -0.0049) reduction in HRV for the quadratic model, -0.0085 (-0.0124, -0.0046) reduction in HRV for the cubic model, and -0.0084 (-0.0124, -0.0042) reduction in HRV for the quartic model. This shows consistency across the parameterization. These results are consistent with the moving average results of Zanobetti et al. 2000.

T-wave alternans

Figure 1.4 also contains plots of the distributed lag function for the relationship between TWA and $PM_{2.5}$. Each plot represents a quadratic, cubic, and quartic



(a) Quadratic Model (HRV) (b) Cubic Model (HRV) (c) Quartic Model (HRV)



(d) Quadratic Model (TWA) (e) Cubic Model (TWA) (f) Quartic Model (TWA)

Figure 1.4: Top Row: SINGLE OUTCOME MODEL HRV - Each graph represents the DL function of the relationship between $PM_{2.5}$ and HRV adjusting for subject, day of week, average heart rate, mean temperature, hour of the day, and date. Bottom Row: SINGLE OUTCOME MODEL TWA - Each graph represents the DL function of the relationship between $PM_{2.5}$ and TWA adjusting for subject, day of week, average heart rate, mean temperature, hour of the day, and date.

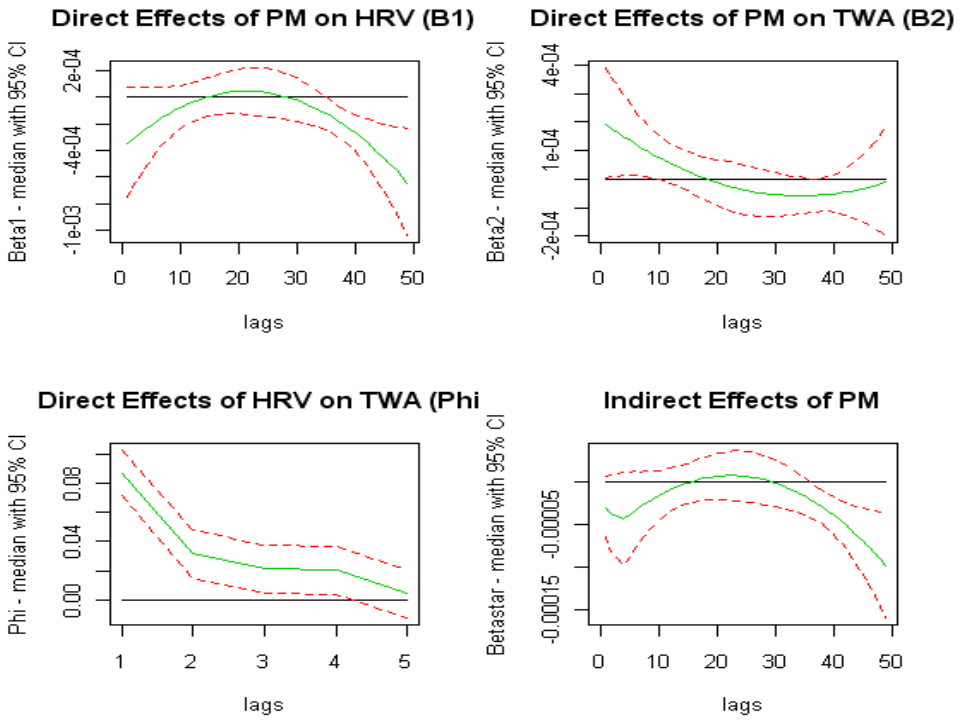
parameterization of this relationship. Figure 5 reveals that the effect of $PM_{2.5}$ on TWA has a curvilinear shape that begins with a highly positive effect for early lags and approaches zero for later lags. The overall impact of a unit change in $PM_{2.5}$ over 48 lags (24 hours) was associated with a 0.0023 (0.00042, 0.00421) increase in TWA for the quadratic model, 0.0030 (0.0012, 0.0050) increase in TWA for the cubic model, and 0.0032 (0.0013, 0.0051) increase in TWA for the quartic model. Each of the estimates show a significant and positive relationship between $PM_{2.5}$ and TWA-. Zanobetti et al., 2009 shows that with increasing moving averages, there is an increase in the TWA which means that the two approaches are in accord (See Appendix Figure 8).

1.8.3 Path Analysis

The path analyses in Figures 1.5, 1.6, and 1.7 reflect the real data from the Exposure, Epidemiology, and Risk Program in Boston. The outcomes of specific interest are HRV (measured through r-MSSD) , TWA, and the exposure is $PM_{2.5}$ just as in sections 1.8.1 and 1.8.2. In the previous sections, the distributed lag models were used to show the univariate relationships between exposure and outcome controlling for potential confounders. In the current section, the path models seek to estimate these effects simultaneously and in aggregate along with the inclusion of an indirect effect. The path model is included below:

$$Y_{2,it} = \eta_{2,0} + \sum_{l=0}^{q} \beta_{2,l} x_{i,t-l} + \sum_{l=0}^{q} \varphi_l Y_{1,i,t-l} + \boldsymbol{\gamma}_2^T w_{1,it} + U_{2,i} + \epsilon_{2,it}, \tag{1.15}$$

where $Y_{1,i,t-l}$ is given by:

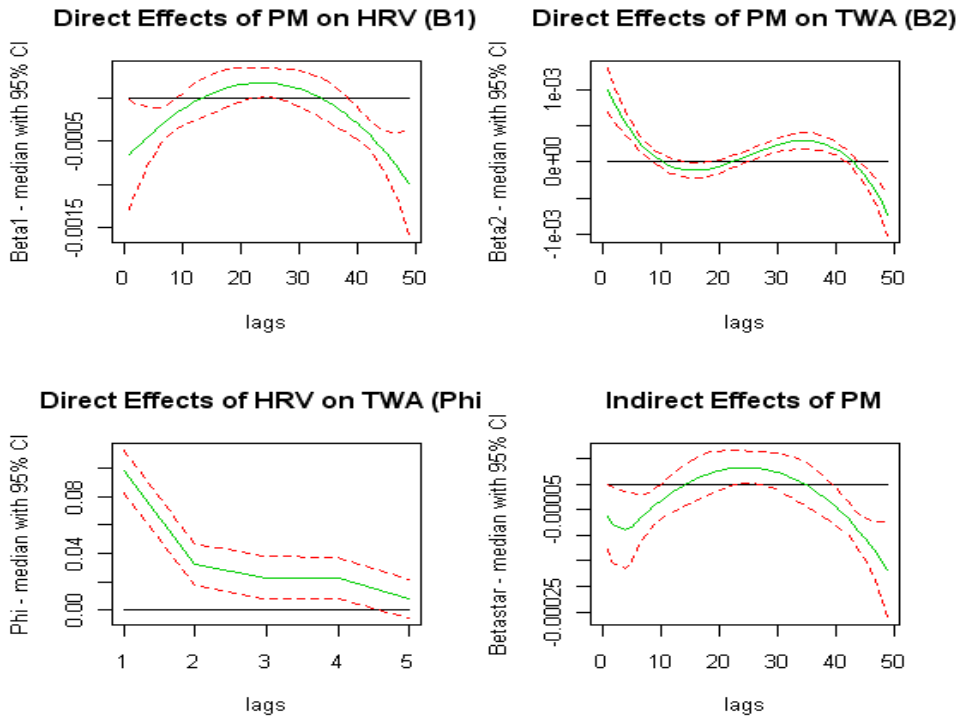


..

Figure 1.5: QUADRATIC PATH MODEL - Clockwise : 1) $PM_{2.5}$ on HRV, 2) $PM_{2.5}$ on TWA, 3) $PM_{2.5}$ on TWA indirectly through HRV, and 4) HRV on TWA.

$$Y_{1,it} = \eta_{1,0} + \sum_{l=0}^{q} \beta_{1,l} x_{i,t-l} + \boldsymbol{\gamma}_1^T w_{it} + U_{1,i} + \epsilon_{1,it}, \tag{1.16}$$

In order for the path model to be fully implemented a lag structure needed to be created for $Y_{1,i,t-l}$ (HRV). 4 lags were used to describe the relationship between HRV and TWA because the largest effects were seen within that time. As a result of the creation of these new lags 4 observations per study id had to be removed. We also investigated 8 lags for the HRV vs.TWA relationship and the results were comparable. Figures 1.5 - 1.7 show the distributed lag function for the following relationships in a clockwise fashion: 1) $PM_{2.5}$ on HRV, 2) $PM_{2.5}$ on TWA, 3) $PM_{2.5}$ on TWA indirectly through HRV, and 4) HRV on TWA. Each plot consists of a median posterior curve and a corresponding 95% credible interval. The distributed

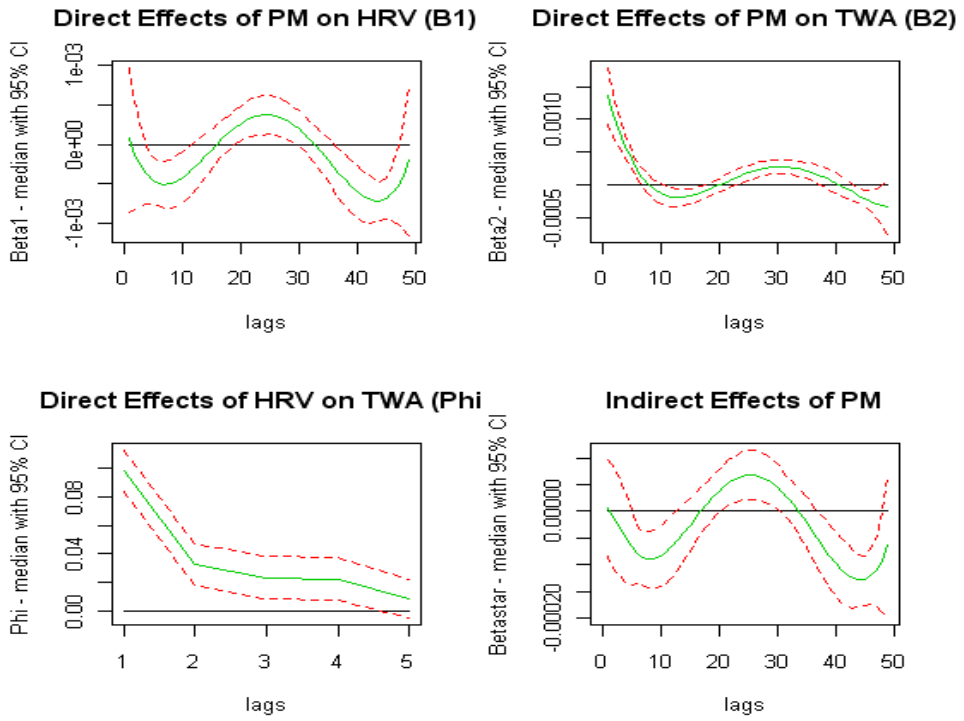


..

Figure 1.6: CUBIC PATH MODEL - Clockwise : 1) $PM_{2.5}$ on HRV, 2) $PM_{2.5}$ on TWA, 3) $PM_{2.5}$ on TWA indirectly through HRV, and 4) HRV on TWA.

lag functions in position 1) and 2) from each of the path models are similar in shape to their separate model counterparts respectively.

In Table1. 2 we see the overall estimated effects of the separate models juxtaposed with overall estimated effects of the pathway models. We found significant associations between HRV and $PM_{2.5}$ in both the separate and pathway models in all cases, and the overall estimates were similar. All of the estimates show a negative relationship whereby increases in $PM_{2.5}$ are associated with decreases in HRV. Further, the distributed lag functions of the quadratic, cubic, and quartic path models are similar. Significant associations were found between TWA and $PM_{2.5}$ in both the separate and pathway models in all cases as well. All of the estimates show a positive overall relationship whereby increases in $PM_{2.5}$ are associated with in-



..

Figure 1.7: QUARTIC PATH MODEL - Clockwise : 1) $PM_{2.5}$ on HRV, 2) $PM_{2.5}$ on TWA, 3) $PM_{2.5}$ on TWA indirectly through HRV, and 4) HRV on TWA.

creases in TWA. Once again, the distributed lag functions of the quadratic, cubic, and quartic path models are similar. The effect estimates from the separate models were attenuated by approximately 37-46% in comparison with the pathway model estimates for TWA. The direct and indirect effects are intrinsically included when conducting the single outcome models, which leads to the dimmed effect estimate.

The indirect effects cannot be estimated in the separate models because they would not have been simultaneously done. The pathway model estimates a negative overall effect for the indirect relationship between $PM_{2.5}$ and TWA through the intermediary HRV in the quadratic, cubic, and quartic cases. The overall impact of a unit change in $PM_{2.5}$ over 48 lags (24 hours) is associated with a -0.0086 (-0.0126, -0.0046), -0.0085 (-0.0124, -0.0044), and -0.0083 (-0.0124, -0.0042) decreases in HRV

Outcome	Quad. Separate Model Est. and 95% CI	Quad. Path Model Est. and 95% CI
HRV($\beta_{1,l}$):	-0.00872 (-0.013, -0.0049)†	-0.0086 (-0.0126, -0.0046) †
TWA($\beta_{2,l}$):	0.0023 (0.00042, 0.00421)†	0.0043 (0.0025, 0.0063)†
Indirect(β^*) :	N/A	-0.0014 (-0.0022, -0.00071) †
Outcome	**Cubic Separate Model Est. and 95% CI**	**Cubic Path Model Est. and 95% CI**
HRV($\beta_{1,l}$):	-0.0085 (-0.0124, -0.0046)†	-0.0085 (-0.0124, -0.0044)†
TWA($\beta_{2,l}$):	0.0030 (0.0012, 0.0050)†	0.0050 (0.0031, 0.0069)†
Indirect(β^*):	N/A	-0.0014 (-0.0021, -0.00065) †
Outcome	**Quartic Separate Model Est. and 95% CI**	**Quartic Path Model Est. and 95% CI**
HRV($\beta_{1,l}$):	-0.0084 (-0.0124, -0.0042)†	-0.0083 (-0.0124, -0.0042)†
TWA($\beta_{2,l}$):	0.0032 (0.0013, 0.0051)†	0.0051 (0.0032, 0.0070)†
Indirect(β^*):	N/A	-0.0015 (-0.0023, -0.00070)†

Table 1.2: This table represents the parameter estimates from separate parametric distributed lag models and the parameter estimates from the pathway models for 48 lags (24 hours). The intermediary φ is 4 lagged time-points. †Denotes significance at α = 0.05.

for the quadratic, cubic, and quartic models respectively. The overall impact of a unit change in $PM_{2.5}$ over 48 lags (24 hours) is associated with a 0.0043 (0.0025, 0.0063), 0.0050 (0.0031, 0.0069), and 0.0051 (0.0032, 0.0070) increase in log(TWA) μV for the quadratic, cubic, and quartic models respectively. Finally, the overall indirect impact of a unit change in $PM_{2.5}$ over 48 lags (24 hours) is associated with a -0.0014 (-0.0022, -0.00071), -0.0014 (-0.0021, -0.00065), and -0.0015 (-0.0023, -0.00070) decrease in log(TWA) μV through HRV for the quadratic, cubic, and quartic models respectively. We see that the distributed lag function for the indirect effects mimics the the DL function for HRV except for the the first 6 time points.

This study provides evidence that exposure to ambient air pollution in the form of $PM_{2.5}$ increases cardiac electrical instability over a 24 hour period capturing effects during sleep, morning hours, as well as normal activity. This finding offers a possible parsing of the mechanisms that lead to cardiac events.

1.9 DISCUSSION

In this paper, we considered methods to assess specific health effects of $PM_{2.5}$ as they relate to electrical cardiac outcomes. One objective was to detail the path by which particulate matter effected cardio-vascular outcomes and to parse out the effects between multiple outcomes. In a simulation study and corresponding analysis, we showed that the path way distributed lag model was able to estimate the effects of particulate matter on multiple outcomes simultaneously with relative accuracy when compared to separate models.

As an alternative to the moving average approach and the separate lag model method initially which allowed us to model a function rather than a single point estimate at different times. The advantage is that the relationship can be viewed with a fine, continuous resolution over the entire time period. We proposed a path way distributed lag model to account for multiple effects at different time intervals simultaneously. Simulations suggest that the proposed distributed lag pathway model is effective in separating the effects of multiple outcomes, which when done separately could be biased. The pathway models showed that the relationship between $PM_{2.5}$ and TWA-MAX was underestimated by greater than 37%(37-46%) compared to models being separately done. The pathway model was able to estimate the relationship between $PM_{2.5}$ and HRV with relative accuracy. All of the indirect effects were significant which lends evidence to the hypothesis that there are alternate/complementary/indirect biological pathways that can influence the direct relationship. Our results suggest that the magnitude of the indirect effect was highly dependent on the direction and length of the intermediate distributed lag function of HRV.

We also demonstrated the flexibility of the pathway approach to accommodate

different parametric modes such as quadratic, cubic, and quartic, set at different initial values, and remain consistent/accurate in estimation. The overall estimates were similar when moving from quadratic to cubic, from cubic to quartic, and from quadratic to quartic and the highest effects ere seen at early lags. In our data set and subsequent simulations, the distributed lag function for HRV nearly always mimicked the indirect effect distributed lag function although the effect was slightly attenuated. The changes in the DL function of the indirect effect would occur most notably in the first 4,8, or 12 time points depending on the number of lags in the HRV variable. Lastly, when including lagged HRV as a confounder in the single outcome models, the estimates remain the same as those given in the pathway models although the indirect effects could not be estimated. One limitation of these analyses is the time resolution of HRV and TWA. There were half-hour periods over which the outcomes measures for HRV and TWA were averaged. Therefore if shorter time resolution were needed for HRV in defining the pathway toTWA the model may not pick it up. Therefore it would be beneficial to not only have an increased sample size but also a much more finely measured time resolution which would allow effects to seen at the appropriate time points.

1.10 APPENDIX

1.10.1 Proof Direct and Indirect Effects

$$
\begin{aligned}
ST_{it} &= \eta_{1,0} + \sum_{l'=0}^{q'} \beta_{1,l'} x_{i,t-l'} + \sum_{j=1}^{d} f_j(s_{itj}) + \boldsymbol{\gamma_1^{\mathbf{T}} \mathbf{w_{it}}} + U_{1,i} + \epsilon_{1,it} \\
TWA_{it} &= \eta_{2,0} + \sum_{l=0}^{q} \beta_{2,l} x_{i,t-l} + \sum_{l=0}^{q} \varphi_l ST_{1,i,t-l} + \sum_{j=1}^{d} g_j(s_{itj}) + \boldsymbol{\gamma_2^{\mathbf{T}} \mathbf{w_{1,it}}} + U_{2,i} + \epsilon_{2,it}
\end{aligned}
$$

Now we substitute ST_{it} into equation for TWA_{it} and rearrange the terms.

$$
\begin{aligned}
TWA_{it} &= \eta_{2,0} + \sum_{l=0}^{q} \beta_{2,l} x_{i,t-l} + \sum_{l=0}^{q} \varphi_l \left[\eta_{1,0} + \sum_{l'=0}^{q'} \beta_{1,l'} x_{i,t-l'-l} + \cdots + \epsilon_{1,i,t-l}, \right] \\
&+ \sum_{j=1}^{d} g_j(s_{itj}) + \boldsymbol{\gamma_2^{\mathbf{T}} \mathbf{w_{1,it}}} + U_{2,i} + \epsilon_{2,it} \\
&= \eta_{2,0} + \sum_{l=0}^{q} \beta_{2,l} x_{i,t-l} + \sum_{l=0}^{q} \varphi_l \left[\eta_{1,0} \right] + \sum_{l=0}^{q} \varphi_l \left[\sum_{l'=0}^{q'} \beta_{1,l'} x_{i,t-l'-l} \right] \\
&+ \sum_{l=0}^{q} \varphi_l \left[\sum_{j=1}^{d} f_j(s_{i,t-l,j}) \right] + \sum_{l=0}^{q} \varphi_l \left[\boldsymbol{\gamma_1^{\mathbf{T}} \mathbf{w_{i,t-l}}} \right] + \sum_{l=0}^{q} \varphi_l \left[U_{1,i} \right] + \sum_{l=0}^{q} \varphi_l \left[\epsilon_{1,i,t-l} \right] \\
&+ \sum_{j=1}^{d} g_j(s_{itj}) + \boldsymbol{\gamma_2^{\mathbf{T}} \mathbf{w_{1,it}}} + U_{2,i} + \epsilon_{2,it} \\
&= \left[\eta_{2,0} + \sum_{l=0}^{q} \varphi_l \eta_{1,0} \right] + \left[\sum_{l=0}^{q} \beta_{2,l} x_{i,t-l} + \sum_{l=0}^{q} \sum_{l'=0}^{q'} \varphi_l \beta_{1,l'} x_{i,t-l'-l} \right] + \\
&+ \left[\sum_{l=0}^{q} \sum_{j=1}^{d} \varphi_l f_j(s_{i,t-l,j}) + \sum_{j=1}^{d} g_j(s_{itj}) \right] + \left[\sum_{l=0}^{q} \varphi_l \boldsymbol{\gamma_1^{\mathbf{T}} \mathbf{w_{i,t-l}}} + \boldsymbol{\gamma_2^{\mathbf{T}} \mathbf{w_{1,it}}} \right] \\
&+ \left[\sum_{l=0}^{q} \varphi_l U_{1,i} + U_{2,i} \right] + \left[\sum_{l=0}^{q} \varphi_l \epsilon_{1,i,t-l} + \epsilon_{2,it} \right]
\end{aligned}
$$

Since we are only interested in the direct and indirect effects of exposure and their interpretations we will use the following as our model of interest where $[**]$ represents the other terms/confounders in the model.

$$TWA_{it} = \left[\sum_{l=0}^{q} \beta_{2,l} x_{i,t-l} + \sum_{l=0}^{q}\sum_{l'=0}^{q'} \varphi_l \beta_{1,l'} x_{i,t-l'-l}\right] + [**]\,.$$

The $\beta_{2,l}$ parameters represent the direct effects of lagged exposure on our outcome. Next we will expand the expression of the double sum to get the indirect effect of exposure on our outcome Taking this expression we have:

$$\begin{aligned}
\sum_{l=0}^{q}\sum_{l'=0}^{q'} \varphi_l \beta_{1,l'} x_{i,t-l'-l} &= \varphi_0 \left[\beta_{10} x_{i,t-0} + \beta_{11} x_{i,t-1} + \beta_{12} x_{i,t-2} + \cdots + \beta_{1q'} x_{i,t-q'}\right] \\
&+ \varphi_1 \left[\beta_{10} x_{i,t-1} + \beta_{11} x_{i,t-2} + \beta_{12} x_{i,t-3} + \cdots + \beta_{1q'} x_{i,t-q'}\right] \\
&+ \varphi_2 \left[\beta_{10} x_{i,t-2} + \beta_{11} x_{i,t-3} + \beta_{12} x_{i,t-4} + \cdots + \beta_{1q'} x_{i,t-q'}\right] \\
&\vdots \\
&+ \varphi_q \left[\beta_{10} x_{i,t-q} + \beta_{11} x_{i,t-q-1} + \beta_{12} x_{i,t-q-2} + \cdots + \beta_{1q'} x_{i,t-q-q'}\right]
\end{aligned}$$

By distributing and rearranging terms we have:

$$
\begin{aligned}
&= [\varphi_0\beta_{10}x_{i,t-0} + \varphi_0\beta_{11}x_{i,t-1} + \varphi_0\beta_{12}x_{i,t-2} + \varphi_0\beta_{13}x_{i,t-3} + \cdots + \varphi_0\beta_{1q'}x_{i,t-q'}] \\
&+ [\varphi_1\beta_{10}x_{i,t-1} + \varphi_1\beta_{11}x_{i,t-2} + \varphi_1\beta_{12}x_{i,t-3} + \varphi_1\beta_{13}x_{i,t-4} + \cdots + \varphi_1\beta_{1q'}x_{i,t-1-q'}] \\
&+ [\varphi_2\beta_{10}x_{i,t-2} + \varphi_2\beta_{11}x_{i,t-3} + \varphi_2\beta_{12}x_{i,t-4} + \varphi_2\beta_{13}x_{i,t-5} + \cdots + \varphi_2\beta_{1q'}x_{i,t-2-q'}] \\
&\vdots \\
&+ [\varphi_q\beta_{10}x_{i,t-q} + \varphi_q\beta_{11}x_{i,t-q-1} + \varphi_q\beta_{12}x_{i,t-q-2} + \varphi_q\beta_{13}x_{i,t-q-3} + \cdots + \varphi_q\beta_{1q'}x_{i,t-q-q'}]
\end{aligned}
$$

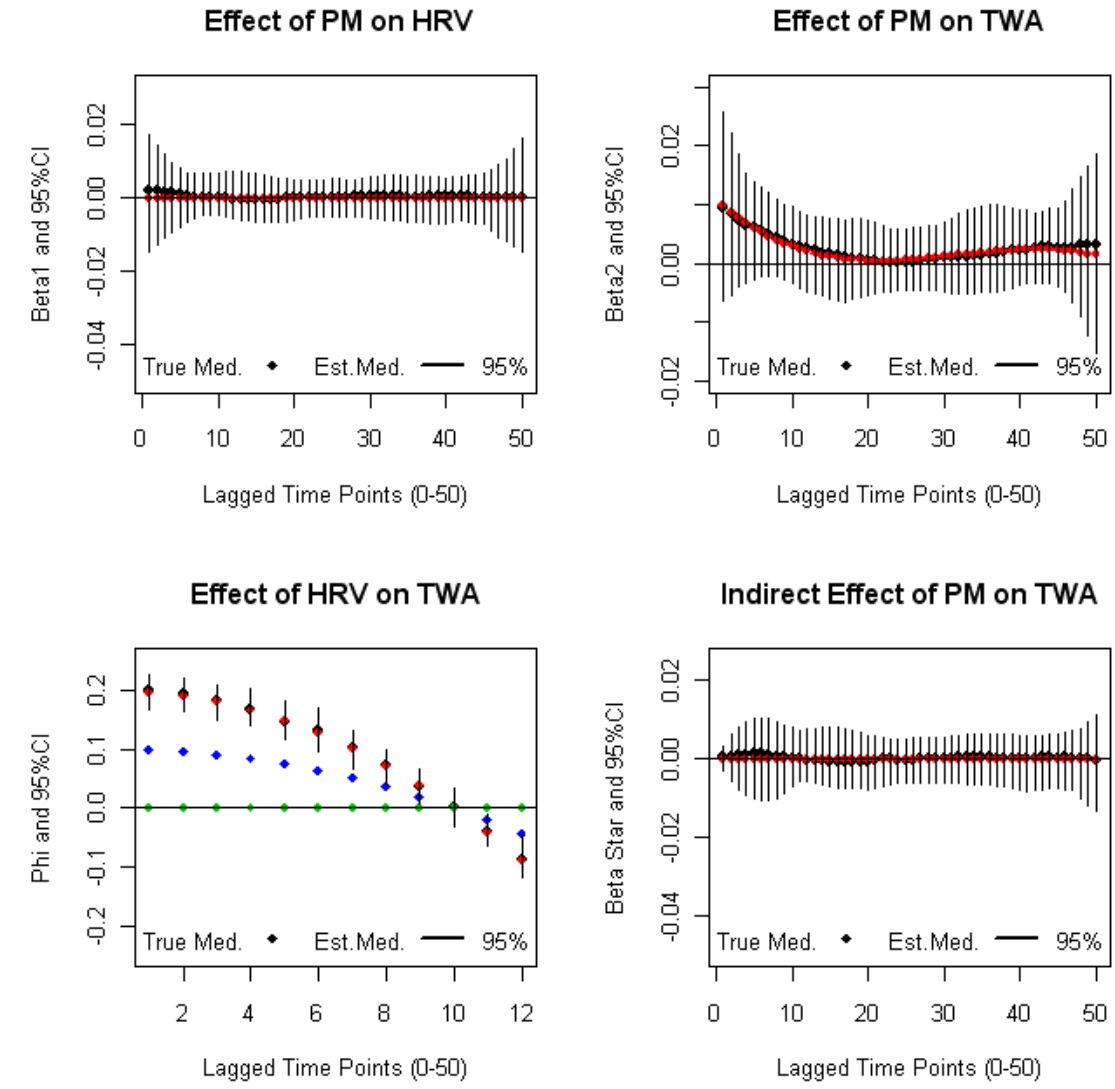
$$
\begin{aligned}
&= \overbrace{[\varphi_0\beta_{10}]}^{\beta_0^*} x_{i,t-0} + \overbrace{[\varphi_0\beta_{11} + \varphi_1\beta_{10}]}^{\beta_1^*} x_{i,t-1} + \overbrace{[\varphi_0\beta_{12} + \varphi_1\beta_{11} + \varphi_2\beta_{10}]}^{\beta_2^*} x_{i,t-2} \\
&+ \cdots + [\varphi_0\beta_{1q'} + \varphi_1\beta_{1,q'-1} + \cdots + \varphi_q\beta_{10}]\, x_{i,t-q-q'}
\end{aligned}
$$

$$
\begin{aligned}
&= \beta_0^* x_{i,t-0} + \beta_1^* x_{i,t-1} + \beta_2^* x_{i,t-2} + \beta_3^* x_{i,t-3} + \beta_4^* x_{i,t-4} + \cdots + \beta_{q+q'}^* x_{i,t-q-q'} \\
&= \sum_{k=0}^{q+q'} \beta_k^* x_{i,t-k}
\end{aligned}
$$

where $\beta_k^* = \sum_{l+l'=k} [\varphi_l\beta_{1l'}]$ which represents the indirect effect of the lagged exposure on outcome. The final models is as follows:

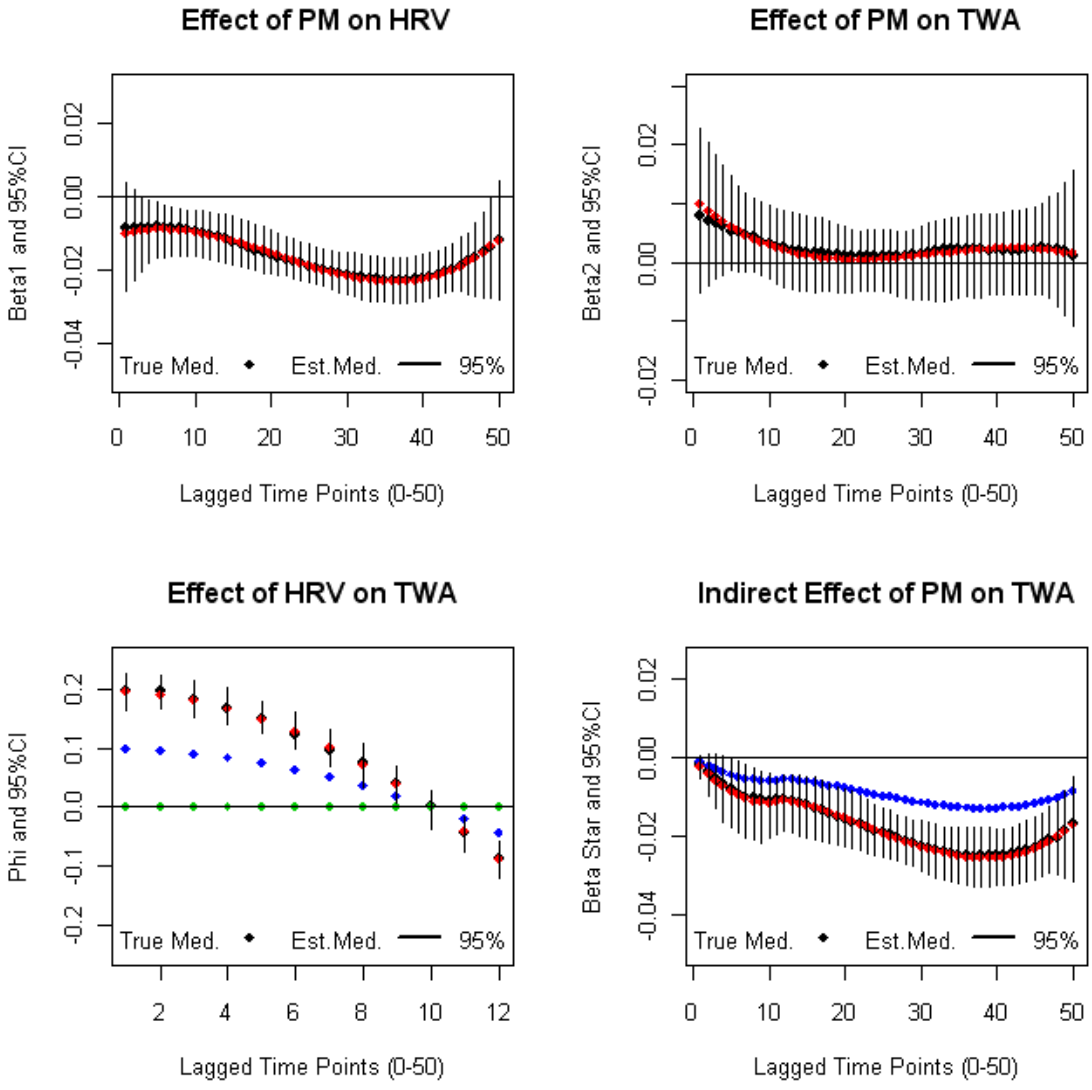
$$
\begin{aligned}
TWA_{it} &= \overbrace{\left[\eta_{2,0} + \sum_{l=0}^{q} \varphi_l \eta_{1,0}\right]}^{Intercepts} + \overbrace{\left[\sum_{l=0}^{q} \beta_{2,l} x_{i,t-l} + \sum_{k=0}^{q+q'} \beta_k^* x_{i,t-k}\right]}^{DirectandIndirectEffects} \\
&+ \overbrace{\left[\sum_{l=0}^{q}\sum_{j=1}^{d} \varphi_l f_j(s_{i,t-l,j}) + \sum_{j=1}^{d} g_j(s_{itj})\right]}^{NonlinearConfounders} + \overbrace{\left[\sum_{l=0}^{q} \varphi_l \gamma_{\mathbf{1}}^{\mathbf{T}} \mathbf{w_{i,t-l}} + \gamma_{\mathbf{2}}^{\mathbf{T}} \mathbf{w_{1,it}}\right]}^{LinearConfounders} \\
&+ \overbrace{\left[\sum_{l=0}^{q} \varphi_l U_{1,i} + U_{2,i}\right]}^{RandomEffects} + \overbrace{\left[\sum_{l=0}^{q} \varphi_l \epsilon_{1,i,t-l} + \epsilon_{2,it}\right]}^{Error}
\end{aligned}
$$

1.10.2 Simulations Scenarios 1-7



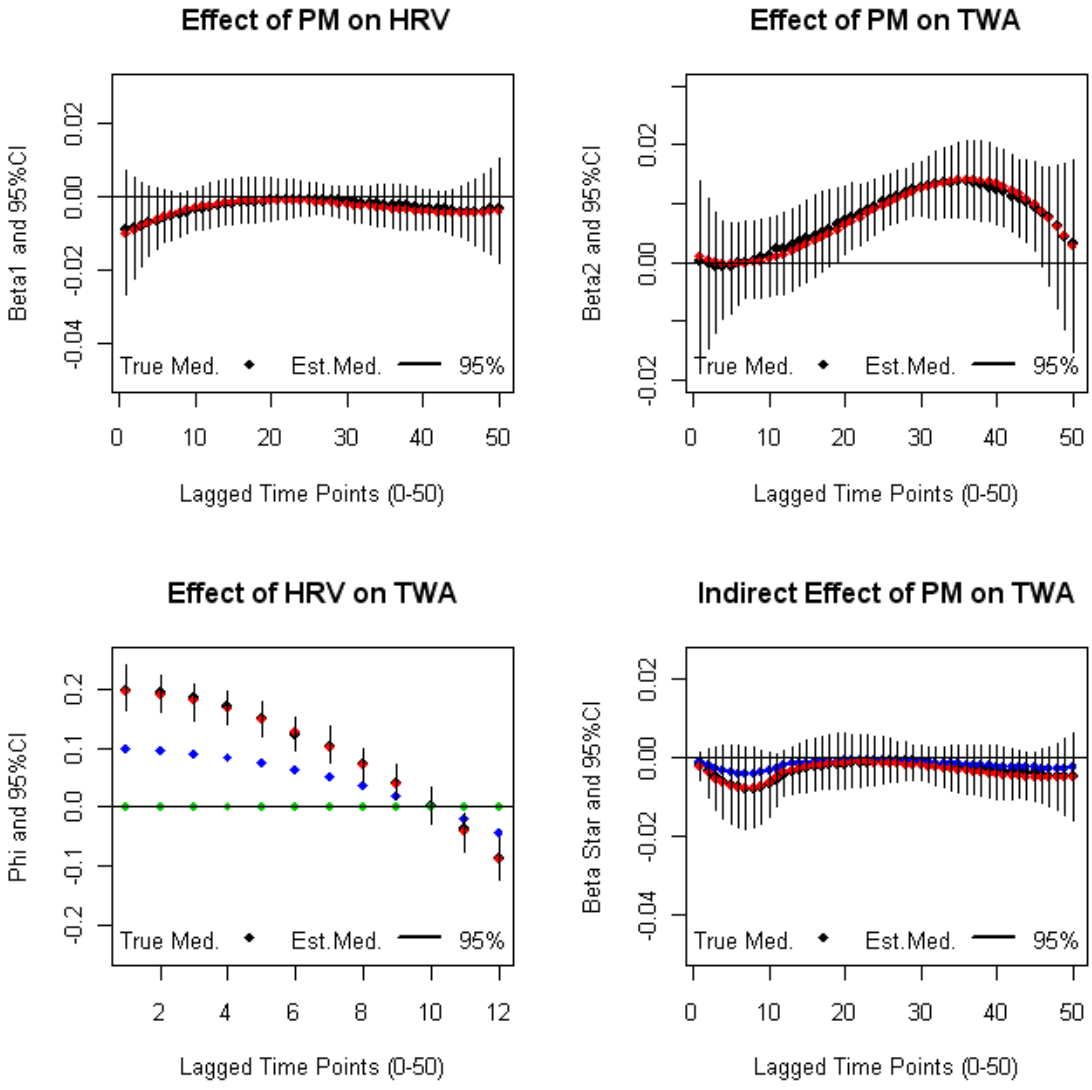
..

Figure 1.8: Model 2-No Effect HRV; Regular TWA



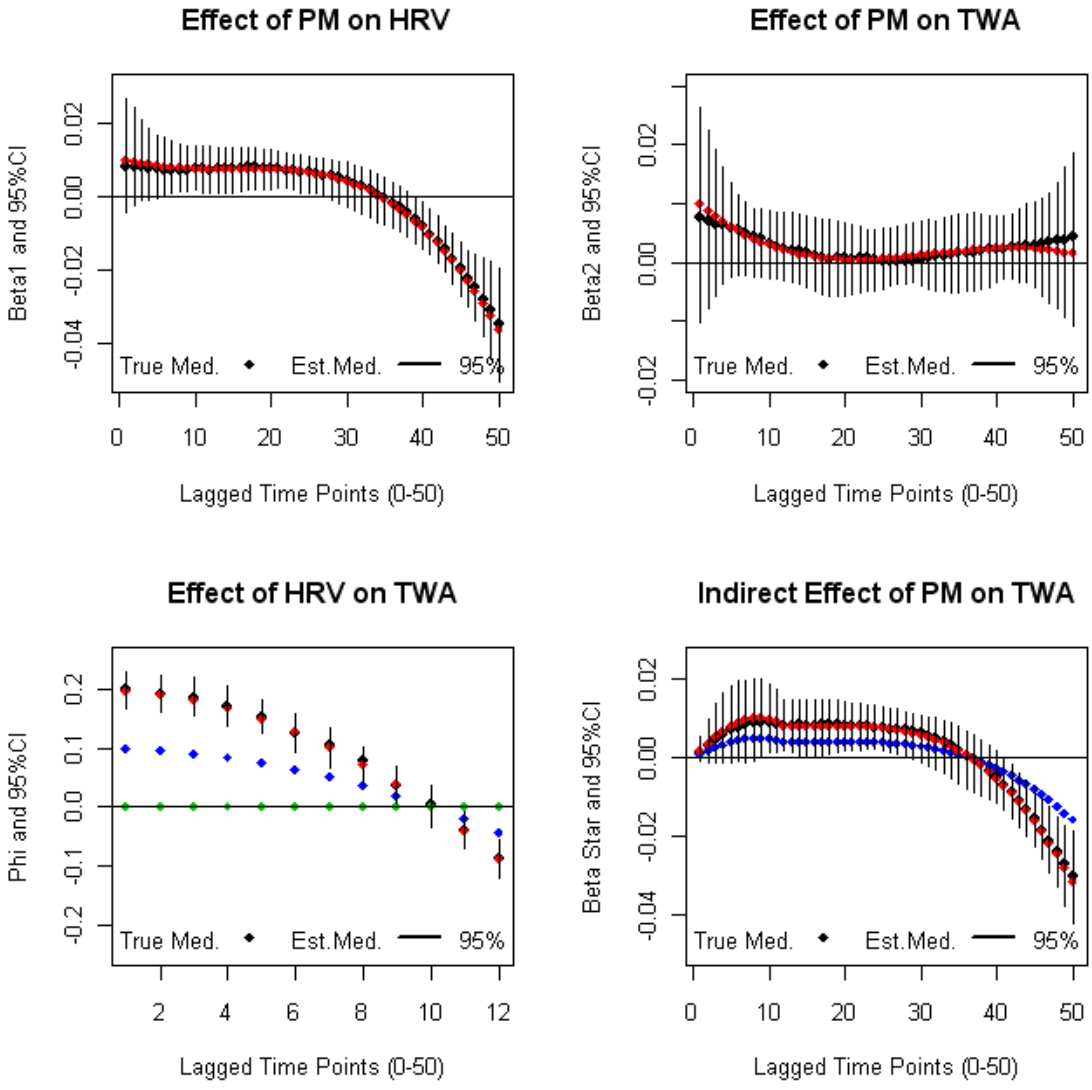
..

Figure 1.9: Model 3-Shifted Effect HRV; Regular TWA



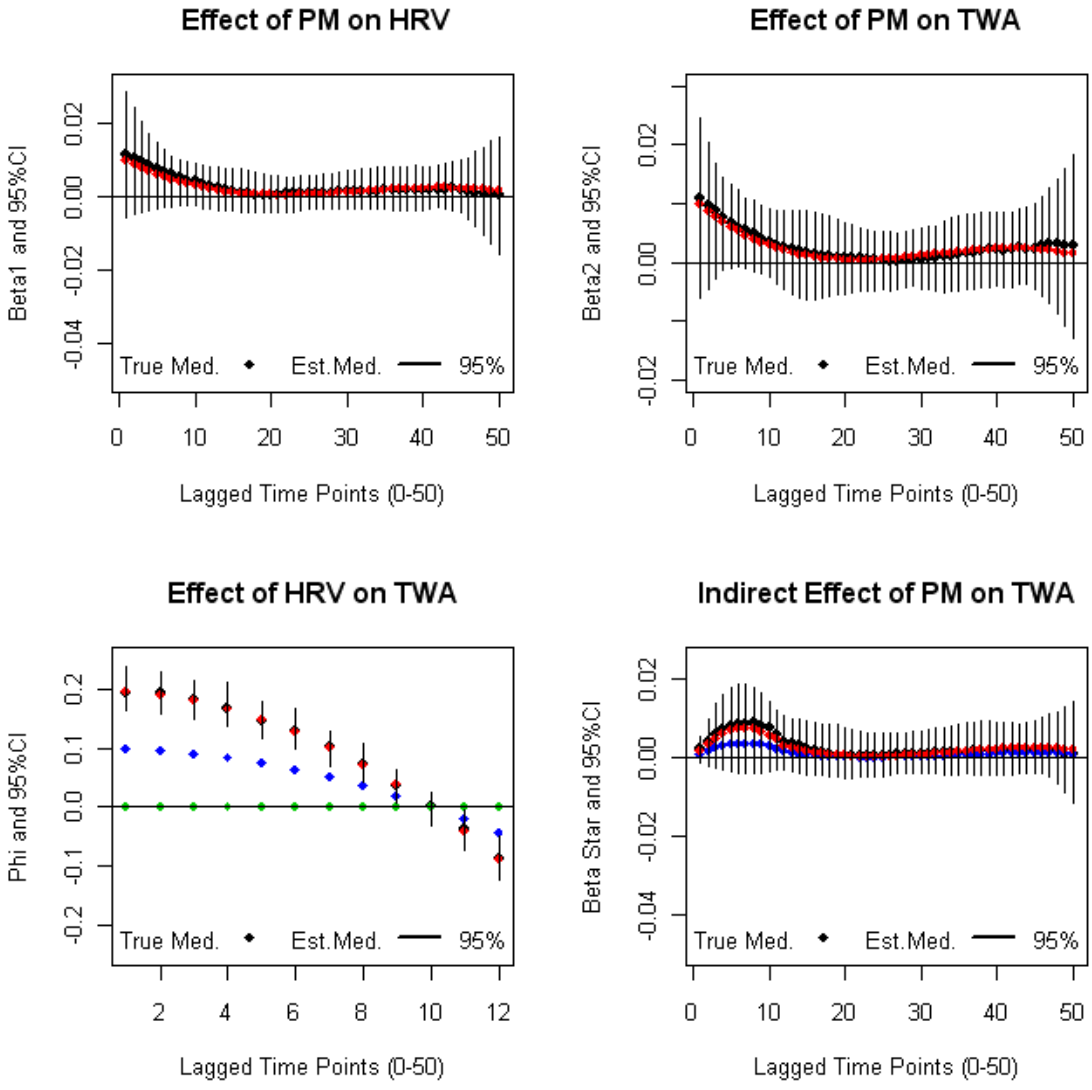
..

Figure 1.10: Model 4-Regular HRV; Shifted Effect TWA



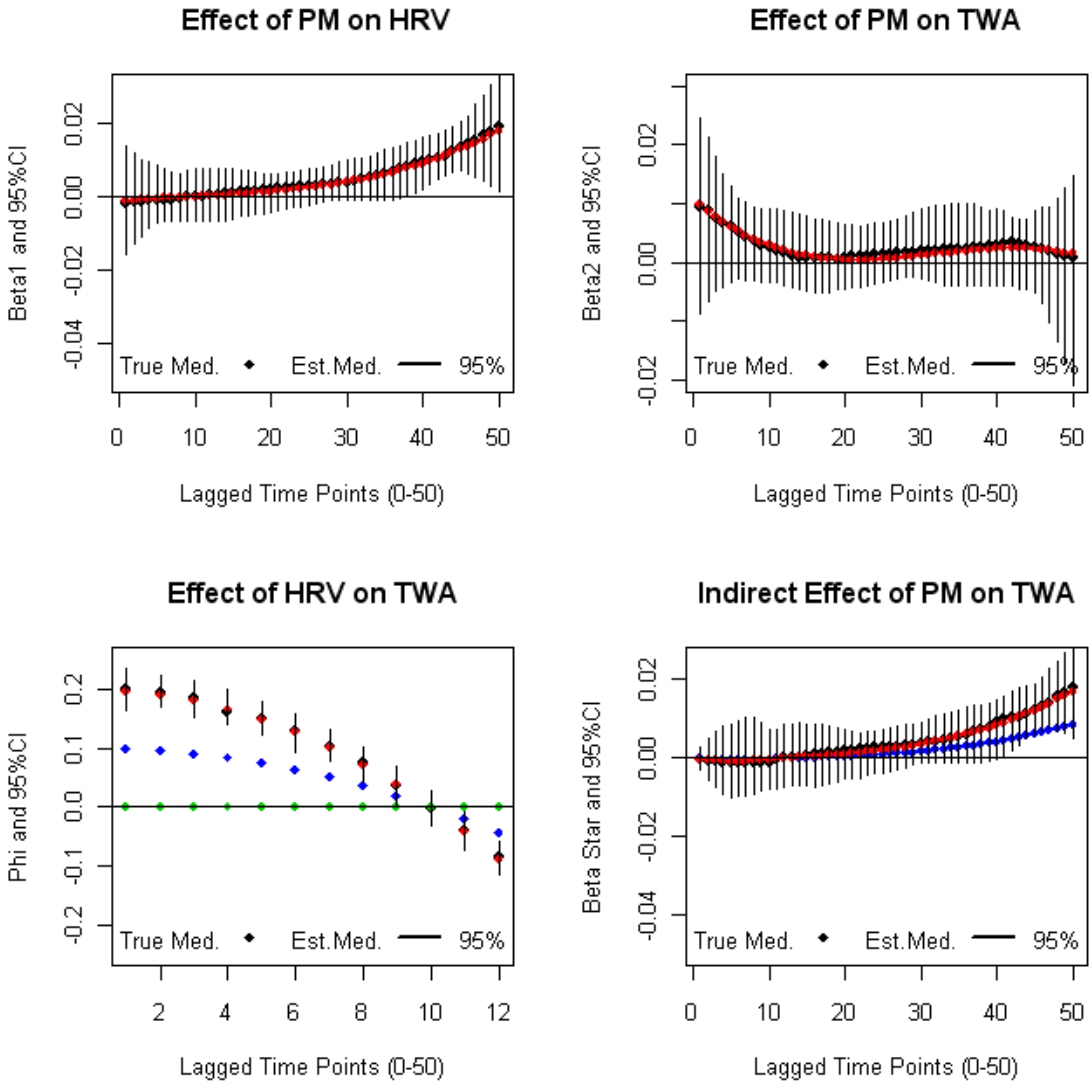
..

Figure 1.11: Model 5-Heavy end Effect HRV; Regular TWA



..

Figure 1.12: Model 6-Positive Effect HRV; Regular TWA



..

Figure 1.13: Model 7-Positive Effect HRV (downward); Regular TWA

New regression Calibration Approaches for Missing Exposure Data in Panel Studies

[1]Alfa I. Yansané, [2,3]Diane R. Gold, [1,4]Paul J. Catalano,[1] Brent A. Coull, [2]Petros Koutrakis

[1]Department of Biostatistics, Harvard School of Public Health

[2]Department of Environmental Health, Harvard School of Public Health

[3]Department of Medicine, Brigham and Women's Hospital/Harvard Medical School and

[4]Department of Biostatistics, Dana-Farber Cancer Institute

2.1 ABSTRACT

Epidemiological studies have shown that an individual's cumulative exposure to pollutants is a result of both acute and lagged effects. Therefore, accurate measures of exposure patterns are instrumental to conducting a health effects analysis. The variables traditionally studied in environmental epidemiology, such as exposure to ambient outdoor pollution, indoor pollutants, and environmental hazards are often subject to measurement errors. Although, there has been considerable research conducted on missing data and measurement error, little work has been done to address missing data in the moving average time series setting. In this article we propose a new regression calibration approach that focuses on estimating missing moving averages by conditioning on the observed data. We compare via simulation the performance of our "moving average" approach in estimating missing exposure data to that of 2 existing approaches, a daily value imputation method and a reduced method where missing observations are deleted. Simulation results suggest that the proposed regression calibration using the moving average lead to robust and less biased estimates of the parameters of interest.

2.2 INTRODUCTION

Variables commonly studied in environmental epidemiology, such as exposure to ambient outdoor pollution, indoor pollutants, and environmental hazards are often subject to measurement error. Such pollutants have many sources of variability such as instrument error, recording error, and missing observations. Misclassification of exposures is a well-recognized inherent limitation of panel studies linking diseases to the environment. For many agents of interest, exposures have both spatial and temporal influences. As a result, it is often a daunting task for investigators to accurately represent the relevant exposures for each participant. Researchers continue to take steps to control the consequences of measurement error through conscientious study designs and data collection, and by making adjustments for the error in the statistical analyses (Zeger et al., 2000).

In many panel studies, access to ambient exposure information for several consecutive periods (e.g. days, months, years previous) are available due to central site monitoring but the precision of the measurements have inconsistencies. Alternatively, access to indoor exposure measurements requires more resources for accurate and complete data collection, include increased participant burdens, and require higher costs for the longer term measurements. As the issue of exposure errors has become well recognized in the literature, researchers have shown that human activities impact the timing, location, and degree of pollutant exposure. According to the National Human Activity Pattern Survey (NHAPS), respondents reported spending an average of 87% of their time in enclosed buildings and about 6% of their time in enclosed vehicles (Klepeis N et al., 2001). Therefore human behaviors are key contributors in explaining exposure variations. These statistics underscore the need to detail pollutant exposures using indoor methods so that the

most comprehensive arch of vulnerability and susceptibility could be captured. In this paper, we seek to develop methods that correct covariate/exposure measurement error due to missing data through the use of regression calibration.

Existing analyses, when confronted with missing indoor exposure data may attempt to either remove participants/observations with incomplete data or conduct regression calibration methods by imputing the individual exposure observations. Since indoor exposure data tend to be sparse, discarding valuable observations may not be an optimal option. In this paper, we propose to develop methods that allow one to perform imputation methods on the moving averages of unobserved covariate exposures conditional on the observed data. This is achieved using large sample methods that require statistical models for the distribution of exposure $\mathbf{X}$ conditional on observed covariates. In addition to these considerations, time series models are used so that temporal components can be introduced and accounted for in the exposure-outcome relationship. Typically, such data represent a sequence of observations at successive times and spaced at uniform time intervals. An intrinsic statistical issue in the collection of time series data (lagged data) is having the appropriate amount of sequenced information collected for each subject. More specifically, the value of some of the the variables of interest may not be observable for all study participants. For example, a variable may be observed for 80% of the study, but unobserved for the other 20%. This presents a unique problem when dealing with pollution exposure data and potentially any longitudinal study where missing data are involved. A mechanism/procedure must accurately and efficiently estimate the missing information or risk conducting an analysis with insufficient data. This missing data issue could be more easily illustrated through the following example:

At the initial start date of a study, there are no indoor exposure measurements for time lags previous to the start date. As a result, the data are incomplete and those rows of data must be deleted in order to proceed further with an analysis. For example, Table 2.1 shows sample exposure data for 7 lagged days. In order to calculate the 4-day moving average the data are insufficient. In the sample data, it can be readily seen that each subject was followed for 7 days but the 4-day moving average can only be calculated for the last 4 measurements of each ID because the rows with NA's must be deleted as seen in Table 2.2. Subsequently, to conduct the analysis investigators must work from a reduced data set like the one seen in Table 2.3. Given that each subject should have a fixed number of measurements taken at successive time points, our proposed regression calibration method seeks to impute the missing moving averages using the existing exposure information. We would like to condition on the current observed information to find estimates for the missing 4-day averages as seen in table four. This would allow investigators to utilize all of their data more efficiently.

	Start Date						
ID	$x_{i,0}$	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	$x_{i,5}$	$x_{i,6}$
1	3.6	4.7	5.2	3.8	4.1	3.9	5.6
2	3.5	5.6	4.3	4.7	3.2	2.9	6.4
3	2.6	6.5	3.4	5.6	2.3	3.8	5.5
4	3.7	5.2	4.5	5.2	3.6	4.8	6.5

Table 2.1: SAMPLE OBSERVED DATA - This table represents the observed exposure data over a 7 day period.

The data that motivates the proposed research comes from the analyses conducted through the Electric Power Research Institute through the Harvard School of Public Health on the effects of indoor and outdoor air pollution (particulate matter, black carbon, carbon monoxide, ozone, nitrogen dioxide, and sulfur dioxide) on children with respiratory illnesses. Investigators have conducted a number of re-

	Start Date						
ID	$X_{i,0}^{(4)}$	$X_{i,1}^{(4)}$	$X_{i,2}^{(4)}$	$X_{i,3}^{(4)}$	$X_{i,4}^{(4)}$	$X_{i,5}^{(4)}$	$X_{i,6}^{(4)}$
1	NA	NA	NA	4.325	4.45	4.25	4.35
2	NA	NA	NA	4.525	4.45	3.775	4.3
3	NA	NA	NA	4.525	4.45	3.775	4.3
4	NA	NA	NA	4.65	4.625	4.525	5.025

Table 2.2: SAMPLE MOVING AVERAGE DATA - This table represents the observed exposure data for a 4-day moving average by id.

ID	$X_{i,3}^{(4)}$	$X_{i,4}^{(4)}$	$X_{i,5}^{(4)}$	$X_{i,6}^{(4)}$
1	4.325	4.45	4.25	4.35
2	4.525	4.45	3.775	4.3
3	4.525	4.45	3.775	4.3
4	4.65	4.625	4.525	5.025

Table 2.3: SAMPLE REDUCED MOVING AVERAGE DATA - This table represents the observed exposure data for a 4-day moving average by id where the missing values were deleted.

gression analyses using moving averages of exposure and these outcomes. Pollutants were measured from two central monitoring sites in New York City, NY while the indoor exposures were measured using indoor sampling apparatus. The outcomes of interests were onset of asthma, eczema, hay fever, and pulmonary function.

This paper is organized as follows: Section 2.3 describes in detail the design and data from a study evaluating the effects of particulate air pollution on outcome. Section 2.4 presents the exposure/error model, subsequent imputation algorithms, and model mis-specification. Section 2.5 presents a simulation study to examine the effectiveness of the imputation methods compared to existing approaches and tests the robustness of the regression calibration and Section 2.6 demonstrates an application of the regression calibration methods to analyze the afore mentioned study from Exposure, Epidemiology, and Risk Program. Finally in Section 2.7 we discuss our findings along with its implications.

ID	$X_{i,0}^{(4)}$	$X_{i,1}^{(4)}$	$X_{i,2}^{(4)}$	$X_{i,3}^{(4)}$	$X_{i,4}^{(4)}$	$X_{i,5}^{(4)}$	$X_{i,6}^{(4)}$
1	$\mu_{X^{(4)}\|X^{(1)}}$	$\mu_{X^{(4)}\|X^{(2)}}$	$\mu_{X^{(4)}\|X^{(3)}}$	4.325	4.45	4.25	4.35
2	$\mu_{X^{(4)}\|X^{(1)}}$	$\mu_{X^{(4)}\|X^{(2)}}$	$\mu_{X^{(4)}\|X^{(3)}}$	4.525	4.45	3.775	4.3
3	$\mu_{X^{(4)}\|X^{(1)}}$	$\mu_{X^{(4)}\|X^{(2)}}$	$\mu_{X^{(4)}\|X^{(3)}}$	4.525	4.45	3.775	4.3
4	$\mu_{X^{(4)}\|X^{(1)}}$	$\mu_{X^{(4)}\|X^{(2)}}$	$\mu_{X^{(4)}\|X^{(3)}}$	4.65	4.625	4.525	5.025

Table 2.4: MOVING AVERAGE IMPUTATION IDEA - This table represents the observed exposure data for a 4-day moving average by id.

2.3 DATA

Subjects were recruited from 6 different mediums; The Pediatric Emergency Department at MSSM, Chest Clinic at MSSM, Asthma Support Group at MSSM, Health Fairs, Referral from outside sources, and Advertisements. Children between the ages of 6 and 14 with moderate to severe asthma as defined by NIH criteria. Each subject had to reside North of 96th Street in Manhattan and South of Cross Bronx Expressway in the Bronx and sleep in the same place at least 5 nights a week. Children with active disease other than asthma requiring daily medications and those with mental retardation were excluded. Further exclusions were smoking in the home and family planning to move from current home within the next six months.

2.4 MOVING AVERAGE IMPUTATION

To begin, let X_{it} represent the value of a particular air pollutant and $X_{i,t}^{(n)}$ is the n-day moving average for subject i at time t. Let Y_{it} be the value of the outcome of interest for subject i at time t. We recognize that pollution measures taken at close intervals in time should be correlated and our estimate of the missing values should take this into consideration. Allow each subject to have covariate expo-

sure measurements for the current day and 6 previous days yielding a total of 7 measurements for each subject. In total one subject may have between 1 and 6 missing values (per visit if needed). For instance, if there is one missing value for the moving average, then we can use the 6 day moving average from the remaining observed values to estimate the 7 day moving average value of air pollution. In the data, it can be readily seen that each subject was followed for 14(2 weeks) days but the full 7-day moving average can only be calculated for the last 8 measurements of each ID as in the example in section 2.2. To conduct the analysis without operating from a reduced data set that deletes unobserved moving averages, a regression calibration can be performed. Each 7-day moving average will be estimated conditional on incomplete but observed exposure information. Therefore each $\mu_{X^{(7)}|X^{(m)}}$ is estimated conditional on the m-day moving average where $m < 7$. Table 2.5 can give a theoretical illustration of the values being estimated. We assume the following mean model and error:

ID	$X_{i,0}^{(7)}$	$\dots$	$X_{i,6}^{(7)}$	$X_{i,7}^{(7)}$	$X_{i,8}^{(7)}$	$\dots$	$X_{i,14}^{(7)}$
1	$\mu_{X^{(7)}\|X^{(1)}}$	$\dots$	$\mu_{X^{(7)}\|X^{(6)}}$	$x_{1,7}^{(7)}$	$x_{1,8}^{(7)}$	$\dots$	$x_{1,14}^{(7)}$
2	$\mu_{X^{(7)}\|X^{(1)}}$	$\dots$	$\mu_{X^{(7)}\|X^{(6)}}$	$x_{2,7}^{(7)}$	$x_{2,8}^{(7)}$	$\dots$	$x_{2,14}^{(7)}$
3	$\mu_{X^{(7)}\|X^{(1)}}$	$\dots$	$\mu_{X^{(7)}\|X^{(6)}}$	$x_{3,7}^{(7)}$	$x_{3,8}^{(7)}$	$\dots$	$x_{3,14}^{(7)}$
4	$\mu_{X^{(7)}\|X^{(1)}}$	$\dots$	$\mu_{X^{(7)}\|X^{(6)}}$	$x_{4,7}^{(7)}$	$x_{4,8}^{(7)}$	$\dots$	$x_{4,14}^{(7)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
30	$\mu_{X^{(7)}\|X^{(1)}}$	$\dots$	$\mu_{X^{(7)}\|X^{(6)}}$	$x_{30,7}^{(7)}$	$x_{30,8}^{(7)}$	$\dots$	$x_{30,14}^{(7)}$

Table 2.5: MOVING AVERAGE IMPUTATION IDEA - This table represents the observed exposure data for a 7-day moving average by id.

2.4.1 Simple Exposure Model Including of Covariates

The simple exposure model is as follows:

$$\mathbf{X}_i = \boldsymbol{\beta} W_i + \boldsymbol{\epsilon}_i \tag{2.1}$$

Where $\mathbf{X}_i$ represents a vector of exposure observations for subject i. Each of its components are of the form X_{it} for subject i at time t. $\mathbf{W}_i$ is a vector of possible confounders that influence the level exposure and it components are of the form W_{it} for the ith subject at time t. For $1 \leq t \leq T$ and $1 \leq i \leq N$. $\boldsymbol{\epsilon_i} \sim N(0, \boldsymbol{\Sigma})$ so that $\mathbf{X_i} \sim \mathbf{N}(\boldsymbol{\beta}\mathbf{W_i}, \boldsymbol{\Sigma})$. Also assume that the errors follow the AR(1) correlation structure where $\rho^{(n)}$, and $\Lambda^{(n)}$ are the correlation matrix and variance-covariance matrix for the n-day moving average respectively.

$$\boldsymbol{\rho}^{(7)} = \begin{bmatrix} 1 & \rho & \rho^{\mathbf{2}} & \dots & \rho^{\mathbf{6}} \\ \rho & 1 & \rho & \dots & \rho^{\mathbf{5}} \\ \rho^{\mathbf{2}} & \rho & 1 & \dots & \rho^{\mathbf{4}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{\mathbf{6}} & \rho^{\mathbf{5}} & \rho^{\mathbf{4}} & \dots & 1 \end{bmatrix},$$

In general, we have the correlation and covariance relation formula that can be rearranged to compute all of the components of the variance covariance matrix for the errors.

$$\rho = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

$$\boldsymbol{\epsilon}^{(7)} \sim MVN\left(0, \boldsymbol{\Sigma}^{(7)} = \begin{bmatrix} \sigma^{\mathbf{2}} & \sigma^2\rho & \sigma^2\rho^{\mathbf{2}} & \dots & \sigma^2\rho^{\mathbf{6}} \\ \sigma^2\rho & \sigma^{\mathbf{2}} & \sigma^2\rho & \dots & \sigma^2\rho^{\mathbf{5}} \\ \sigma^2\rho^{\mathbf{2}} & \sigma^2\rho & \sigma^{\mathbf{2}} & \dots & \sigma^2\rho^{\mathbf{4}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho^{\mathbf{6}} & \sigma^2\rho^{\mathbf{5}} & \sigma^2\rho^{\mathbf{4}} & \dots & \sigma^{\mathbf{2}} \end{bmatrix}\right),$$

2.4.2 Distribution of the Moving Average

$$E\left[X_{it}|W_{1it}, \dots, W_{nit}\right] = \beta_0 + \beta_1 W_{1it} + \cdots + \beta_n W_{nit}$$

Let $S_{it}^{(n)}$ represent the sum of the pollutant measures for the i^{th} subject at time t. Therefore we have $M_{it}^{(n)}$ represents the n-day moving average for the i^{th} subject at time t. The respective distributions are as follows:

$$\begin{aligned} S_{it}^{(n)} &= X_{i1} + X_{i2} + X_{i3} \cdots + X_{in} \\ M_{it}^{(n)} &= \frac{S_{it}^{(n)}}{n} \end{aligned}$$

$$
\begin{aligned}
E\left[\mathbf{M_i^{(n)}}|\mathbf{W_i}\right] &= E\left[\frac{1}{n}\mathbf{S_i^{(n)}}|\mathbf{W_i}\right] \\
&= E\left[\frac{1}{n}\sum_{t=1}^{n}\mathbf{X_i}|\mathbf{W_i}\right] \\
&= \frac{1}{n}\sum_{t=1}^{n}E\left[\mathbf{X_i}|\mathbf{W_i}\right] \\
&= \frac{1}{n}\sum_{t=1}^{n}[\boldsymbol{\beta}\boldsymbol{W}_i] \\
&= \boldsymbol{\beta}\frac{1}{n}\sum_{t=1}^{n}\mathbf{W_i} \\
&= \boldsymbol{\beta}\mathbf{W_i^{(n)}}
\end{aligned}
$$

We assume the following AR(1) correlation structure as before. Below is the distribution of the moving averages.

$$
\boldsymbol{\epsilon}^{(7)} \sim MVN\left(0, \boldsymbol{\Sigma}^{(7)} = \begin{bmatrix}
\sigma^{\mathbf{2}} & \sigma^2\rho & \sigma^2\rho^{\mathbf{2}} & \dots & \sigma^2\rho^{\mathbf{6}} \\
\sigma^2\rho & \sigma^2 & \sigma^2\rho & \dots & \sigma^2\rho^{\mathbf{5}} \\
\sigma^2\rho^{\mathbf{2}} & \sigma^2\rho & \sigma^{\mathbf{2}} & \dots & \sigma^2\rho^{\mathbf{4}} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\sigma^2\rho^{\mathbf{6}} & \sigma^2\rho^{\mathbf{5}} & \sigma^2\rho^{\mathbf{4}} & \dots & \sigma^{\mathbf{2}}
\end{bmatrix}\right),
$$

$$
\begin{aligned}
M_{it}^{(7)} &= \frac{1}{7}\sum_{l=t-6}^{t} X_{il} \sim MVN(\boldsymbol{\beta W_{it}^{(7)}}, \mathbf{L_7 \Sigma^{(7)} L_7^T}) \\
M_{it}^{(6)} &= \frac{1}{6}\sum_{l=t-5}^{t} X_{il} \sim MVN(\boldsymbol{\beta W_{it}^{(6)}}, \mathbf{L_6 \Sigma^{(6)} L_6^T}) \\
&\vdots \\
M_{it}^{(1)} &= \sum_{l=t-0}^{t} X_{it} \sim MVN(\boldsymbol{\beta W_{it}^{(1)}}, \mathbf{L_1 \Sigma^{(1)} L_1^T})
\end{aligned}
$$

Where L and L^T are vectors of scalar values reflecting the moving average. Since we know that $M^{(7)}$ and $M^{(6)}$ are both normally distributed, then the conditional distribution is also normal. To formalize this notion we want the conditional distribution of $M^{(7)}|M^{(6)}$. Given the following general formulas for the conditional normal distribution:

$$
\begin{aligned}
\mu_{i|j} &= \mu_i + \Lambda_{ij}\Lambda_{jj}^{-1}(X_j - \mu_j) && (2.2) \\
\Lambda_{i|j} &= \Lambda_{jj} - \Lambda_{ij}^T\Lambda_{ii}^{-1}\Lambda_{ij} && (2.3)
\end{aligned}
$$

In particular we have the following equations related to our example in matrix notation:

$$
\begin{aligned}
\mu_{M^{(7)}|M^{(6)}} &= \mu_{[7]} + \Lambda_{[7,6]}\Lambda_{[6]}^{-1}(M^{(6)} - \mu_{[6]} \\
\Lambda_{M^{(7)}|M^{(6)}} &= \Lambda_{[6]} - \Lambda_{[7,6]}^T\Lambda_{[7]}^{-1}\Lambda_{[7,6]}
\end{aligned}
$$

As scalars:

$$\begin{aligned}
\mu_{M^{(7)}|M^{(6)}} &= \mu_{[7]} + \frac{\sigma_{[7,6]}}{\sigma^2_{[6]}}(M^{(6)} - \mu_{[6]}) \\
\Lambda_{M^{(7)}|M^{(6)}} &= \sigma^2_{[6]} - \frac{\sigma^2_{[7,6]}}{\sigma^2_{[7]}} = \sigma^2_{[6]}(1-\rho^2)
\end{aligned}$$

Now, we must solve for each of the missing values $\mu_{[7]}$, $\sigma_{[7,6]}$, $\sigma^2_{[6]}$, and $\sigma^2_{[7]}$ using the following equations.

$$\begin{aligned}
\mu_{[n]} &= E(M^{(n)}) = E(\frac{1}{n}\sum_{t=1}^{n} X_{it}) = \boldsymbol{\beta}\mathbf{W_i} \\
\sigma^2_{[n]} &= V(M^{(n)}) = V(\frac{1}{n}\sum_{t=1}^{n} X_{it}) = \frac{1}{n^2}V(\sum_{t=1}^{n} X_{it}) = \frac{1}{n^2}\left[\sum_{t=1}^{n} V(X_{it}) + \sum\sum_{t\neq t'} Cov(X_{it}, X_{it'})\right] \\
&= \frac{\sigma^2}{n^2}\left[\sum_{j=1}^{n}(1) + 2\sum_{k=1}^{n-1}\sum_{j=1}^{k}\rho^j\right] \\
\sigma_{[n,m]} &= Cov(M^{(n)}, M^{(m)}) = \frac{\sigma^2}{nm}\left[\sum_{j=1}^{m}(1) + 2\sum_{k=1}^{m-1}\sum_{j=1}^{k}\rho^j + \sum_{k=1}^{n-m}\sum_{j=k}^{m+k-1}\rho^j\right]
\end{aligned}$$

For the covariance between the 7-day and 6-day moving average, we have the following derivation:

$$
\begin{aligned}
Cov(M^{(7)}, M^{(6)}) &= Cov(\frac{1}{7}\sum_{t=1}^{7} X_{it}, \frac{1}{6}\sum_{t'=2}^{7} X_{it'}) \\
&= \frac{1}{42}\left[\sum_{t=1}^{7}\sum_{t'=2}^{7} Cov(X_{it}, X_{it'})\right] \\
&= \frac{1}{42}\left[\sum_{t=2}^{7}\sum_{t'=2}^{7} Cov(X_{it}, X_{it'}) + \sum_{t'=2}^{7} Cov(X_{i1}, X_{it'})\right] \\
&= \frac{1}{42}\left[V(\sum_{t=2}^{7} X_{it}) + \sum_{j=1}^{6}\sigma^2\rho^j\right] \\
&= \frac{1}{42}\left[\sum_{t=2}^{7} V(X_{it}) + 2 * [\sum_{t=2}^{7}\sum_{t'=2}^{7} Cov(X_{it}, X_{it'})] + \sum_{j=1}^{6}\sigma^2\rho^j\right] \\
&= \frac{1}{42}\left[\sum_{j=1}^{6}\sigma^2 + 2[\sum_{j=1}^{5}\sigma^2\rho^j + \sum_{j=1}^{4}\sigma^2\rho^j + \cdots + \sum_{j=1}^{1}\sigma^2\rho^j] + \sum_{j=1}^{6}\sigma^2\rho^j\right] \\
&= \frac{1}{42}\left[\sum_{j=1}^{6}\sigma^2 + 2\sigma^2\sum_{k=1}^{5}\sum_{j=1}^{k}\rho^j + \sum_{k=1}^{7-6}\sum_{j=1}^{6+1-1}\sigma^2\rho^j\right] \\
&= \frac{\sigma^2}{42}\left[11\rho + 9\rho^2 + 7\rho^3 + 5\rho^4 + 3\rho^5 + 1\rho^6 + 6\right]
\end{aligned}
$$

The rest of the needed values can be derived in the same fashion and their values are given below:

$$
\begin{aligned}
\mu_{[7]} &= \boldsymbol{\beta}\mathbf{W_i} \\
\mu_{[6]} &= \boldsymbol{\beta}\mathbf{W_i} \\
\mu_{[5]} &= \boldsymbol{\beta}\mathbf{W_i} \\
\mu_{[4]} &= \boldsymbol{\beta}\mathbf{W_i} \\
\mu_{[3]} &= \boldsymbol{\beta}\mathbf{W_i} \\
\mu_{[2]} &= \boldsymbol{\beta}\mathbf{W_i} \\
\mu_{[1]} &= \boldsymbol{\beta}\mathbf{W_i} \\
\sigma_{[7,6]} &= \frac{\sigma^2}{42}(11\rho + 9\rho^2 + 7\rho^3 + 5\rho^4 + 3\rho^5 + 1\rho^6 + 6) \\
\sigma_{[7,5]} &= \frac{\sigma^2}{35}(9\rho + 8\rho^2 + 6\rho^3 + 4\rho^4 + 2\rho^5 + 1\rho^6 + 5) \\
\sigma_{[7,4]} &= \frac{\sigma^2}{28}(7\rho + 6\rho^2 + 5\rho^3 + 3\rho^4 + 2\rho^5 + 1\rho^6 + 4) \\
\sigma_{[7,3]} &= \frac{\sigma^2}{21}(5\rho + 4\rho^2 + 3\rho^3 + 3\rho^4 + 2\rho^5 + 1\rho^6 + 3) \\
\sigma_{[7,2]} &= \frac{\sigma^2}{14}(3\rho + 2\rho^2 + 2\rho^3 + 2\rho^4 + 2\rho^5 + 1\rho^6 + 2) \\
\sigma_{[7,1]} &= \frac{\sigma^2}{7}(1\rho + 1\rho^2 + 1\rho^3 + 1\rho^4 + 1\rho^5 + 1\rho^6 + 1) \\
\sigma^2_{[7]} &= \frac{\sigma^2}{49}(12\rho + 10\rho^2 + 8\rho^3 + 6\rho^4 + 4\rho^5 + 2\rho^6 + 7) \\
\sigma^2_{[6]} &= \frac{\sigma^2}{36}(10\rho + 8\rho^2 + 6\rho^3 + 4\rho^4 + 2\rho^5 + 6) \\
\sigma^2_{[5]} &= \frac{\sigma^2}{25}(8\rho + 6\rho^2 + 4\rho^3 + 2\rho^4 + 5) \\
\sigma^2_{[4]} &= \frac{\sigma^2}{16}(6\rho + 4\rho^2 + 2\rho^3 + 4) \\
\sigma^2_{[3]} &= \frac{\sigma^2}{9}(4\rho + 2\rho^2 + 3) \\
\sigma^2_{[2]} &= \frac{\sigma^2}{4}(2\rho + 2) \\
\sigma^2_{[1]} &= \sigma^2
\end{aligned}
$$

Conditional Distribution of the Moving Average

So finally, we know that the conditional distributions for the 7-day moving average are given below where $\mu_{M^{(n)}|M^{(m)}} = \boldsymbol{\beta W^{(n)}} + \frac{\sigma_{[n,m]}}{\sigma^2_{[m]}}(M^{(m)} - \boldsymbol{\beta W^{(n)}})$ and $\Lambda_{M^{(n)}|M^{(m)}} = \sigma^2_{[m]}(1-\rho^2)$ for $n > m$.

$$\begin{aligned}
M^{(7)}|M^{(6)} &\sim MVN(\boldsymbol{\beta W^{(7)}} + \frac{\sigma_{[7,6]}}{\sigma^2_{[6]}}(M^{(6)} - \boldsymbol{\beta W^{(7)}}), \sigma^2_{[6]}(1-\rho^2)) \\
M^{(7)}|M^{(5)} &\sim MVN(\boldsymbol{\beta W^{(6)}} + \frac{\sigma_{[7,5]}}{\sigma^2_{[5]}}(M^{(5)} - \boldsymbol{\beta W^{(6)}}), \sigma^2_{[5]}(1-\rho^2)) \\
&\vdots \\
M^{(7)}|M^{(1)} &\sim MVN(\boldsymbol{\beta W^{(1)}} + \frac{\sigma_{[7,1]}}{\sigma^2_{[1]}}(M^{(1)} - \boldsymbol{\beta W^{(1)}}), \sigma^2_{[1]}(1-\rho^2))
\end{aligned}$$

Given the original mean model, the AR(1) correlation structure assumption, and the resulting conditional distribution we can estimate the missing moving average values from table 2.5 in section 2.3. By running a regression model, estimates for $\hat{\beta}$, $\hat{\rho}$, and $\hat{\sigma}^2$ would be obtained and those resulting values can be used to impute/recreate missing estimates given the observed data. Consequently, a regression analysis can be conducted on the resulting data.

2.4.3 Nonparametric Moving Average Imputation

Initially, we assumed an exposure model, correlation of the errors, and the corresponding exposure distribution. The resulting mean functions were linear in form where each 7-day moving average of exposure was a linear function of the previ-

ous days values. For example take

$$\begin{aligned}
\mu_{M^{(n)}|M^{(m)}} &= \boldsymbol{\beta W^{(n)}} + \frac{\sigma_{[n,m]}}{\sigma^2_{[m]}}(M^{(m)}_{im} - \boldsymbol{\beta W^{(n)}}) \\
&= \boldsymbol{\beta W^{(n)}} + \frac{\sigma_{[n,m]}}{\sigma^2_{[m]}} * M^{(m)}_{im} - \frac{\sigma_{[n,m]}}{\sigma^2_{[m]}} * \boldsymbol{\beta W^{(n)}} \\
&= \underbrace{(\boldsymbol{\beta W^{(n)}} - \frac{\sigma_{[n,m]}}{\sigma^2_{[m]}} * \boldsymbol{\beta W^{(n)}}}_{\alpha^*_0} + \underbrace{\frac{\sigma_{[n,m]}}{\sigma^2_{[m]}}}_{\alpha^*_1} * M^{(m)}_{im} \\
&= \alpha^*_0 + \alpha^*_1 * M^{(m)}_{im}.
\end{aligned}$$

For Imputation method 2, we choose to assume an unspecified linear relationship between the 7-day moving averages and the previous days. This relationship is defined by the estimates α_0 and α_1.

$$\begin{aligned}
M^{(7,r)}_{it} &= \alpha_{06} + \alpha_{16} M^{(6,r)}_{it} \\
M^{(7,r)}_{it} &= \alpha_{05} + \alpha_{15} M^{(5,r)}_{it} \\
M^{(7,r)}_{it} &= \alpha_{04} + \alpha_{14} M^{(4,r)}_{it} \\
M^{(7,r)}_{it} &= \alpha_{03} + \alpha_{13} M^{(3,r)}_{it} \\
M^{(7,r)}_{it} &= \alpha_{02} + \alpha_{12} M^{(2,r)}_{it} \\
M^{(7,r)}_{it} &= \alpha_{01} + \alpha_{11} M^{(1,r)}_{it}
\end{aligned}$$

Where $M^{(nr)}_{it}$ is the moving average from the reduced data set. Once, this simple regression is conducted, the estimates of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ can be extracted and used to them to generate values for the missing moving averages.

$$
\begin{aligned}
\mu_{M^{(7)}|M^{(6)}} &= M_{i6}^{(7)} = \hat{\alpha}_{06} + \hat{\alpha}_{16} M_{i6}^{(6)} \\
\mu_{M^{(7)}|M^{(5)}} &= M_{i5}^{(7)} = \hat{\alpha}_{05} + \hat{\alpha}_{15} M_{i5}^{(5)} \\
\mu_{M^{(7)}|M^{(4)}} &= M_{i4}^{(7)} = \hat{\alpha}_{04} + \hat{\alpha}_{14} M_{i4}^{(4)} \\
\mu_{M^{(7)}|M^{(3)}} &= M_{i3}^{(7)} = \hat{\alpha}_{03} + \hat{\alpha}_{13} M_{i3}^{(3)} \\
\mu_{M^{(7)}|M^{(2)}} &= M_{i2}^{(7)} = \hat{\alpha}_{02} + \hat{\alpha}_{12} M_{i2}^{(2)} \\
\mu_{M^{(7)}|M^{(1)}} &= M_{i1}^{(7)} = \hat{\alpha}_{01} + \hat{\alpha}_{11} M_{i1}^{(1)}
\end{aligned}
$$

The values obtained could replace any absent moving averages that had been deleted in the reduced data. The results are explained in the next section through the use of a comprehensive table.

2.4.4 Data Reduction and Daily Imputation

Many investigators remove those subjects with missing exposure/covariate observations. This will represent the "reduced method". Imputation method $\#3$ represents the conventional means of imputing data in panel studies. Rather than using the the moving average to reproduce the missing moving averages, this calibration strategy seeks to impute each of the individual missing observations and calculate the resulting moving averages to conduct the analyses. This is accomplished by regressing the original outcome data on the original exposure data and using the predicted values to impute each missing value. Once the missing values are computed and the unobserved covariates are then replaced by their predicted values from the calibration model, merged with the existing data set, and regression anal-

ysis can be undertaken. Finally, the standard errors are adjusted to account for the estimation of the unknown covariates. The typical approach is to calculate standard errors using bootstrap or sandwich methods, but asymptotic standard errors are available as well.

$$\begin{aligned}
\hat{X_{i6}} &= \hat{\beta}_0 + \hat{\beta}_1 W_{1i6} \\
\hat{X_{i5}} &= \hat{\beta}_0 + \hat{\beta}_1 W_{1i5} \\
\hat{X_{i4}} &= \hat{\beta}_0 + \hat{\beta}_1 W_{1i4} \\
\hat{X_{i3}} &= \hat{\beta}_0 + \hat{\beta}_1 W_{1i3} \\
\hat{X_{i2}} &= \hat{\beta}_0 + \hat{\beta}_1 W_{1i2} \\
\hat{X_{i1}} &= \hat{\beta}_0 + \hat{\beta}_1 W_{1i1}
\end{aligned}$$

Then we substitute $\hat{X_{i1}}, \ldots, \hat{X_{i6}}$ in for the missing daily values and subsequently construct the moving averages.

2.4.5 Regression Calibration

Each one of the afore mentioned methods represents a form of regression calibration. In a conventional analyses, investigators would like to detail the relationship between a particular response **Y** and its predictors. For simplicity we will distinguish between two different types of predictors **X** and **Z**. **Z** represents those predictors that can be measured without error and **X** represents those that cannot be observed for all subjects. We may be able to observe a variable **W** which is related to **X**. Since the parameters of the model relating **Y** to (**Z**,**X**) cannot be fit-

ted accurately due to the unobservable **X**, hence the surrogate relationship of **Y** on (**Z**,**W**) must be modeled. The structure of the error model relating **X** to **W** is the basis of regression calibration. Once the error model has been developed, the distribution and conditional distributions of **X** can be derived. Lastly, we replace the unobserved exposures **X** by its mean function $m(\mathbf{Z}, \mathbf{W}, \boldsymbol{\beta})$ and run the appropriate standard regression analysis(Carroll, 1995). The following simulation will help to illustrate these ideas.

2.5 SIMULATION STUDY

We conducted a simulation study to examine the effectiveness of the imputation method in estimating missing values in time series data. Upon estimation of the missing values, the next goal was to produce a parameter estimate for the effect of exposure on outcome. The simulations were to be conducted in three phases; 1a) We created three data sets where the first will be a reduced data set where rows of data were deleted due to the missing observations. 1b) The second data set consisted of a full data set where the previously missing values were imputed. 1c) The last data set was simulating true data where no missing values existed or were imputed. Our next step was to run mixed effects models on each of these data sets and compare the effect estimates with the true given values. The intended outcomes, the moving averages, and the initial coefficients needed to be simulated in order to test consistency. We began by simulating exposure data with 30 unique subjects and 20 lagged (24 hour) observations for each.

2.5.1 Simulating The Moving Average Data

We assumed 30 unique "ID"(subject) each had 20 measurements (T=20) taken at consecutive days. The assumed exposure model:

$$X_{it} = \beta_0 + \beta_1 W_{1it} + \epsilon_{it},$$

where X_{it} is the exposure for subject i at time t. W_{1it} represents simulated, ambient pollutant exposure for weekday and weekends for subject i at time t. For example, W_{it} represent ambient pollution on the weekday(WD) and weekend(WE).

$$W_{1it} \sim \begin{cases} N(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_{w_1^2}) & if \quad WD \\ N(\boldsymbol{\mu}_2, \boldsymbol{\sigma}_{w_1^2}) & if \quad WE \end{cases},$$

$\mu_1 = 1$, $\mu_2 = 0$, and $\boldsymbol{\sigma}_{\boldsymbol{w_1^2}}, \boldsymbol{\sigma}_{\boldsymbol{w_2^2}} = 1$. The errors were normally distributed $\epsilon_{it} \sim N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$. There were 3 error correlation structures assumed for simulations: $AR(1)$, $AR(2)$, and $ARMA(1,1)$. To simulate the daily exposure data we used the following distribution: $X \sim N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$. We assumed,

$$\boldsymbol{\Sigma_x} = \begin{bmatrix} \sigma^{\mathbf{2}} & \sigma^2\rho & \sigma^2\rho^{\mathbf{2}} & \dots & \sigma^2\rho^{\mathbf{20}} \\ \sigma^2\rho & \sigma^{\mathbf{2}} & \sigma^2\rho & \dots & \sigma^2\rho^{\mathbf{19}} \\ \sigma^2\rho^{\mathbf{2}} & \sigma^2\rho & \sigma^{\mathbf{2}} & \dots & \sigma^2\rho^{\mathbf{18}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho^{\mathbf{20}} & \sigma^2\rho^{\mathbf{19}} & \sigma^2\rho^{\mathbf{18}} & \dots & \sigma^{\mathbf{2}} \end{bmatrix},$$

where $\rho = 0.4, 0.8$, $\sigma^2 = 0.3, 0.9, 1.5$, and $\mu_{\mathbf{x}}{}^{(n)} = \beta_0 + \beta_1 W_{1it}^{(n)}$. The initial value of β_0 was 0 and β_1 was 1. This structure was used to reflect the fact that exposure measurements taken at closer intervals are more highly correlated. These initial values were taken from averages in the real data.

2.5.2 Simulating Outcome

The linear model of interest could be written in the following way:

$$Y_{it} = \gamma_0 + \gamma_1 M_{it}^{(7)} + b_i + \delta_{it}$$

where $M_{it}^{(7)}$ is the moving average for the i^{th} subject at time t.

- $b_i \sim N(0, \sigma_b^2)$ are the random coefficients due to i^{th} subject.
- $\delta_{it} \sim N(0, \sigma_\delta^2)$ is the error term for the i^{th} subject and time t.
- Initial values were chosen for $\gamma_0 = 0$ and $\gamma_1 = 1$
- Y_{it} was simulated from the following distribution:

$$Y_{it}|M_{it}^{(7)}, \gamma_0, \gamma_1 \sim N(\gamma_0 + \gamma_1 M_{it}^{(7)}, \sigma_{y_1}^2)$$

where $\sigma_{y_{it}}^2 = \sigma_{b_i}^2 + \sigma_{\delta_{it}}^2$.

The simulations were to be conducted in three phases where 5 data sets of exposure were created. First, a reduced data set where rows of data were deleted due to the missing observations. In this case, there would be 30 unique subjects

at 7 time points (210 observations) remaining. This clearly decreases the amount of exposure data available for modeling. Second, 3 imputed data sets where the previously missing moving averages/values were imputed where the parametric moving average imputation (PMA), the nonparametric moving average imputation (NPMA), and the daily value imputation (DIMA) are conducted. This will amount to 30 unique subjects at 14 time points (420 observations) consisting of 180 imputed values for the 7-day moving average and 120 imputed values for the 4-day moving average. Lastly, we simulated true data "gold standard/truth" where no missing values existed or were imputed. This will be represented by a fully simulated data set where there are no missing values and the moving averages can be calculated for each observation. This corresponds to 30 unique subjects at 14 time points (420 observations). In total, this amounts to each subject having 14 days of available exposure information to be used in the modeling process.

Our next step was to run mixed effects models on each of these data sets and compare the effect estimates with the true given values. The mixed effects models conducted:

$$
\begin{aligned}
Y_{it} &= \gamma_0 + \gamma_1 M_{it}^{(7,reduced)} + b_i + \delta_{it} \\
Y_{it} &= \gamma_0 + \gamma_1 M_{it}^{(7,PMA)} + b_i + \delta_{it} \\
Y_{it} &= \gamma_0 + \gamma_1 M_{it}^{(7,NPMA)} + b_i + \delta_{it} \\
Y_{it} &= \gamma_0 + \gamma_1 M_{it}^{(7,DIMA)} + b_i + \delta_{it} \\
Y_{it} &= \gamma_0 + \gamma_1 M_{it}^{(7,truth)} + b_i + \delta_{it}
\end{aligned}
$$

These models were simulated for 200 iterations and no covariates were included.

The results are explained in section 4.5 through the use of a comprehensive table.

2.5.3 Relaxing the AR(1) Assumption

The conditional distributions are accurate under the given assumptions, but when the AR(1) covariance structure is relaxed new derivations must be developed. The AR(1) or ARIMA(1,0,0) covariance structure is a simplest case of ARIMA models. ARIMA models are, in theory, they are the most general class of models for forecasting time series. It is of interest to determine whether the three afore mentioned methods of imputation are robust when dealing with new correlation structures and hence model mis-specification. In order to investigate this issue, alternative correlation structures for simulating exposure data were constructed. The correlation structures of interest were AR(2) and ARMA(1,1). Data was simulated from these two distributions respectively and subsequently imputation methods were performed and compared.

Simulating ARMA(1,1) Exposure Data

Suppose that the errors followed a different correlation pattern had the following ARIMA(1,0,1) or ARMA(1,1) covariance structure. In order to simulate the exposure data we first need to derive the variance covariance matrix. This was achieved using the general equations for producing ARMA(p, q) errors in closed form by Jan van der Leeuw where p are the autoregressive terms and q are the moving average terms. Assume $p = 1$ and $q = 1$ for ARMA(1,1).

$$V = [NM][\bar{P}^T\bar{P} - \bar{Q}\bar{Q}^T]^{-1}[NM]^T$$

where $\bar{P}, \bar{Q}$,M,N are well-defined toeplitz matrices of the following form:

$$\bar{P}_{[21\times 21]} = \left[\begin{array}{c|c} \bar{P}_1 & 0 \\ \hline \bar{P}_2 & \bar{P}_3 \end{array}\right] = \left[\begin{array}{c|cccc} 1 & 0 & 0 & \dots & 0 \\ \hline \rho_1 & 1 & 0 & \dots & 0 \\ 0 & \rho_1 & 1 & \dots & 0 \\ 0 & 0 & \rho_1 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ & & & & \end{array}\right],$$

$$\bar{Q}_{[21\times 1]} = \left[\begin{array}{c} \bar{Q}_1 \\ \hline 0 \end{array}\right] = \left[\begin{array}{c} \rho_1 \\ \hline 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{array}\right],$$

$$\bar{M}_{[20\times 20]} = \left[\begin{array}{c|c} \bar{M}_1 & 0 \\ \hline \bar{M}_2 & \bar{M}_3 \end{array}\right] = \left[\begin{array}{c|cccc} 1 & 0 & 0 & \dots & 0 \\ \hline \theta_1 & 1 & 0 & \dots & 0 \\ 0 & \theta_1 & 1 & \dots & 0 \\ 0 & 0 & \theta_1 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ & & & & \end{array}\right],$$

$$\bar{N}_{[20\times 1]} = \left[\begin{array}{c} \bar{N}_1 \\ \hline 0 \end{array}\right] = \left[\begin{array}{c} \theta_1 \\ \hline 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{array}\right],$$

We define P to be a (square) $T \times T$ lower band matrix where T (T=20) represents the total number of days being simulated. Q is $T \times p$ and is partitioned into an upper $p \times p$ part and a lower $(T-p) \times p$ part, which consists of only zeros. M and N have the same structure as P and Q with r_i replaced by θ and p replaced by q. $\bar{P}$ is the same as P but the dimensions are $(T+p) \times (T+p)$ while $\bar{Q}$ has dimensions $(T+p) \times p$. Performing these matrix calculations result in the correlation matrix for ARMA($p = 1, q = 1$) errors [van der Leeuw,1994]. Initial values for ρ_1 and θ_1 can be chosen according to previous analyses or subject specific estimates. The final form of the ARMA(1,1) correlation structure is $\rho_{\mathbf{X}}$ below.

$$\rho_{\mathbf{X}} = \begin{bmatrix} 1 & \theta & \theta\rho^{\mathbf{1}} & \dots & \theta\rho^{\mathbf{18}} \\ \theta & 1 & \theta & \dots & \theta\rho^{\mathbf{17}} \\ \theta\rho^{\mathbf{1}} & \theta & 1 & \dots & \theta\rho^{\mathbf{16}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta\rho^{\mathbf{18}} & \theta\rho^{\mathbf{17}} & \theta\rho^{\mathbf{16}} & \dots & 1 \end{bmatrix},$$

The corresponding covariance matrix will have the following form:

$$\epsilon_{(\mathbf{X})} \sim MVN\left(0, \mathbf{\Lambda}_{(\mathbf{X})} = \begin{bmatrix} \sigma^{\mathbf{2}} & \sigma^2\theta & \sigma^2\theta\rho^{\mathbf{1}} & \dots & \sigma^2\theta\rho^{\mathbf{18}} \\ \sigma^2\theta & \sigma^{\mathbf{2}} & \sigma^2\theta & \dots & \sigma^2\theta\rho^{\mathbf{17}} \\ \sigma^2\theta\rho^{\mathbf{1}} & \sigma^2\theta & \sigma^{\mathbf{2}} & \dots & \sigma^2\theta\rho^{\mathbf{16}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma^2\theta\rho^{\mathbf{18}} & \sigma^2\theta\rho^{\mathbf{17}} & \sigma^2\theta\rho^{\mathbf{16}} & \dots & \sigma^{\mathbf{2}} \end{bmatrix}\right),$$

To simulate the daily exposure data we used the following distribution: $\mathbf{X} \sim N(\boldsymbol{\mu}_x, \mathbf{\Lambda}_{\mathbf{x_{AR(2)}}})$. Now that the exposure data has been simulated the outcome data can be simulated in the same fashion as "Simulating Outcome".

Simulating AR(2) Exposure Data

Assume $p = 2$ for the autoregressive coefficient. For the ARIMA(p,0,0) or AR(p) cases the covariance structure equations reduce to the following formula.

$$V = [P^T P - QQ^T]^{-1}$$

where P,Q are well-defined toeplitz matrices as before but are of different dimensions of the following form:

$$P_{[20\times 20]} = \left[\begin{array}{c|c} P_1 & 0 \\ \hline P_2 & P_3 \end{array}\right] = \left[\begin{array}{cc|ccc} 1 & 0 & 0 & \dots & 0 \\ \rho_1 & 1 & 0 & \dots & 0 \\ \hline \rho_2 & \rho_1 & 1 & \dots & 0 \\ 0 & \rho_2 & \rho_1 & \dots & 0 \\ 0 & 0 & \rho_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ & & & & \end{array}\right],$$

$$Q_{[20\times 2]} = \left[\begin{array}{c} Q_1 \\ \hline 0 \end{array}\right] = \left[\begin{array}{cc} \rho_2 & \rho_1 \\ 0 & \rho_2 \\ \hline 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{array}\right],$$

The variance covariance matrix for AR(2) does not have a discernable pattern like the others therefore the elements of the correlation structure will be left unspeci-

fied.

$$\epsilon^{(7)} \sim MVN\left(0, \Lambda^{(7)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{17} \\ a_{21} & a_{22} & a_{23} & \dots & a_{27} \\ a_{31} & a_{32} & a_{33} & \dots & a_{37} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{71} & a_{72} & a_{73} & \dots & a_{77} \end{bmatrix}\right),$$

$$Cov(M^{(n)}, M^{(m)}) = \frac{1}{nm}\left[\sum_{t=n-m}^{7} V(X_{it}) + 2[\sum_{t=n-m}^{7}\sum_{t'=n-m}^{7} Cov(X_{it}, X_{it'})] + \sum_{t'=n-m}^{7} Cov(X_{i1}, X_{it'})\right]$$

where $n > m$. So for the covariance between $M^{(7)}$ and $M^{(6)}$ we would have:

$$\begin{aligned} Cov(M^{(7)}, M^{(6)}) &= \frac{1}{42}\left[\sum_{t=2}^{7} V(X_{it}) + 2[\sum_{t=2}^{7}\sum_{t'=2}^{7} Cov(X_{it}, X_{it'})] + \sum_{t'=2}^{7} Cov(X_{i1}, X_{it'})\right] \\ &= \frac{1}{42}\left[\sum_{j=1}^{7} a_{jj} + 2[\sum^{7}\sum_{j>k}^{7} a_{jk}] + \sum_{j=2}^{7} a_{j1}\right] \end{aligned}$$

2.5.4 Simulation Results

We conducted a simulation study to examine the effectiveness of three regression calibration methods of estimation for missing data. We would also like to perform a direct comparison of the point estimates between each of the relevant methods through calculation of the mean squared errors. The initial values for the corre-

lation coefficients and variances must be simulated under varying assumptions in order to get a complete picture of the regression calibration effectiveness. We began with 30 unique subjects with 20 observations taken on 20 consecutive days. The 20 lagged time-points were created to simulate 3 weeks of exposure data. Table 6 represents the initial values chosen for the simulation study. For each correlation structure, AR(1), AR(2), and ARMA(1,1), the correlation coefficient, moving average coefficient, and variance were needed. Each scenario produced 5 data sets and linear mixed models were run all assuming AR(1) error structures. We adjusted the resulting standard errors to account for the estimation of γ by using the bootstrap method. Both 7-day and 4-day moving averages were conducted over the 20 day period. Under the 7-day moving average 6 conditional exposure measurements must estimated per subject while in 4-day moving averages only 3 must be estimated per subject.

AR(1) Simulation Results

We found that when the model was properly specified meaning that the simulated data was in accord with linear mixed model, all the point estimates among all data sets are very close to the true values of $\gamma_0 = 0$ and $\gamma_1 = 1$. This shows that there was little bias found in the well specified cases although there were clear differences in efficiency. Model efficiency was determined by the mean squared error. Using the first regression calibration method(PMA) yielded the lowest MSE as can be seen in Table 2.7. The models that used the reduced had the highest MSE values and hence the lowest efficiency because the standard errors and standard deviations were the highest. Using PMA increased efficiency by approximately $25\% - 40\%$ over the reduced method depending on the initial values for σ^2. Further, PMA was more

efficient than the other two calibration methods (NPMA and DIMA) but by smaller margins. PMA tends to be around $5\% - 10\%$ more efficient than the other NPMA methods when the model is properly specified. For example, in the simulation scenario for AR(1) where$\phi = .8$ and $\sigma^2 = .9$ for the 4-day moving average we have the following output which can be seen in Table 2.7. All of the γ_1 point estimates are quite close to the simulated starting value of 1. But looking at the MSE of $\hat{\gamma}_{1,reduced}$ = 0.019 while for $\hat{\gamma}_{1,PMA}$ = 0.009. This reflects more than a50% decrease in the MSE which means the PMA is at an advantage above deleting incomplete data. When comparing PMA to conventional regression calibration method 3 (DIMA), there is no reduction in MSE when moving from $\hat{\gamma}_{1,reduced}$ = 0.009 to $\hat{\gamma}_{1,DIMA}$ = 0.009 for $\sigma^2 = .9$ but for $\sigma^2 = 1.5$ and $\sigma^2 = 2.0$ there are small gains in efficiency. These patterns are consistent throughout all of the proposed AR(1) cases.

AR(2) and ARMA(1,1) Simulation Results

We found that when the model was mis-specified, there were new patterns revealed. There were two cases of mis-specification that were investigated. The first was simulating data with an AR(2) error correlation structure but modeled with an AR(1). The second was simulating data with an ARMA(1,1) error correlation structure but modeled with and AR(1). Each of the these cases are different representations of the larger class of ARIMA models with small parameter shifts.

PMA and NPMA methods performed with very little bias in most cases while the conventional methodology used in DIMA experienced high levels of bias. Small deviations from AR(1) were examined such as AR(2) with $\phi_1 = .8$,$\phi_2 = .1$ and ARMA(1,1) with$\phi_1 = .8$,$\theta_2 = .2$ each of which were done with both 4 and 7 day moving averages. These cases reflect marginal departures from AR(1) error

structure but elicit large responses in bias of the $\hat{\gamma}_1$ estimates. Table 7 shows that the conventional DIMA for AR(2) has the following values for $\gamma_{\mathbf{AR(2)}}$: $\hat{\gamma}_{1,DIMA}$ = 0.946(0.053), 0.943(0.022), 0.924(0.020), 0.927(0.015) for $\sigma^2 = .3, .9, 1.5, 2.0$ respectively. Each point estimate underestimates the simulated value of 1 by greater than 5%. This trend is echoed in the ARMA(1,1) case for DIMA. The values for $\gamma_{\mathbf{ARMA(1,1)}}$ are as follows: $\hat{\gamma}_{1,DIMA}$ = 0.9478(0.0495), 0.9359(0.0215), 0.9291(0.0180), 0.9347(0.0161) for $\sigma^2 = .3, .9, 1.5, 2.0$ respectively which also underestimate the simulated value of 1 by $> 5\%$. On the other hand, the PMA and NPMA methods show very low bias in the point estimates as they are all very close to 1. The MSE's are also lower for PMA and NPMA methods than for the DIMA method for increasing variability. For example, when $\sigma^2 = 2.0$ the $MSE_{AR(2)}(DIMA) = 0.015$ while $MSE_{AR(2)}(PMA) = 0.010$. Alternatively for $\sigma^2 = .3$ the $MSE_{AR(2)}(DIMA) = 0.053$ while the $MSE_{AR(2)}(PMA) = 0.055$.

Large deviations from AR(1) were explored as well AR(2) with $\phi_1 = .8$,$\phi_2 = .5$ and ARMA(1,1), $\phi_1 = .8$,$\theta_2 = .5$ and findings are similar. These cases reflect larger departures from AR(1) error structure but elicit smaller responses in bias. Table 8 shows that the conventional DIMA method for AR(2) has the following values for $\gamma_{\mathbf{AR(2)}}$: $\hat{\gamma}_{1,DIMA}$ = 0.992 (0.046), 0.969(0.013), 0.964(0.014), 0.964(0.011) for $\sigma^2 = .9, 1.5, 2.0$ respectively. Each point estimate underestimates the simulated value of 1 by greater than 3% accept when $\sigma^2 = 2$. This trend is echoed in the ARMA(1,1) case for the DIMA method. The values for $\gamma_{\mathbf{ARMA(1,1)}}$ are as follows: $\hat{\gamma}_{1,DIMA}$ = 0.961 (0.045),0.973(0.0190), 0.964(0.014), 0.972(0.009) for $\sigma^2 = .9, 1.5, 2.0$ respectively which also underestimate the simulated value of 1 by $> 5\%$. On the other hand PMA and NPMA methods show very low bias in the point estimates as they are all very close to 1.

2.6 DISCUSSION

These results show that both the parametric and nonparametric imputations using moving averages yield more robust health effects estimates than the reduced and daily value imputation methods. This is the trend in all of the simulations regardless of the correlation structure chosen. Further, the simulations show that the daily value imputation methods (DIMA) experience higher levels of bias when the models are mis-specified. This means that in the circumstance where the correlations structures are unknown and are very different than AR(1), the DIMA approach may have clear issues with accuracy. On the other hand, the moving average imputation methods (PMA and NPMA) maintain high accuracy and efficiency in comparison with the gold standard (truth). Our methods were conducted on 7-day moving averages which required 6 imputed moving averages to be estimated while the 4-day moving averages required 3 moving average estimates. The same patterns were maintained irrespective of the length of the imputation. We also attempted the bootstrap method to calculate the standard errors but due to the small sample sizes in each panel, the values were higher than normal.

The reason for the differences in both bias an efficiency are a result of the PMA and NPMA approaches consideration of specific correlation structures of the exposure measurements. The DIMA imputation methods assume a linear relationship between the exposure measurements at different times and use estimated predicted values to estimate missing exposure data. On the other hand, the PMA and NPMA methods use correlations that allow the methods to draw upon pollution information from previous days in order to estimate the missing moving averages. With the inclusion of the assumed error structures for the exposures, the estimates from the health effects model are closer to the best linear unbiased predictor (BLUP).

2.7 APPENDIX

γ Estimate(MSE), MA4 Continuous Covariate				
AR(1), $\phi = .8$	$\sigma^2 = .3$	$\sigma^2 = .9$	$\sigma^2 = 1.5$	$\sigma^2 = 2.0$
AR(1), $\phi = .4$	$\sigma^2 = .3$	$\sigma^2 = .9$	$\sigma^2 = 1.5$	$\sigma^2 = 2.0$
AR(2), $\phi_1 = .4$,$\phi_2 = .2$	$\sigma^2 = .3$	$\sigma^2 = .9$	$\sigma^2 = 1.5$	$\sigma^2 = 2.0$
AR(2), $\phi_1 = .8$,$\phi_2 = .1$	$\sigma^2 = .3$	$\sigma^2 = .9$	$\sigma^2 = 1.5$	$\sigma^2 = 2.0$
AR(2), $\phi_1 = .8$,$\phi_2 = .4$	$\sigma^2 = .3$	$\sigma^2 = .9$	$\sigma^2 = 1.5$	$\sigma^2 = 2.0$
AR(2), $\phi_1 = .8$,$\phi_2 = .5$	$\sigma^2 = .3$	$\sigma^2 = .9$	$\sigma^2 = 1.5$	$\sigma^2 = 2.0$
ARMA(1,1), $\phi_1 = .4$,$\theta_2 = .2$	$\sigma^2 = .3$	$\sigma^2 = .9$	$\sigma^2 = 1.5$	$\sigma^2 = 2.0$
ARMA(1,1), $\phi_1 = .8$,$\theta_2 = .2$	$\sigma^2 = .3$	$\sigma^2 = .9$	$\sigma^2 = 1.5$	$\sigma^2 = 2.0$
ARMA(1,1), $\phi_1 = .8$,$\theta_2 = .5$	$\sigma^2 = .3$	$\sigma^2 = .9$	$\sigma^2 = 1.5$	$\sigma^2 = 2.0$

Table 2.6: 30 Unique IDs - This table represents the initial values chosen for the simulation study. For each correlation structure, AR(1), AR(2), and ARMA(1,1), the correlation coefficient, moving average coefficient, and variance were needed. Each scenario produced 5 data sets.

γ_1 Estimate(MSE), MA4 Continuous Covariate				
	$\sigma^2 = .3$	$\sigma^2 = .9$	$\sigma^2 = 1.5$	$\sigma^2 = 2.0$
AR(1), $\phi = .8$				
$\hat{\gamma}_{1,reduced}$	0.966 (0.051)	0.986 (0.019)	0.998 (0.017)	0.986 (0.008)
$\hat{\gamma}_{1,PMA}$	0.992 (0.023)	0.997 (0.009)	1.009 (0.011)	0.995 (0.004)
$\hat{\gamma}_{1,NPMA}$	0.996 (0.025)	0.996 (0.009)	1.008 (0.011)	0.994 (0.004)
$\hat{\gamma}_{1,DIMA}$	1.005 (0.025)	1.006 (0.009)	1.019 (0.015)	1.006 (0.005)
$\hat{\gamma}_{1,true}$	0.988 (0.020)	0.992 (0.008)	1.006 (0.006)	0.996 (0.004)
AR(2), $\phi_1 = .8$,$\phi_2 = .1$				
$\hat{\gamma}_{1,reduced}$	0.986 (0.099)	1.026 (0.040)	0.992 (0.029)	1.001 (0.018)
$\hat{\gamma}_{1,PMA}$	0.999 (0.055)	1.019 (0.023)	0.993 (0.015)	1.003 (0.010)
$\hat{\gamma}_{1,NPMA}$	1.021 (0.058)	1.030 (0.023)	1.001 (0.017)	1.007 (0.011)
$\hat{\gamma}_{1,DIMA}$	0.946 (0.053)	0.943 (0.022)	0.924 (0.020)	0.927 (0.015)
$\hat{\gamma}_{1,true}$	1.005 (0.049)	1.023 (0.020)	0.994 (0.012)	1.0030 (0.089)
ARMA(1,1), $\phi_1 = .8$,$\theta_2 = .2$				
$\hat{\gamma}_{1,reduced}$	0.972 (0.099)	0.997 (0.036)	1.009 (0.025)	1.020 (0.022)
$\hat{\gamma}_{1,PMA}$	0.993 (0.049)	1.001 (0.020)	1.003 (0.014)	1.010 (0.014)
$\hat{\gamma}_{1,NPMA}$	1.013 (0.056)	1.007 (0.021)	1.008 (0.015)	1.015 (0.014)
$\hat{\gamma}_{1,DIMA}$	0.948 (0.050)	0.936 (0.022)	0.929 (0.018)	0.935 (0.016)
$\hat{\gamma}_{1,true}$	0.998 (0.046)	1.000 (0.017)	1.006 (0.012)	1.011 (0.012)

Table 2.7: Small Deviations from AR(1) for 4-day Moving Average - This table represents the parameter estimates for a 4-day moving average from 5 separate linear mixed model using simulated data from the reduced, imputed, and true data sets. Linear mixed mean models were conducted for 200 iterations. The subsequent estimates were aggregated into means with accompanying MSE. Each model included one continuous covariate for weekend. All of the simulation standard errors are < 0.019.

γ_1 Estimate(MSE), MA4 Continuous Covariate				
	$\sigma^2 = .3$	$\sigma^2 = .9$	$\sigma^2 = 1.5$	$\sigma^2 = 2.0$
AR(2), $\phi_1 = .8$,$\phi_2 = .5$				
$\hat{\gamma}_{1,reduced}$	1.004 (0.072)	1.021 (0.021)	1.001 (0.022)	0.992 (0.018)
$\hat{\gamma}_{1,PMA}$	1.002 (0.047)	1.006 (0.013)	0.989 (0.014)	0.999 (0.010)
$\hat{\gamma}_{1,NPMA}$	1.032 (0.051)	1.024 (0.014)	0.997 (0.013)	1.016 (0.011)
$\hat{\gamma}_{1,DIMA}$	0.992 (0.046)	0.969 (0.013)	0.964 (0.014)	0.964 (0.011)
$\hat{\gamma}_{1,true}$	1.007 (0.040)	1.012 (0.017)	0.991 (0.011)	1.005 (0.009)
ARMA(1,1), $\phi_1 = .8$,$\theta_2 = .5$				
$\hat{\gamma}_{1,reduced}$	0.939 (0.084)	1.001 (0.034)	1.001 (0.022)	0.998 (0.019)
$\hat{\gamma}_{1,PMA}$	0.980 (0.046)	1.003 (0.019)	0.989 (0.014)	0.997 (0.009)
$\hat{\gamma}_{1,NPMA}$	0.992 (0.050)	1.007 (0.020)	0.997 (0.013)	0.999 (0.009)
$\hat{\gamma}_{1,DIMA}$	0.961 (0.045)	0.973 (0.0190)	0.964 (0.014)	0.972 (0.009)
$\hat{\gamma}_{1,true}$	0.980 (0.042)	1.006 (0.016)	0.991 (0.011)	0.991 (0.008)

Table 2.8: Large Deviations From AR(1) for 4-day Moving Average - This table represents the parameter estimates for a 4-day moving average from 5 separate linear mixed model using simulated data from the reduced, imputed, and true data sets. Linear mixed mean models were conducted for 200 iterations. The subsequent estimates were aggregated into means with accompanying MSE. Each model included one continuous covariate for weekend. All of the simulation standard errors are < 0.019.

γ_1 Estimate(MSE), MA7 Continuous Covariate				
	$\sigma^2 = .3$	$\sigma^2 = .9$	$\sigma^2 = 1.5$	$\sigma^2 = 2.0$
AR(1), $\phi = .8$				
$\hat{\gamma}_{1,reduced}$	1.013 (0.082)	0.989 (0.030)	1.008 (0.011)	0.988 (0.013)
$\hat{\gamma}_{1,PMA}$	1.015 (0.042)	0.997 (0.017)	1.020 (0.004)	1.016 (0.009)
$\hat{\gamma}_{1,NPMA}$	0.986 (0.046)	0.987 (0.019)	1.011 (0.004)	1.006 (0.009)
$\hat{\gamma}_{1,DIMA}$	1.016 (0.052)	0.989 (0.023)	1.004 (0.005)	1.007 (0.013)
$\hat{\gamma}_{1,true}$	1.015 (0.030)	0.988 (0.010)	1.004 (0.004)	1.001 (0.005)
AR(2), $\phi_1 = .8$,$\phi_2 = .1$				
$\hat{\gamma}_{1,reduced}$	1.015 (0.111)	1.026 (0.040)	1.015 (0.071)	0.998 (0.049)
$\hat{\gamma}_{1,PMA}$	1.025 (0.070)	1.019 (0.023)	1.010 (0.041)	1.003 (0.033)
$\hat{\gamma}_{1,NPMA}$	1.048 (0.084)	1.030 (0.023)	1.0289(0.0450)	1.0151(0.0336)
$\hat{\gamma}_{1,DIMA}$	0.964 (0.064)	0.943 (0.022)	0.942 (0.039)	0.935 (0.035)
$\hat{\gamma}_{1,true}$	1.021 (0.057)	1.023 (0.020)	0.996 (0.030)	0.997 (0.022)
ARMA(1,1), $\phi_1 = .8$,$\theta_2 = .2$				
$\hat{\gamma}_{1,reduced}$	0.948 (0.260)	0.958 (0.095)	1.015 (0.051)	1.026 (0.035)
$\hat{\gamma}_{1,PMA}$	0.978 (0.134)	0.989 (0.063)	1.024 (0.040)	1.037 (0.026)
$\hat{\gamma}_{1,NPMA}$	0.995 (0.163)	0.996 (0.063)	1.028 (0.041)	1.037 (0.026)
$\hat{\gamma}_{1,DIMA}$	0.946 (0.124)	0.945 (0.061)	0.973 (0.038)	0.984 (0.023)
$\hat{\gamma}_{1,true}$	0.971 (0.109)	0.984 (0.046)	1.007 (0.023)	1.008 (0.018)

Table 2.9: Small Deviations From AR(1) for 7-day Moving Average - This table represents the parameter estimates for a 7-day moving average from 5 separate linear mixed model using simulated data from the reduced, imputed, and true data sets. Linear mixed mean models were conducted for 200 iterations. The subsequent estimates were aggregated into means with accompanying MSE. Each model included one continuous covariate for weekend. All of the simulation standard errors are < 0.036.

γ_1 Estimate(MSE), MA7 Continuous Covariate				
	$\sigma^2 = .3$	$\sigma^2 = .9$	$\sigma^2 = 1.5$	$\sigma^2 = 2.0$
AR(2), $\phi_1 = .8$,$\phi_2 = .5$				
$\hat{\gamma}_{1,reduced}$	1.027 (0.272)	0.985 (0.121)	1.017 (0.076)	0.994 (0.051)
$\hat{\gamma}_{1,PMA}$	1.037 (0.168)	0.950 (0.089)	1.001 (0.047)	0.977 (0.031)
$\hat{\gamma}_{1,NPMA}$	1.089 (0.232)	1.005 (0.097)	1.051 (0.059)	1.019 (0.034)
$\hat{\gamma}_{1,DIMA}$	0.996 (0.156)	0.903 (0.089)	0.949 (0.045)	0.924 (0.034)
$\hat{\gamma}_{1,true}$	1.058 (0.151)	0.968 (0.059)	1.022 (0.040)	0.995 (0.022)
ARMA(1,1), $\phi_1 = .8$,$\theta_2 = .5$				
$\hat{\gamma}_{1,reduced}$	1.036 (0.214)	0.983 (0.074)	1.006 (0.050)	0.991 (0.037)
$\hat{\gamma}_{1,PMA}$	1.051 (0.126)	0.998 (0.044)	1.010 (0.035)	0.999 (0.025)
$\hat{\gamma}_{1,NPMA}$	1.051 (0.134)	0.999(0.047)	1.011(0.033)	0.999(0.026)
$\hat{\gamma}_{1,DIMA}$	1.050 (0.119)	0.976(0.043)	0.986(0.034)	0.974(0.024)
$\hat{\gamma}_{1,true}$	1.018 (0.099)	1.001(0.033)	0.996(0.025)	0.988(0.015)

Table 2.10: Large Deviations From AR(1) for 7-day Moving Average - This table represents the parameter estimates for a 7-day moving average from 5 separate linear mixed model using simulated data from the reduced, imputed, and true data sets. Linear mixed mean models were conducted for 200 iterations. The subsequent estimates were aggregated into means with accompanying MSE. Each model included one continuous covariate for weekend. All of the simulation standard errors are < 0.036.

Health Effects of Multipollutant Mixtures: Testing Properties of Source Apportionment and Two-Stage Hierarchical Regression Methods

[1]Alfa I. Yansané, [2,3]Diane R. Gold, [1,4]Paul J. Catalano,[1] Brent A. Coull

[1]Department of Biostatistics, Harvard School of Public Health

[2]Department of Environmental Health, Harvard School of Public Health

[3]Department of Medicine, Brigham and Women's Hospital/Harvard Medical School and

[4]Department of Biostatistics, Dana-Farber Cancer Institute

3.1 ABSTRACT

One of the core questions in environmental health and pollution research is to identify the health effects associated with specific pollution sources and their constituents. The requirements of such studies include determining air pollution emissions, ambient concentrations by pollution type and particle composition, and the associated health impacts. Since most pollution studies are unable to directly observe the pollution contributions of specific sources, determining the source specific health risk can be difficult. Conventional approaches such as, source apportionment, principle components analysis, and two-stage hierarchical regression have been widely used in the analysis of multipollutant mixtures. Little work has been done to evaluate the appropriate use of each method and characterizing the premier approach has not been resolved. The purpose of this article is to develop a simulation study that compares the source apportionment and two-stage hierarchical regression methods by detailing the power and type 1 errors of the related health affects.

3.2 INTRODUCTION

Exposure to air pollutants has been linked with adverse health outcomes, including increased premature mortality and morbidity, respiratory and cardiovascular disease, and increases in hospital admissions (Dockery et al., 1993; Pope et al., 1999; Brooks et al., 2004; Zanobetti et al., 2004). Recent studies have shown that these outcomes differ according to specific pollutant mixtures, sources of pollution, and particle composition. Given these concerns, recent studies have attempted to examine the risks associated with pollutant mixtures as a whole as opposed to single pollutants. Analyses from Brook et al. 2009 shows that health effects can vary by air pollution mixtures. Alternatively, single pollutant approaches do not adequately estimate cumulative or joint effects of multiple pollutants. In addition to exploration of pollutant mixtures, it is of interest to understand the relative toxicity of individual pollutants to identify the toxic sources. These are important scientific goals, but the methods used to accomplish these aims have not been carefully vetted. Conventional strategies for analyzing such data include: 1) fitting a full model that contains a collection of individual pollutant concentrations; 2) using stepwise model selection; and 3) conducting a number of separate models each containing a single exposure. Each of these approaches do not provide satisfactory solutions to the multiple, correlated exposure problem because of multiple testing issues and the fact that they do not address pollution mixtures (Witte et al., 1996; Momoli et al., 2009).

Two methods have primarily been used in the estimation of effects from multiple correlation exposures. The first approach is source apportionment. Most PM health studies do not directly observe the contributions of the specific pollution sources. Given the knowledge of the chemical characteristics of known sources,

investigators infer pollution source contributions via a source apportionment or multivariate receptor analysis (Nikolov et al., 2006). These approaches typically begin by sampling pollution composition and inferring the likely pollution sources by matching common chemical and physical characteristics between source and air pollution samples. These methods have been shown to effectively quantify the relative contribution of the different sources to ambient air pollution. In aggregate, source apportionment begins by ambient sampling of the concentrations of individual PM constituents. Second, investigators must conduct source profiling where each of the appropriate chemical constituents (receptors/markers) are grouped according to emission source. Third, based on the source profiling one can construct estimates of the contributions to ambient pollution levels from each identified source. This approach is advantageous if the pollutants responsible for the health effects are emitted from only one source. If the health effect is associated with a single element, grouping elements with varying toxicities into a single source can attenuate the health effect (Suh et al., 2011).

The second approach is the use of a hierarchical regression model. Research has shown that hierarchical models outperform conventional regression approaches such as multiple linear regression with multiple exposures, when analyzing epidemiologic data on multiple exposures (Witte et al., 2000; Thomas et al., 2007). The models attempt to measure the relationship between the health outcome and exposure when the exposure variables have meaningful structures or groupings (Witte et al., 1998; Young et al., 2008). For example, rather than estimating independent effects of each pollution constituent separately, this approach seeks to estimate the association between a health outcome and groups of elements that might be defined by chemical properties or other characteristics of the individual exposures. At the first stage, the model contains all of the elements/pollution constituents

as covariates so there is no need for preliminary variable selection. The approach then relates the independent effects of individual pollution constituents to characteristics of these individual exposures in a second stage regression. Therefore the hypotheses that risks differ by pollutant chemical property could be tested (Suh et al., 2011).

The purpose of this paper is to compare the statistical performance of the two approaches by conducting a simulation study. Several existing studies have used one of these approaches to analyze multi-pollutant health effects. (Lall et al., 2011; Ito et al.; Laden et al, 2000). Alternatively, the hierarchical regression approach has been used by Suh et al, 2011. None of these studies have sought to compare the testing performances of these methods in various time series settings. For our analyses, each method will be constructed under varying initial settings established by previous study estimates (Hopke et al., 2006). For the source apportionment, an exploratory factor analysis was applied to the exposure data collected at Harvard School of Public Health to get estimates of the source contributions (Nikolov et al., 2006). For the hierarchical regression models, the choice of second stage covariates were also pre-specified. The power calculations and type I errors will be reported and analyzed. Lastly, each approach will be compared using simulated data to determine the appropriateness of its use.

The remainder of this paper is as follows: Section 3.3 reviews the pollution study data and the experimental design. Section 3.4 describes factor analysis for times series data and then presents the two-stage hierarchical modeling approach. Section 3.5 details the simulation study to examine the statistical properties of the health effects estimates obtained from the two-stage and factor analysis approaches along with the corresponding results. Section 3.6 discusses the findings and the potential

implications of the research.

3.3 DATA AND STUDY DESIGN

Investigators often use time series methods and data to describe the relationship between pollutants and health outcomes. In this study design the day-to-day variations in elemental concentrations are employed to identify the sources of pollution and are correlated with daily mortality or hospital admission counts. Time series methods also allow for time varying confounders to be included in the analyses through adjustments for temperature, humidity, year, season, and days of the week. The daily sampling schedule provides greater power and allows the investigation of distributed lag effects that may not be possible in other analyses (Lall et al., 2011). Some investigators have also been able to include pollution source (source related $PM_{2.5}$) information through factor analyses in their times series models (Lall et al., 2011).

The data that motivates the proposed research are from environmental health time series studies where $PM_{2.5}$ composition data are collected on each day. The primary exposure of interest is daily source-related fine particulate matter or aerodynamic diameter $\leq$ 2.5 μm ($PM_{2.5}$) and its relation to mortality or hospital admissions.

3.4 MODEL AND NOTATION

3.4.1 Factor Analysis Modeling Framework

We consider a full-likelihood approach that estimates the source contributions from the receptor model and subsequently substitutes the estimates into the health effects model taking the form of a generalized linear regression model. The factor model is as follows:

Receptor Model:

$$\mathbf{X}_t = \mathbf{\Lambda}\mathbf{S}_t + \boldsymbol{\epsilon}_t^X \tag{3.1}$$

$$g(\boldsymbol{\mu}_t) = \alpha_0 + \boldsymbol{\alpha}^T\mathbf{S}_t \tag{3.2}$$

Notationally, $\mathbf{X}_t$ is the vector of $[p \times 1]$ elemental concentrations for a given time t(day). $\mathbf{\Lambda}$ is the $[p \times k]$ matrix of factor loadings, $\mathbf{S}_t$ is a $[k \times 1]$ vector of unobserved source contributions. $\mathbf{Y}_t$ is the health outcome for a given time t. The $\boldsymbol{\alpha}$ represent $[1 \times k]$ effect estimates for the k pollution sources. Each component (elemental concentration) of $\mathbf{X}_t$ is represented by equation (3.1).

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \dots & \dots & \lambda_{1k} \\ \lambda_{21} & \lambda_{22} & \dots & \dots & \lambda_{2k} \\ \lambda_{31} & \lambda_{32} & \dots & \dots & \lambda_{3k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \dots & \dots & \lambda_{pk} \end{bmatrix}_{[p \times k]} ,$$

$$X_{it} = \sum_{j=1}^{k} \lambda_{ij} s_{jt} + \epsilon_{it}^{X} \tag{3.3}$$

For the i^{th} element $(1 \leq i \leq p)$, the j^{th} factor loading $(1 \leq j \leq k)$, for day t $(1 \leq t \leq T)$. The distributional assumptions are as follows:

$\mathbf{S}_t \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$

$\boldsymbol{\epsilon}_t^X \sim N(0, \mathbf{\Psi})$

where,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & 0 & \\ 0 & 0 & \sigma_3^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_k^2 \end{bmatrix}, \boldsymbol{\Psi} = \begin{bmatrix} \psi_1^2 & 0 & 0 & \dots & 0 \\ 0 & \psi_2^2 & 0 & \dots & 0 \\ 0 & 0 & \psi_3^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \psi_p^2 \end{bmatrix}$$

3.4.2 Two-Stage Hierarchical Regression Modeling Framework

The Greenland (1993) method of hierarchical modeling seeks to perform dimension reduction on the effect estimates associated with multiple exposures rather than the exposures variables themselves. Given equation (3.4) from section 3.3, the corresponding health effects estimates ($\boldsymbol{\beta}$) are calculated. Since the exposure variables may be correlated or if there are not enough events to accurately estimate the $\boldsymbol{\beta}$'s, these estimates are often unstable. A hierarchical approach can often remedy some of these issues. The two-stage approach takes the form:

$$g(\boldsymbol{\mu}_t) = \alpha_0 + \boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{X}_t \tag{3.4}$$

$$\beta_i = \boldsymbol{\omega}^{\boldsymbol{T}}\mathbf{Z}_i + \delta_i, i \in 1, \dots, p. \tag{3.5}$$

By substituting equation (3.5) into equation (3.4) we have the following model:

$$g(\boldsymbol{\mu}_t) = \alpha + \mathbf{X}_t\mathbf{Z}\boldsymbol{\omega} + \mathbf{X}_t\boldsymbol{\delta}, \tag{3.6}$$

where the $\boldsymbol{\omega}$ are treated as a vector of fixed coefficients while δ is treated as a vec-

tor of random coefficients with mean 0 and variance τ^2. The i^{th} row of $\mathbf{Z}$ contains second-stage covariates for the i^{th} exposure in β_i. $\boldsymbol{\omega}$ is a vector of second-stage regression coefficients, and the elements of δ_i are independent normal random variables with zero means and variances τ^2. Hence, as described above, we use the second-stage covariates (that is, columns of $\mathbf{Z}$) to model similarities among the β_i in an attempt to improve conventional estimates from equation (3.4).

For cases when we are interested in assessing the health effects of PM sources, the question arises of which second stage covariate matrix ($\mathbf{Z}$) should be chosen. If one were to fit model (3.5) in two distinct stages, then the estimates for the second stage coefficients $\boldsymbol{\omega}$ would take the following form:

$$\widehat{\boldsymbol{\omega}} = (\mathbf{Z^T Z})^{-1}\mathbf{Z^T}\widehat{\boldsymbol{\beta}}.$$

This shows that the second stage effect estimates ($\widehat{\boldsymbol{\omega}}$) are a function of the $\mathbf{Z}$ matrix and the health effects estimates ($\boldsymbol{\beta}$) from the elemental concentrations. This relationship provides a motivation for some connection between the source apportionment and two stage approaches. In the next section, to further motivate our choices for $\mathbf{Z}$, we consider a special case of a source apportionment model with $X_{it} = \sum_{j=1}^{k} \lambda_{ij} s_{jt} + \epsilon_{it}^{X}$, where we allow the second stage covariates $(\mathbf{Z})$ to become different variations of the factor loadings $\boldsymbol{\Lambda}$.

Dimension Reduction

Factor analysis seeks dimension reduction of the receptor model by reducing the dimension of the exposure in the health effects model. By substituting equation (3.3) into equation (3.4) below, the model is reduced from p to k dimensions for $(k < p)$ different source contributions.

$$\begin{aligned}
g(\boldsymbol{\mu}_t) &= \alpha_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_p X_{pt} \\
&= \alpha_0 + \beta_1 \left[\sum_{j=1}^{k} \lambda_{1j} \mathbf{s}_{jt} \right] + \beta_2 \left[\sum_{j=1}^{k} \lambda_{2j} \mathbf{s}_{jt} \right] + \cdots + \beta_p \left[\sum_{j=1}^{k} \lambda_{pjt} \mathbf{s}_{jt} \right] \\
&= \alpha_0 + \beta_1 \left[\lambda_{11} \mathbf{s}_{1t} + \cdots + \lambda_{1k} \mathbf{s}_{kt} \right] + \cdots + \beta_p \left[\lambda_{p1} \mathbf{s}_{1t} + \cdots + \lambda_{pk} \mathbf{s}_{kt} \right] \\
&= \alpha_0 + \left[\beta_1 \lambda_{11} \mathbf{s}_{1t} + \cdots + \beta_1 \lambda_{1k} \mathbf{s}_{kt} \right] + \cdots + \left[\beta_p \lambda_{p1} \mathbf{s}_{1t} + \cdots + \beta_p \lambda_{pk} \mathbf{s}_{kt} \right] \\
&= \alpha_0 + \left[\beta_1 \lambda_{11} \mathbf{s}_{1t} + \cdots + \beta_p \lambda_{p1} \mathbf{s}_{1t} \right] + \cdots + \left[\beta_1 \lambda_{1k} \mathbf{s}_{kt} + \beta_2 \lambda_{2k} \mathbf{s}_{kt} + \cdots + \beta_p \lambda_{pk} \mathbf{s}_{kt} \right]
\end{aligned}$$

Therefore, the equations become:

$$\begin{aligned}
g(\boldsymbol{\mu}_t) &= \alpha_0 + \overbrace{\left[\beta_1 \lambda_{11} + \beta_2 \lambda_{21} + \cdots + \beta_p \lambda_{p1} \right]}^{\alpha_1} \mathbf{s}_{1t} + \overbrace{\left[\beta_1 \lambda_{12} + \beta_2 \lambda_{22} + \cdots + \beta_p \lambda_{p2} \right]}^{\alpha_2} \mathbf{s}_{2t} + \ldots \\
&+ \overbrace{\left[\beta_1 \lambda_{1k} + \beta_2 \lambda_{2k} + \cdots + \beta_p \lambda_{pk} \right]}^{\alpha_k} \mathbf{s}_{kt} \\
&= \alpha_0 + \alpha_1 \mathbf{s}_{1t} + \alpha_2 \mathbf{s}_{2t} + \cdots + \alpha_k \mathbf{s}_{kt}
\end{aligned}$$

$$\hat{\alpha}_j = \sum_{i=1}^{p} \beta_i \lambda_{ij} \tag{3.7}$$

Which means that the health effects represented by the k PM sources ($\hat{\alpha}_j$) are a linear combination of the element-specific coefficients (β_i), weighted by the loadings of the elements for that specific source. Given these considerations, at least thee choices of **Z** may be reasonable in the hierarchical formulation, each of which allow **Z** to represent a variation of the factor loading matrix.

1. The non-overlapping case, where $Z_{ij} = 1$ if the source has the highest loading for the particular element and $Z_{ij} = 0$ for the remaining sources for that element.

2. A moderately overlapping case, where $Z_{ij} = 1$ if the source loading is greater than some constant threshold and $Z_{ij} = 0$ for the remaining sources for that element.

3. A general case, where $Z_{ij} = z_{ij}$ where the z_{ij} could be representation of real factor loadings measured or those given be previous or existing studies. Each source has its own contribution for each element. In our case, we allow $z_{ij} = \lambda_{ij}$, where each λ_{ij} is the factor loading for each element and source combination from concentrator data.

Case 1 (No Overlap):

Assume the following sample **Z**:

$$\mathbf{Z} = \begin{bmatrix} \lambda_{11} & 0 & 0 & 0 \\ \lambda_{21} & 0 & 0 & 0 \\ \lambda_{31} & 0 & 0 & 0 \\ 0 & \lambda_{42} & 0 & 0 \\ 0 & \lambda_{52} & 0 & 0 \\ 0 & \lambda_{62} & 0 & 0 \\ 0 & 0 & \lambda_{73} & \dots \\ \vdots & \vdots & \vdots & \ddots \\ 0 & 0 & \dots & \lambda_{pk} \end{bmatrix}_{[\mathbf{p}\times\mathbf{k}]} , \tag{3.8}$$

where $\mathbf{Z}$ represents the p elemental concentrations for the k_{th} source contribution and $\widehat{\boldsymbol{\beta}}$ is a $p \times 1$ vector of coefficient estimates given from the health effects model in equation (3.4). In this simple form of the factor loadings given by equation (3.8), we choose $\mathbf{Z}$ to contain indicators reflecting which source is most responsible for an element. In essence the $Z_{ij} = \{1, 0\}$ and each elemental concentration is assumed to be given fully by one source contribution. Each λ_{ij} does not have to be exactly like (3.8) the only restriction is that there is one non-zero factor loading per row. Therefore, the normal equations and corresponding health effects estimates are given by the following $k \times 1$ matrix :

$$\widehat{\boldsymbol{\omega}} = (\mathbf{Z^T Z})^{-1}\mathbf{Z^T}\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \frac{\sum_{i=1}^{p} \lambda_{i1}\hat{\beta}_i}{\sum_{i=1}^{p} \lambda_{i1}} \\ \frac{\sum_{i=1}^{p} \lambda_{i2}\hat{\beta}_i}{\sum_{i=1}^{p} \lambda_{i2}} \\ \vdots \\ \frac{\sum_{i=1}^{p} \lambda_{ik}\hat{\beta}_i}{\sum_{i=1}^{p} \lambda_{ik}} \end{bmatrix}_{[\mathbf{k}\times\mathbf{1}]} . \tag{3.9}$$

In this simple case $\widehat{\omega}_j = \frac{\sum_{i=1}^{p} \lambda_{ij}\hat{\beta}_i}{\sum_{i=1}^{p} \lambda_{ij}}$ whereas $\widehat{\alpha}_j = \sum_{i=1}^{p} \lambda_{ij}\hat{\beta}_i$. In this case there are overlapping terms and $\widehat{\omega}_j$ proportional to $\widehat{\alpha}_j$. An example of the simple case of $\mathbf{Z}$ could be the following where $p = 13$, $k = 4$:

$$\mathbf{Z}_{non-overlapexample} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}_{[p \times k]}$$

Given this particular scenario and using the general equations for $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\omega}}$ the health effects estimates can be calculated as seen in Table 3.1.

Estimates	$\widehat{\boldsymbol{\alpha}}$	$\widehat{\boldsymbol{\omega}}$
1	$\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$	$\frac{\hat{\beta}_1+\hat{\beta}_2+\hat{\beta}_3}{3}$
2	$\hat{\beta}_4 + \hat{\beta}_5 + \hat{\beta}_6$	$\frac{\hat{\beta}_4+\hat{\beta}_5+\hat{\beta}_6}{3}$
3	$\hat{\beta}_7 + \hat{\beta}_8 + \hat{\beta}_9$	$\frac{\hat{\beta}_7+\hat{\beta}_8+\hat{\beta}_9}{3}$
4	$\hat{\beta}_{10} + \hat{\beta}_{11} + \hat{\beta}_{12} + \hat{\beta}_{13}$	$\frac{\hat{\beta}_{10}+\hat{\beta}_{11}+\hat{\beta}_{12}+\hat{\beta}_{13}}{4}$

Table 3.1: Health effects estimates for $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\omega}}$

Case 2 (Moderately Overlapping Case):

In general, an element is not generated from a single source, which is why the source profiles of a typical $(\mathbf{\Lambda})$ are not non-zero for only one entry per row as in case 1. In these settings where a single element is spread across many sources, we can extend the choice of $\mathbf{Z}$ from case 1 to include such settings. We propose 2 new variations: In the first variation we use indicators wherever the source profiles are greater than some threshold C and will be called the "moderately overlapping case". In this case, an element's loadings can be distributed across different sources but indicator values (0,1) are still used. Assume $\mathbf{\Lambda}$ has the following form:

$$\mathbf{\Lambda_1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.88 & 0.00 & 0.01 & 0.00 \\ 0.83 & 0.08 & 0.34 & 0.09 \\ 0.91 & 0.02 & 0.31 & 0.17 \\ 0.00 & 0.95 & 0.00 & 0.01 \\ 0.02 & 0.65 & 0.05 & 0.26 \\ 0.16 & 0.04 & 1.02 & 0.03 \\ 0.18 & 0.58 & 0.26 & 0.43 \\ 0.17 & 0.41 & 0.44 & 0.65 \\ 0.13 & 0.27 & 0.51 & 0.81 \end{bmatrix}_{[p\times k]} \tag{3.10}$$

Given these values of $\mathbf{\Lambda_1}$, an appropriate "moderately overlapping" $\mathbf{Z}$ matrix can be constructed. By allowing each value of $\lambda_{ij} \geq .30$ to be valued at 1 and $\lambda_{ij} < .30$ to be valued

at 0 we are left with the following $\mathbf{Z}$ matrix.

$$\mathbf{Z}_{mod.overlap} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}_{[p \times k]}$$

Case 3 (General Overlapping Case):

In the third and last variation of the second stage covariates we choose $\mathbf{Z}$ to be exactly equal to the true $\mathbf{\Lambda_1}$ given in equation (3.10). In this case, each of the elemental concentrations can be attributed to more than one source and the values represent real-valued factor loadings that are not restricted to indicators. With this assumption, the normal equations for both case 2 and case 3 can be generalized to:

$$\widehat{\boldsymbol{\omega}} = (\mathbf{Z^T Z})^{-1}\mathbf{Z^T}\widehat{\boldsymbol{\beta}} = \begin{bmatrix} c_{11}\alpha_1 + c_{12}\alpha_2 + c_{13}\alpha_3 + c_{14}\alpha_4 + \cdots + c_{1k}\alpha_k \\ c_{21}\alpha_1 + c_{22}\alpha_2 + c_{23}\alpha_3 + c_{24}\alpha_4 + \cdots + c_{2k}\alpha_k \\ \vdots \\ c_{k1}\alpha_1 + c_{k2}\alpha_2 + c_{k3}\alpha_3 + c_{k4}\alpha_4 + \cdots + c_{kk}\alpha_k \end{bmatrix}_{[\mathbf{k}\times\mathbf{1}]}$$

Therefore, in the overlapping case $\widehat{\omega}_j = \sum_{i=1}^{k} c_{ji}(\lambda)\widehat{\boldsymbol{\alpha}}(\lambda)$ whereas $\widehat{\alpha}_j = \sum_{i=1}^{p} \beta_i \lambda_{ij}$. The two estimates differ by a factor of c_{ji} so there are some overlapping terms in the estimates. The c_{ji} represent the components of the $(\mathbf{Z}^T\mathbf{Z})^{-1}$ matrix and are functions of the λ_{ij}. Again the α_i are linear combinations of the health effects estimates and the specific factor loadings given by λ_{ij} .

3.5 SIMULATION STDY

We conducted a simulation study to examine the statistical properties of the health effects estimates obtained from the two-stage and factor analysis approaches. The objective of this simulation is to compare the two methods and determine the settings where each scheme is most powerful in detecting differences and maintaining low type 1 errors. The standard errors from the two models are also reported so that comparisons in efficiency can be made.

3.5.1 Simulating Source Data

We assume $k = 4$ source contributions. Although the source contributions are not directly observed or measured in the studies motivating this research, we simulate them in this study so that they are effectively known. First we simulate the source contributions $\mathbf{S}_t$

according to the following normal distribution:

$$\mathbf{S}_t \sim MVN_4(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu} = 0$. Given each set of source contributions we are then able to simulate the corresponding elemental concentrations from

$$\mathbf{X}_t|\mathbf{S}_t \sim MVN_{13}(\boldsymbol{\Lambda}\mathbf{S}_t, \boldsymbol{\Psi})$$

$$diag(\boldsymbol{\Sigma}) = \begin{bmatrix} RoadDust \\ PowerPlant \\ OilComb \\ Vehicles \end{bmatrix} = \begin{bmatrix} 2.36 \\ 1.60 \\ 1.49 \\ 1.62 \end{bmatrix}$$

,

$$diag(\boldsymbol{\Psi}) = \begin{bmatrix} \psi_{Si} \\ \psi_{S} \\ \psi_{Ni} \\ \psi_{OC} \\ \psi_{Al} \\ \psi_{Ti} \\ \psi_{Ca} \\ \psi_{SULF} \\ \psi_{Se} \\ \psi_{V} \\ \psi_{Br} \\ \psi_{BC} \\ \psi_{EC} \end{bmatrix} = \{\boldsymbol{\Psi}_{\mathbf{given}} = \begin{bmatrix} 0.08 \\ 0.05 \\ 0.22 \\ 0.45 \\ 0.05 \\ 0.28 \\ 0.35 \\ 0.05 \\ 0.31 \\ 0.05 \\ 0.31 \\ 0.11 \\ 0.10 \end{bmatrix}, \boldsymbol{\Psi}_{\mathbf{13}} = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.3 \\ 0.3 \\ 0.1 \\ 0.3 \\ 0.1 \\ 0.3 \\ 0.3 \\ 0.3 \end{bmatrix}, \boldsymbol{\Psi}_{\mathbf{31}} = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \\ 0.1 \\ 0.1 \\ 0.3 \\ 0.1 \\ 0.3 \\ 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}\}$$

Both $\mathbf{\Psi}$ and $\mathbf{\Sigma}$ represent realistic variance structures for the source contributions and the exposure data respectively given by settings from concentrator data (Nikolov et al., 2000). We assume the factor loadings of $\mathbf{\Lambda}$ follow the given form. Because the source profiles are unknown and the source contributions are unobserved the model may not be indentifiable and will not have a unique solution. The model can be made indentifiable by constraining some of the factor loadings in $\mathbf{\Lambda_1}$. The elements silicon (Si), sulfur (S), Nickel (Ni), and organic carbon (OC) were chosen to be constrained because they were most unilaterally emitted by single sources.

$$\mathbf{\Lambda_1} = \left[\begin{array}{c|cccc} & S_1 & S_2 & S_3 & S_4 \\ \hline Si & 1 & 0 & 0 & 0 \\ S & 0 & 1 & 0 & 0 \\ Ni & 0 & 0 & 1 & 0 \\ OC & 0 & 0 & 0 & 1 \\ Al & 0.88 & 0.00 & 0.01 & 0.00 \\ Ti & 0.83 & 0.08 & 0.34 & 0.09 \\ Ca & 0.91 & 0.02 & 0.31 & 0.17 \\ Sulf & 0.00 & 0.95 & 0.00 & 0.01 \\ Se & 0.02 & 0.65 & 0.05 & 0.26 \\ V & 0.16 & 0.04 & 1.02 & 0.03 \\ Br & 0.18 & 0.58 & 0.26 & 0.43 \\ BC & 0.17 & 0.41 & 0.44 & 0.65 \\ EC & 0.13 & 0.27 & 0.51 & 0.81 \end{array}\right]_{[p\times k]}$$

3.5.2 Simulating Health Outcome Data

The health outcome was simulated using a poisson distribution, since data represented counts. The outcomes of interest were mortality and number of hospital admissions. Further, we generated health outcome assuming a health effect from a single source

$$\mathbf{Y}_t \sim Pois(\boldsymbol{\mu}),$$

where k is the source contribution at time t. Therefore,

$$\begin{aligned}
\mathbf{Y}_t^{(1)}|\{\alpha_1 = \alpha_1, \alpha_2 = \alpha_3 = \alpha_4 = 0\} &\sim Pois(\boldsymbol{\mu} = exp(\alpha_0 + \alpha_1 S_1),\\
\mathbf{Y}_t^{(2)}|\{\alpha_2 = \alpha_2, \alpha_1 = \alpha_3 = \alpha_4 = 0\} &\sim Pois(\boldsymbol{\mu} = exp(\alpha_0 + \alpha_2 S_2),\\
\mathbf{Y}_t^{(3)}|\{\alpha_3 = \alpha_3, \alpha_1 = \alpha_2 = \alpha_4 = 0\} &\sim Pois(\boldsymbol{\mu} = exp(\alpha_0 + \alpha_3 S_3),\\
\mathbf{Y}_t^{(4)}|\{\alpha_4 = \alpha_4, \alpha_1 = \alpha_2 = \alpha_3 = 0\} &\sim Pois(\boldsymbol{\mu} = exp(\alpha_0 + \alpha_4 S_4).
\end{aligned}$$

Where $\mathbf{Y}_t^{(1)}|\alpha_1$ represents the outcome associated with only the first source. $\mathbf{Y}_t^{(2)}|\alpha_2$, $\mathbf{Y}_t^{(3)}|\alpha_3$, and $\mathbf{Y}_t^{(4)}|\alpha_4$ are interpreted in the corresponding way. In the simulation we generate the health effect on each of the four sources individually. Therefore we have four sets of health outcomes $\mathbf{Y}_t^{(1)}$, $\mathbf{Y}_t^{(2)}$, $\mathbf{Y}_t^{(3)}$, $\mathbf{Y}_t^{(4)}$, where each health effect corresponds to a different pollution source. The initial health outcome parameters were given by the following values:$\alpha_0 = 3.00$,

$$\boldsymbol{\alpha}_1 = \{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$$

$$\boldsymbol{\alpha}_2 = \{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$$

$$\boldsymbol{\alpha}_3 = \{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$$

$$\boldsymbol{\alpha}_4 = \{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$$

Each model was run for $T = 3000$ days, $N = 13$ elemental concentrations, and $k = 4$ pollution sources. Each simulation was run for 1000 iterations.

3.5.3 Approaches

1. ***Known source contribution model***: We previously noted that the sources were simulated so we estimate the health effects based on the known source contributions using the poisson model.

$$\begin{aligned} \mathbf{X}_t &= \boldsymbol{\Lambda}\mathbf{S}_t + \boldsymbol{\epsilon}_t^X \\ log(\boldsymbol{\mu}_t) &= \alpha_0 + \boldsymbol{\alpha}^T\mathbf{S}_t \end{aligned}$$

2. ***Factor Analyses***: We conducted a confirmatory factor analyses (CFA) using the SEM package in R in order to get estimates for the corresponding factor scores (estimated source contributions). The aforementioned initial values and the constrained factor loadings were used. Next, we performed an exploratory factor analysis (EFA) using the "factanal" package in R and a

principle components analysis (PCA) using the "princomp" package in R. Both the EFA and the PCA models do not assume the factor loading structure nor do they assume a distribution for the source contributions. Once the estimates for the respective source contributions were determined, a poisson model was subsequently fit.

$$\begin{aligned} \mathbf{X}_t &= \mathbf{\Lambda}\mathbf{S}_t + \boldsymbol{\epsilon}_t^X \\ log(\boldsymbol{\mu}_t) &= \alpha_0 + \boldsymbol{\alpha}^T \hat{\mathbf{S}}_t \end{aligned}$$

3. ***Two-Stage Approach***: We begin with the health effects model. In the second stage, we choose covariates $\mathbf{Z}$ to model similarities among the β_i in an attempt to improve conventional estimates from the health effects model.

$$\begin{aligned} log(\boldsymbol{\mu}_t) &= \alpha_0 + \boldsymbol{\beta}^T \mathbf{X}_t \\ \beta_{i,[13\times 1]} &= \mathbf{Z}_{[13\times 4]}\boldsymbol{\omega}_{[4\times 1]} + \boldsymbol{\delta}_{i,[13\times 1]}, i \in 1, \ldots, p \end{aligned}$$

therefore,

$$log(\boldsymbol{\mu}_t) = \alpha_0 + \mathbf{X}_t\mathbf{Z}\boldsymbol{\omega} + \mathbf{X_t}\boldsymbol{\delta}$$

3.5.4 Choice of Second Stage Covariates

We allowed for three cases in the choice for $\mathbf{Z}$.

- An non overlapping case we assume a simple form for the factor loadings

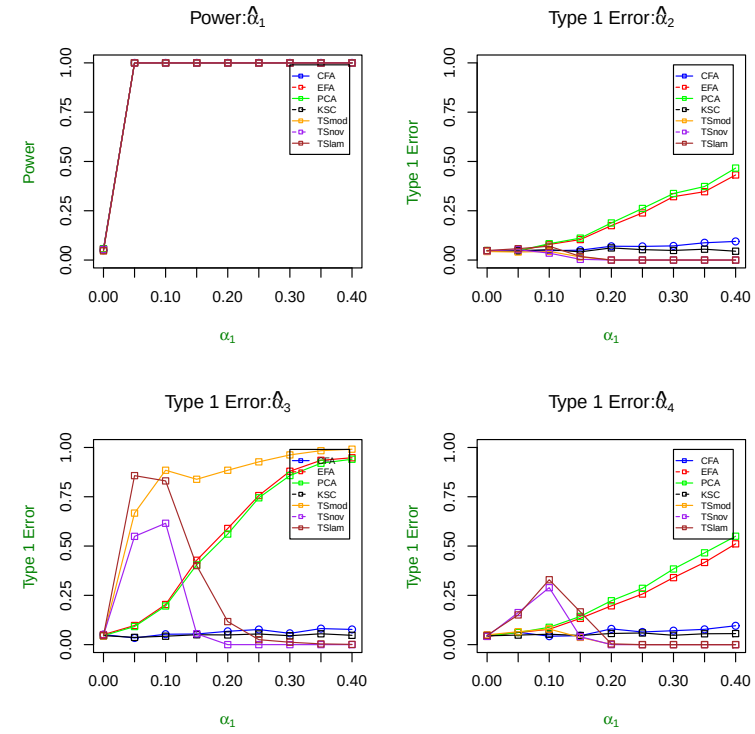
of $\mathbf{Z}$. Each elemental concentration is assumed to be attributed fully to one source contribution. This is done by taking the largest source contribution value for a given elemental concentration.

- A moderately overlapping case where we allowed each value of $\mathbf{\Lambda} : \lambda_{\mathbf{ij}} \geq 0.3$ to be valued at 1 and $\mathbf{\Lambda} : \lambda_{\mathbf{ij}} < 0.3$ to be valued at 0. For some $C > 0$.

- In the overlapping case we assume each of the elemental concentrations can be attributed to more than one source. In this case we allow $\mathbf{Z}$ to be our initial case $\mathbf{\Lambda}$ $(\mathbf{Z} = \mathbf{\Lambda})$.

The three $\mathbf{Z}$ matrices follow:

$$\mathbf{Z}_{no-overlap} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \mathbf{Z}_{mod.overlap} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \mathbf{Z}_{generalcase} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0.88 & 0.00 & 0.01 & 0.00 \\ 0.83 & 0.08 & 0.34 & 0.09 \\ 0.91 & 0.02 & 0.31 & 0.17 \\ 0.00 & 0.95 & 0.00 & 0.01 \\ 0.02 & 0.65 & 0.05 & 0.26 \\ 0.16 & 0.04 & 1.02 & 0.03 \\ 0.18 & 0.58 & 0.26 & 0.43 \\ 0.17 & 0.41 & 0.44 & 0.65 \\ 0.13 & 0.27 & 0.51 & 0.81 \end{bmatrix}$$



..

Figure 3.1: Overlaid Power and Type 1 Error Curves for $Y_t^{(1)}|\mathbf{S_t}$ and $\mathbf{\Psi_{given}}$: A(top left): The power for $\hat{\alpha}_1$ of each health effects model at the given initial value of α_1. B(top right): The type 1 error for $\hat{\alpha}_2$ of each health effects model at a given initial value of α_1. C(bottom left): The type 1 error for $\hat{\alpha}_3$ of each health effects model at a given initial value of α_1. D(bottom right): The type 1 error for $\hat{\alpha}_4$ of each health effects model at a given initial value of α_1.

3.5.5 Simulation Results

Figures 3.1 and 3.2 display power and type 1 error curves for the health effects estimates obtained from five different modeling schemes. The health effects estimates given by the known source contribution model (KSC) are considered the "gold standard". The health effects estimates from the estimated source contributions was determined using three different computing packages. The "CFA" for the confirmatory factor ananlysis using structural equation modeling package in R, "EFA" for the exploratory factor analysis (factanal package in R), and "PCA" for the principle components analysis (princomp package in R). There were three ver-

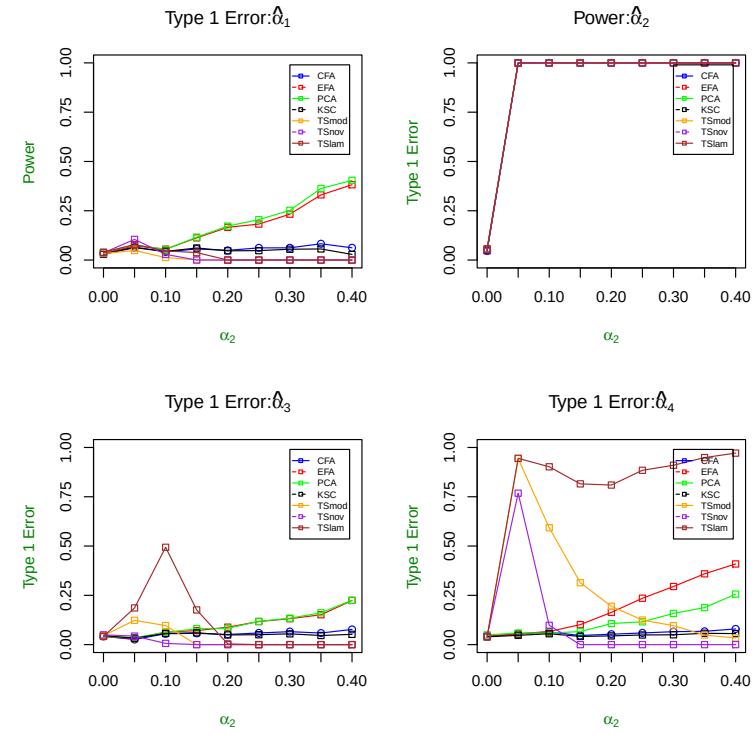


Figure 3.2: Overlaid Power and Type 1 Error Curves for $Y_t^{(2)}|\mathbf{S_t}$ and $\mathbf{\Psi_{given}}$: A(top left): The type 1 error for $\hat{\alpha}_1$ of each health effects model at the given initial value of α_2. B(top right): The power for $\hat{\alpha}_2$ of each health effects model at a given initial value of α_2. C(bottom left): The type 1 error for $\hat{\alpha}_3$ of each health effects model at a given initial value of α_2. D(bottom right): The type 1 error for $\hat{\alpha}_4$ of each health effects model at a given initial value of α_2.

sions of the two-stage regression approach taken; 1) The second-stage covariates did not overlap ("TSnov") and 2) where the second stage covariates do overlap ("TSmod"), and lastly the where the second stage covariates equal $\mathbf{\Lambda}$ ("TSlam"). Each two-stage regression model of was run in glmmPQL package in R.

Each figure consists of four graphs that represent calculations for the health effects estimates $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$, and $\hat{\alpha}_4$ at different initial values given along the x-axis. Each model was simulated so that one specific source would be responsible for the entire health effect. For example, Figure 3.1 represents $Y_t^{(1)}|S_t$ which denotes that the health effect was simulated to be associated with source 1. Simultaneously, sources

2 through 4 were assumed to have no association with the outcome. Therefore, Figure 3.1A (top left) shows six power curves at varying initial values that represent each of the afore mentioned methods. The expectation is that the power curves to detect differences in position A would be quite high because that is where the true association lies. Alternatively, positions 3.1B (top right), 3.1C (bottom left), and 3.1D (bottom right) show seven type 1 error curves each at the varying initial values of α_1. The expectation is that the type 1 errors in positions 3.1B through 3.1D would be quite low (Pr{detect a difference | there is no difference} = 0.05) because these sources were assumed to have no effect.

Figure 3.1A shows the power curves for $\hat{\alpha}_1$ (outcome $Y_t^{(1)}$ given the health effect is only associated with source 1. The 4 source apportionment power curves (CFA, EFA, PCA, and KSC) begin at approximately 0.05 when $\alpha_1 = 0$ but increase to nearly 100% for increasing initial values of the health effect estimate (α_1). This means that the source apportionment approach is able to estimate the appropriate effect with high power. Similarly, the 3 two-stage regression power curves (TSnov, TSov, and TSlam) follow that same pattern and increase in power for larger values of α_1. Figure 3.1B shows type 1 error curves for $\hat{\alpha}_2$ ($Y_t^{(1)}$ given that the source 2 reflects a zero contribution to the health effect). The CFA, EFA, and PCA approaches show increasing $\hat{\alpha}_2$ type 1 errors for increasing initial values of α_1. As the initial health effect estimate (α_1) increases the proportion of false positives for $\hat{\alpha}_2$ rises from approximately 0.05 to 0.30. Alternatively, the type 1 errors for $\hat{\alpha}_2$ from approaches TSnov, TSmod, TSlam, and KSC straddle the 0.05 line for all values of α_1. Figure 3.1C shows type 1 error curves for $\hat{\alpha}_3$ which reflects the outcome ($Y_t^{(1)}$) given that source 3 reflects a zero contribution to the health effect. For the health effect $\hat{\alpha}_3$, the CFA, EFA, and PCA approaches again show increasing type 1 errors with increasing initial values of α_1. As the health effect estimate α_1 increases the

proportion of false positives for $\hat{\alpha}_3$ rises from approximately 0.05 to 0.90. TSmod maintains a high type 1 error rate throughout the varying initial values of α_1. On the other hand, the TSnov and TSlam models have high type 1 error rates in the early values of α_1 but returns to below 0.05 when $\alpha_1 = 0.15$. The KSC model type 1 error curve straddles the 0.05 line as expected because the true simulated sources were used. Figure 3.1D shows type 1 error curves for $\hat{\alpha}_4$ where the outcome ($Y_t^{(1)}$) is given that the source 4 reflects a zero contribution to the health effect. The type 1 error for $\hat{\alpha}_4$ from the CFA, EFA, and PCA approaches show increasing type 1 errors with increasing initial values of α_1. As the initial health effect estimate (α_1) increases the proportion of false positives from $\hat{\alpha}_4$ rise from approximately 0.05 to 0.55. The type 1 errors for $\hat{\alpha}_4$ in the TSnov, TSmod, and TSlam approaches experience slight increases int type 1 errors for initial values before $\alpha_1 = 0.20$. For initial values after $\alpha_1 = 0.20$, the type 1 error rates for $\hat{\alpha}_4$ in models TSnov, TSmod, and TSlam decrease to below 0.05.

Overall, the simulation suggests different patterns for the factor analyses and the two stage approaches. For the factor analyses, the sources with the clear simulated relationship are able to produce effect estimates that are highly powered. For the remaining sources with no association to the health outcome, there are inflated type 1 errors. The source apportionment methods in particular show increasing type 1 errors for increasing initial health effect estimate values, for all settings irrespective of source variance structures. This would indicate that when there is a large signal (health effect estimate) from one source, there is an accompanying spill-over effect into the other sources. Both the EFA and PCA models have the highest type 1 error levels while the CFA has just moderate increases.

On the other hand the two stage approaches showed different patterns. For both

the non-overlapping, moderately overlapping, and the general two stage cases, the power to detect differences is very high for the source with the clear simulated effect. Alternatively, the sources that have no association to the health outcome experience high type 1 error spikes which is in contrast with the factor analysis models. For smaller health effect estimate values between $\alpha = 0.0 - 0.20$, there are large increases in the type 1 error rates. This would indicate that the two-stage approach has difficulties estimating health effects that are very small but once the health effects become larger in magnitude the type 1 error rates come down to normal levels. Another possible predictor of the spikes in type 1 error rates for the two-stage approaches are the distribution of single elements across the many sources. There is a tendency for the spikes to coincide with health effects where the elements are evenly spread across different sources.

3.5.6 Simulation Implications

The confirmatory factor analysis procedures (CFA) are able to estimate the source contributions and health effects estimates with relatively low bias as can be seen in Table 3.2. Table 3.2 shows the health effects estimates from a confirmatory factor analysis along with the model standard errors and the simulation standard deviations. There are 4 sets of initial values represented in Table 3.2: 1) top left: $\{\mathbf{Y}_t^{(1)}|\alpha_1 = 0.05, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = 0\}$which means the health effect associated with source 1 is 0.05 while the effect of the other sources is 0, 2) bottom left: $\{\mathbf{Y}_t^{(2)}|\alpha = 0, \alpha_2 = 0.05, \alpha_3 = 0, \alpha_4 = 0\}$which means the health effect associated with source 2 is 0.05 while the effect of the other sources is 0, 3) top right: $\{\mathbf{Y}_t^{(1)}|\alpha_1 = 0.30, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = 0\}$, and 4) $\{\mathbf{Y}_t^{(2)}|\alpha_1 = 0, \alpha_2 = 0.30, \alpha_3 = 0, \alpha_4 = 0\}$. Since the health effects estimates are unbiased, then the variability of the es-

timates maybe leading to the higher false positive rates due to the bias-variance trade off. Similar evidence of the discrepancy between the model standard errors and the simulation standard deviations can be seen in Tables 3.2 - 3.4. Each of the standard deviation estimates are larger than the standard error estimates for EFA and PCA models while this pattern is present but attenuated in the CFA model. For example in Table 3.2A, the CFA health effect estimate $\hat{\alpha_1}$ is 0.05001 which is quite close to the simulated value of 0.05, the standard error is 0.0027 and the standard deviation is 0.0026. On the other hand, in Tables 3.3 and 3.4 show that the standard deviations are in most cases twice as large as the standard errors for large values of α. For example, in Table 3.3C it is noted that the SE for $\hat{\alpha_1}$ is 0.0039 while the SD is 0.0070 for initial values of 0.30. In order to correct the discrepancy in errors, a bootstrap model can be conducted which will give more accurate estimates of the precision/variability. Subsequently, the number of false positives will be reduced to below 0.05 and the power estimates will be regulated/attenuated.

$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.05, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = 0$

Coef.	CFA Est.	Std Err.	Std. Dev.
$\hat{\alpha_1}$	0.0500	0.0027	0.0026
$\hat{\alpha_2}$	-0.0002	0.0033	0.0033
$\hat{\alpha^3}$	0.0001	0.0034	0.0034
$\hat{\alpha^4}$	0.0000	0.0033	0.0032

$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.30, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = 0$

Coef.	CFA Est.	Std Err.	Std. Dev.
$\hat{\alpha_1}$	0.2999	0.0025	0.0027
$\hat{\alpha_2}$	0.0001	0.0031	0.0032
$\hat{\alpha^3}$	-0.0004	0.0032	0.0035
$\hat{\alpha^4}$	0.0002	0.0031	0.0033

$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.0, \alpha_2 = 0.05, \alpha_3 = 0, \alpha_4 = 0$

Coef.	CFA Est.	Std Err.	Std. Dev.
$\hat{\alpha_1}$	0.0000	0.0027	0.0028
$\hat{\alpha_2}$	0.0500	0.0033	0.0033
$\hat{\alpha^3}$	0.0002	0.0034	0.0034
$\hat{\alpha^4}$	0.0000	0.0033	0.0034

$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.0, \alpha_2 = 0.30, \alpha_3 = 0, \alpha_4 = 0$

Coef.	CFA Est.	Std Err.	Std. Dev.
$\hat{\alpha_1}$	0.0000	0.0026	0.0026
$\hat{\alpha_2}$	0.3001	0.0031	0.0033
$\hat{\alpha^3}$	-0.0003	0.0033	0.0033
$\hat{\alpha^4}$	-0.0001	0.0032	0.0034

Table 3.2: A(top left), B(bottom left), C(top right), D(bottom right): This table represents the parameter estimates and errors for the confirmatory factor analyses (CFA) models conducted on simulated data. Generalized linear models were conducted for 1000 iterations. The subsequent estimates were aggregated into medians. Each model included no covariates.

Alternatively, the two-stage hierarchical regression approach is unable to estimate the health effects estimates without bias. The actual health effects estimates are cannot be compared to the source apportionment models because they represent

different values. Each two-stage model represents the relationship between first and second stage covariates and the health outcome. In addition, the estimates are poorly estimated due to the correlations between sources. Such correlations occur when a single element is distributed among many sources. Table 3.5 shows the health effects estimates and errors from the two-stage model. For this approach, there are discrepancies between the model standard errors and the simulation standard deviations when there are spikes in the type 1 error curves. and the standard errors are very similar and in some cases smaller than those of the standard deviations. Therefore, this does not seem to represent a variance issue and the bootstrap will be ineffective.

$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.05, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = 0$

Coef.	EFA Std Err.	EFA Std Dev.
$\hat{\alpha_1}$	0.0041	0.0044
$\hat{\alpha_2}$	0.0041	0.0041
$\hat{\alpha^3}$	0.0042	0.0043
$\hat{\alpha^4}$	0.0041	0.0041

$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.30, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = 0$

Coef.	EFA Std Err	EFA Std Dev.
$\hat{\alpha_1}$	0.0039	0.0070
$\hat{\alpha_2}$	0.0039	0.0056
$\hat{\alpha^3}$	0.0040	0.0056
$\hat{\alpha^4}$	0.0039	0.0052

$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.0, \alpha_2 = 0.05, \alpha_3 = 0, \alpha_4 = 0$

Coef.	EFA Std Err.	EFA Std Dev.
$\hat{\alpha_1}$	0.0041	0.0041
$\hat{\alpha_2}$	0.0041	0.0041
$\hat{\alpha^3}$	0.0042	0.0043
$\hat{\alpha^4}$	0.0041	0.0039

$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.0, \alpha_2 = 0.30, \alpha_3 = 0, \alpha_4 = 0$

Coef.	EFA Std Err.	EFA Std Dev.
$\hat{\alpha_1}$	0.0040	0.0056
$\hat{\alpha_2}$	0.0040	0.0063
$\hat{\alpha^3}$	0.0040	0.0052
$\hat{\alpha^4}$	0.0040	0.0052

Table 3.3: A(top left), B(bottom left), C(top right), D(bottom right): This table represents the parameter estimates and errors for the confirmatory factor analyses (EFA) models conducted on simulated data. Generalized linear models were conducted for 1000 iterations. The subsequent estimates were aggregated into medians, and 95% CI's. Each model included no covariates.

Bootstrap Option/Solution

The bootstrap procedure for the CFA was chosen so that the precision of the health effects estimate (regression coefficient) could be evaluated. For each data set, b = 100 bootstrap samples $1, \ldots, 100$ was retrieved. T = 3000 days, N = 13 elemental concentrations, K = 4 sources, and r = 100 iterations. First, the CFA model was fit using the observed/simulated data and the following estimates were reported:

$\mathbf{Y}_t \vert \mathbf{S}_t, \alpha_1 = 0.05, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = 0$		
Coef.	**PCA Std Err.**	**PCA Std Dev.**
$\hat{\alpha_1}$	0.0041	0.0043
$\hat{\alpha_2}$	0.0041	0.0040
$\hat{\alpha^3}$	0.0041	0.0042
$\hat{\alpha^4}$	0.0041	0.0041

$\mathbf{Y}_t \vert \mathbf{S}_t, \alpha_1 = 0.30, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = 0$		
Coef.	**PCA Std Err**	**PCA Std Dev.**
$\hat{\alpha_1}$	0.0039	0.0071
$\hat{\alpha_2}$	0.0039	0.0056
$\hat{\alpha^3}$	0.0039	0.0054
$\hat{\alpha^4}$	0.0039	0.0056

$\mathbf{Y}_t \vert \mathbf{S}_t, \alpha_1 = 0.0, \alpha_2 = 0.05, \alpha_3 = 0, \alpha_4 = 0$		
Coef.	**PCA Std Err.**	**PCA Std Dev.**
$\hat{\alpha_1}$	0.0041	0.0041
$\hat{\alpha_2}$	0.0041	0.0041
$\hat{\alpha^3}$	0.0041	0.0042
$\hat{\alpha^4}$	0.0041	0.003

$\mathbf{Y}_t \vert \mathbf{S}_t, \alpha_1 = 0.0, \alpha_2 = 0.30, \alpha_3 = 0, \alpha_4 = 0$		
Coef.	**PCA Std Err.**	**PCA Std Dev.**
$\hat{\alpha_1}$	0.0039	0.0057
$\hat{\alpha_2}$	0.0039	0.0063
$\hat{\alpha^3}$	0.0039	0.0050
$\hat{\alpha^4}$	0.0039	0.0053

Table 3.4: A(top left), B(bottom left), C(top right), D(bottom right):This table represents the parameter estimates and errors for the confirmatory factor analyses (PCA) models conducted on simulated data. Generalized linear models were conducted for 1000 iterations. The subsequent estimates were aggregated into medians, and 95% CI's. Each model included no covariates.

$\mathbf{Y}_t \vert \mathbf{S}_t, \alpha_1 = 0.05, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = 0$			
Coef.	**TSnov Est.**	**Std Err.**	**Std Dev.**
$\hat{\alpha_1}$	0.0137	0.0007	0.0007
$\hat{\alpha_2}$	-0.0001	0.0011	0.0011
$\hat{\alpha^3}$	-0.0036	0.0017	0.0017
$\hat{\alpha^4}$	-0.0012	0.0014	0.0014

$\mathbf{Y}_t \vert \mathbf{S}_t, \alpha_1 = 0.30, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = 0$			
Coef.	**TSnov Est.**	**Std Err.**	**Std Dev.**
$\hat{\alpha_1}$	0.0807	0.0163	0.0009
$\hat{\alpha_2}$	-0.0001	0.0163	0.0016
$\hat{\alpha^3}$	-0.0078	0.0232	0.0027
$\hat{\alpha^4}$	-0.0025	0.0189	0.0021

$\mathbf{Y}_t \vert \mathbf{S}_t, \alpha_1 = 0.0, \alpha_2 = 0.05, \alpha_3 = 0, \alpha_4 = 0$			
Coef.	**TSnov Est.**	**Std Err.**	**Std Dev.**
$\hat{\alpha_1}$	-0.0005	0.0007	0.0007
$\hat{\alpha_2}$	0.0156	0.0011	0.0012
$\hat{\alpha^3}$	-0.0003	0.0018	0.0017
$\hat{\alpha^4}$	-0.0038	0.0014	0.0016

$\mathbf{Y}_t \vert \mathbf{S}_t, \alpha_1 = 0.0, \alpha_2 = 0.30, \alpha_3 = 0, \alpha_4 = 0$			
Coef.	**TSnov Est.**	**Std Err.**	**Std Dev.**
$\hat{\alpha_1}$	-0.0007	0.0188	0.0010
$\hat{\alpha_2}$	0.0780	0.0189	0.0017
$\hat{\alpha^3}$	-0.0015	0.0267	0.0026
$\hat{\alpha^4}$	-0.0026	0.0218	0.0022

Table 3.5: A(top left), B(bottom left), C(top right), D(bottom right):This table represents the parameter estimates and errors for the two-stage hierarchical regression models conducted on simulated data. Generalized linear models were conducted for 1000 iterations. The subsequent estimates were aggregated into medians. Each model included no covariates.

$\hat{\mathbf{\Lambda}}$, $\hat{\mathbf{\Psi}}$,$\hat{\mathbf{\Sigma}}$, and $\hat{\beta}$. Second, using these initial values, bootstrapped source data was generated from $\mathbf{S^{(b)}} \sim \mathbf{N}(\mathbf{0}, \hat{\mathbf{\Sigma}})$. Third, bootstrapped exposure (elemental concentrations) data were generated from $\mathbf{X_t^{(b)}} \sim \mathbf{N}(\hat{\mathbf{\Lambda}}\mathbf{S^{(b)}}, \hat{\mathbf{\Sigma}})$. Lastly, health outcome data were generated from $\mathbf{Y^{(b)}} \sim \mathbf{Pois}(\mu^{(\mathbf{b})})$, where $\mu^{(\mathbf{b})} = \mathbf{exp}(\hat{\beta}\mathbf{S^{(b)}})$. As a result of this simulation structure poisson linear models could be fit using $\mathbf{Y^{(b)}}$ and $\mathbf{X^{(b)}}$ as real data. From each of the 100 bootstrap samples the $\hat{\alpha^{(b)}}$ regression coefficients for each source were stored and the 2.5% and 97.5% bootstrap confidence intervals were calculated. The simulation was run for r =100 iterations which yielded 100

confidence intervals for each source. Subsequently, the power and type 1 errors were calculated.

The bootstrap procedure for the two-stage model begins with b = 100 bootstrap samples $1, \ldots, 100$ for each data set. T = 3000 days, N = 13 elemental concentrations, K = 4 sources, and r = 200 iterations. The two-stage model was fit using observed/simulated data and $\hat{\omega}$ for each source, $\hat{\sigma_{\delta}^2}$, and $\hat{\delta}$. Given the original simulated exposure data (elemental concentrations) $\mathbf{X}_t$ and the second stage covariate matrix $\mathbf{Z}$ the bootstrap mean outcome can be estimated. The mean is given by $\mu^{(b)} = exp(\hat{\mu}_0 + \mathbf{X}_t\mathbf{Z}\hat{\omega} + \mathbf{X}_t\hat{\delta}\}$. Lastly, the health outcome data is generated from $\mathbf{Y}^{(b)} \sim Poisson(\mu^{(b)})$. The two stage model is refit and the from each of the 100 bootstrap samples and the $\hat{\omega^{(b)}})$ regression coefficients for each source were stored. The 2.5% and 97.5% bootstrap confidence intervals were calculated. The simulation was run for r =200 iterations which yielded 200 confidence intervals for each source. Subsequently, the power and type 1 errors were calculated.

Bootstrap Results

Figure 3.3 shows the CFA bootstrap power and type 1 error curves. Figure 3.3A shows the power curves for outcome Y_t given the health effect is only associated with source 1 (S_t). The power ranges from 0.65 to 0.85. Figures 3.3B, 3.3C, and 3.3D show the type 1 error curves and they all are below 0.05 which is expected. Figure 4 shows the power curves for outcome Y_t given the health effect is only associated with source 2(S_2) and the patterns are the same.

Figures 3.5 and 3.6 show the two-stage bootstrap power and type 1 error curves. Figure 3.5A shows the power curves for outcome Y_{1t} given the health effect is only associated with source 1 (S_{1t}). The power estimates are near 100% for all coefficient

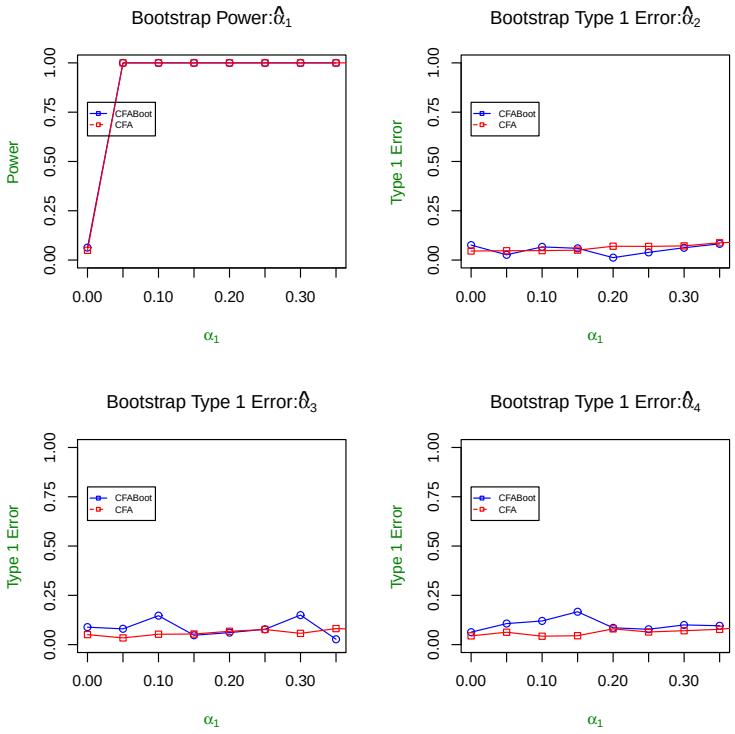


Figure 3.3: CFA Bootstrap Power and Type 1 Error Curves for $Y_{1t}|S_{1t}$ and $\mathbf{\Psi_{given}}$: A(top left): The power for each health effects model at the given initial value of α_{11}. B(top right): The type 1 error for each of the health effects models at a given initial value of α_{12}. C(bottom left): The type 1 error for each of the health effects models at a given initial value of α_{13}. D(bottom right): The type 1 error for each of the health effects models at a given initial value of α_{14}.

values which is the same as the non-bootstrapped values. Figures 3.5B, 3.5C, and 3.5D show the type 1 error curves and they all similar to the non-bootstrapped values which is expected. Figure 3.6 shows the power curves for outcome Y_{2t} given the health effect is only associated with source 2(S_{2t}) and the patterns are the same.

3.6 DISCUSSION

In this paper we evaluated the factor analysis and two-stage hierarchical model estimation procedures. Each model included all relevant exposures. Issues arise

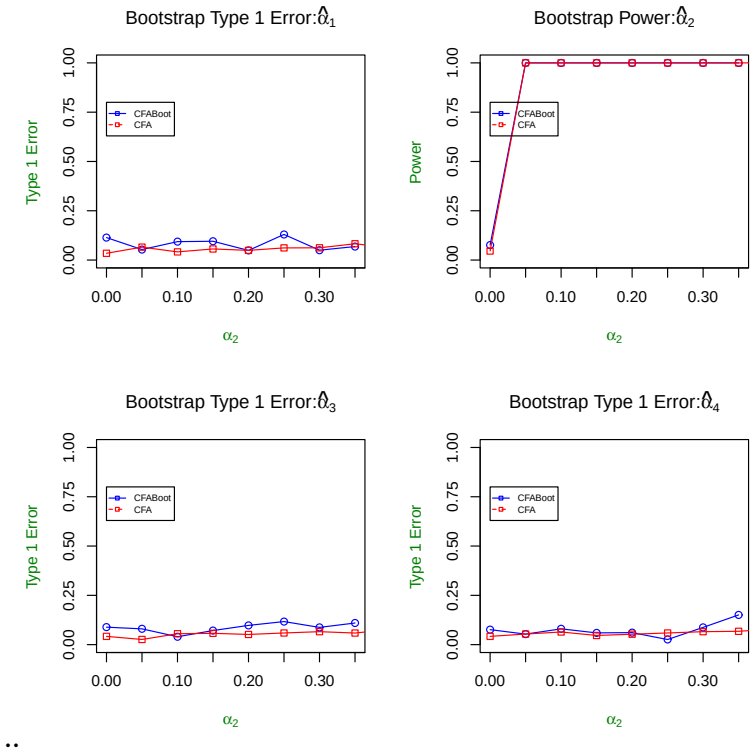
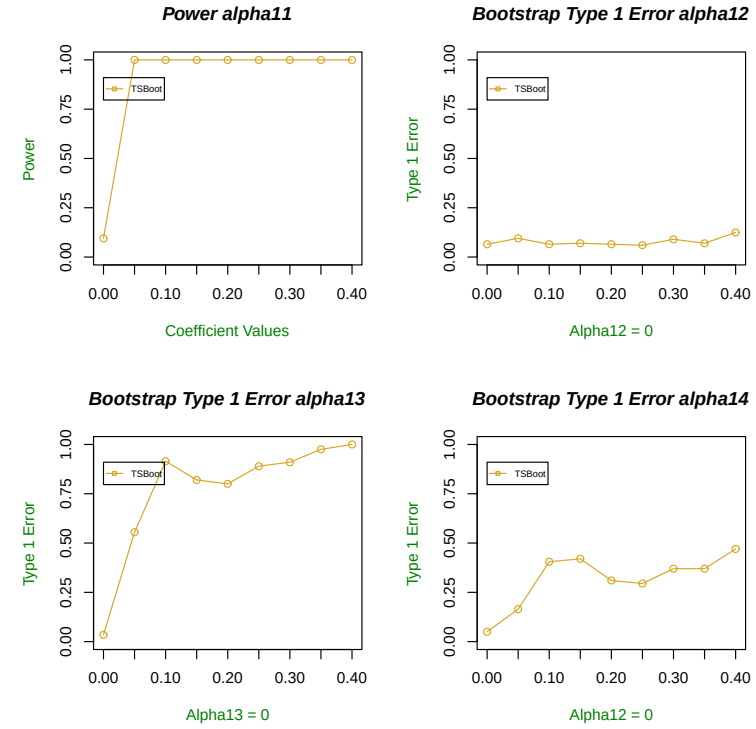


Figure 3.4: CFA Overlaid Power and Type 1 Error Curves for $Y_{2t}|S_{2t}$ and $\mathbf{\Psi_{given}}$: A(top left): The type 1 error for each of the health effects models at a given initial value of α_{21}. B(top right): The power for each health effects model at the given initial value of α_{22}. C(bottom left): The type 1 error for each of the health effects models at a given initial value of α_{23}. D(bottom right): The type 1 error for each of the health effects models at a given initial value of α_{24}.

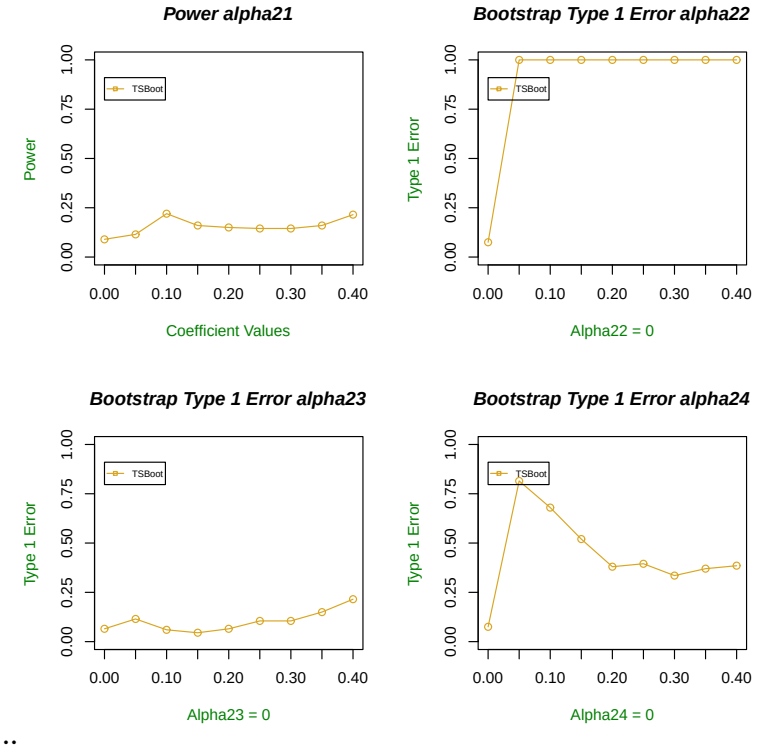
when estimating the effects of sources with no assumed effect on the outcome. The expectation is that the type 1 error rates are to be ≈ 0.05. For each of the factor analyses approaches the type I error rate was inflated for increasing values of the corresponding effect estimate but the differences are larger in the EFA and PCA models. This means that this source apportionment approach falsely rejects the null hypothesis more than the allotted $\alpha = 0.05$ for the sources that have no effect. These patterns are echoed when the variance structures of the source contributions are varied. Conducting a bootstrapped CFA model maintains the variability of the effect estimates and the type 1 error rates so there is no effect. The EFA and PCA bootstrapped models have similar results.



..

Figure 3.5: CFA Bootstrap Power and Type 1 Error Curves for $Y_{1t}|S_{1t}$ and $\mathbf{\Psi_{given}}$: A(top left): The power for each health effects model at the given initial value of α_{11}. B(top right): The type 1 error for each of the health effects models at a given initial value of α_{12}. C(bottom left): The type 1 error for each of the health effects models at a given initial value of α_{13}. D(bottom right): The type 1 error for each of the health effects models at a given initial value of α_{14}.

The two-stage hierarchical model also has issues when estimating effects from sources with no assumed association with the outcome. Much of the data from these models have instances where there are spikes in the type I error rates for low effect estimate values and then the rates go back to ≈ 0.05. The spikes occur when there are elements that occur predominantly in more than one source (overlap). The simulation study seemed to suggest that the two-stage approach where one element was associated with a particular source (no-overlap) yielded the best results. When the variance structures of the source contributions were changed, the type I error results from the two-stage models increased and were greater than 0.05. Conducting the two-stage bootstrapped model is ineffective be-



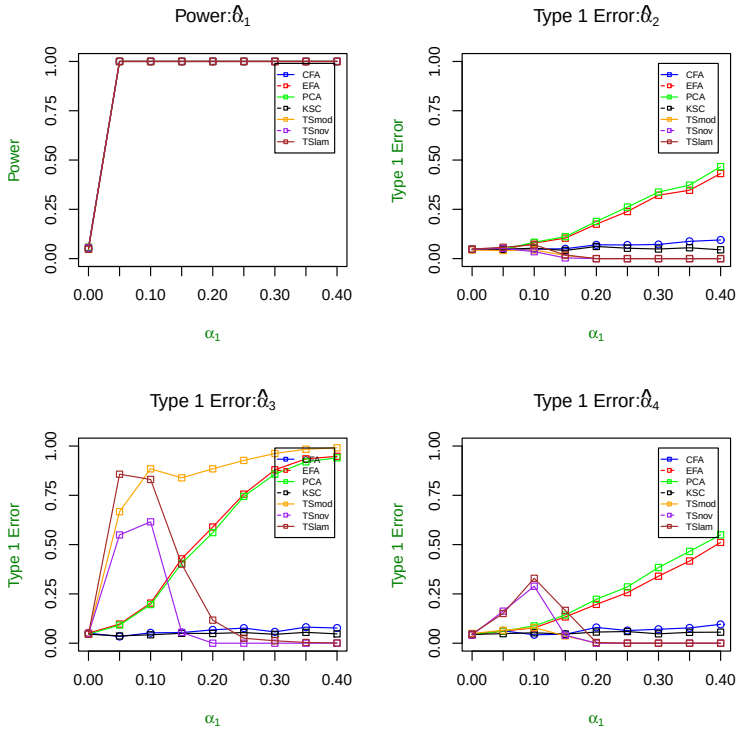
..

Figure 3.6: CFA Overlaid Power and Type 1 Error Curves for $Y_{2t}|S_{2t}$ and $\mathbf{\Psi_{given}}$: A(top left): The type 1 error for each of the health effects models at a given initial value of α_{21}. B(top right): The power for each health effects model at the given initial value of α_{22}. C(bottom left): The type 1 error for each of the health effects models at a given initial value of α_{23}. D(bottom right): The type 1 error for each of the health effects models at a given initial value of α_{24}.

cause the variability of the effect estimates did not need to be adjusted rather the estimates themselves were inaccurate.

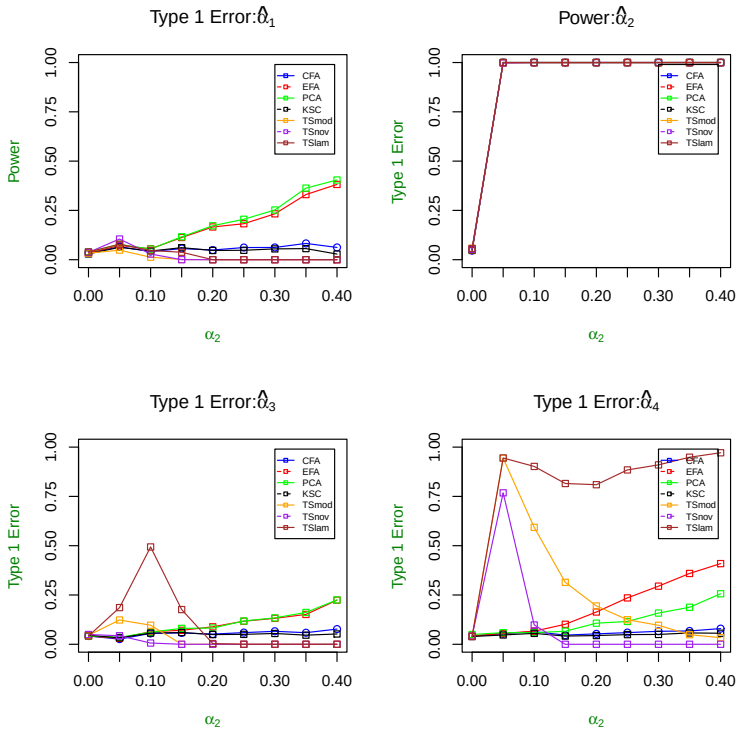
These findings would indicate that both of these modeling approaches have the ability to account for multiple exposures, estimate independent effects from correlated exposures, but each of the strategies has difficulty in accurately estimating the unobserved source contributions which consequently lead to health effects estimates with high false positive probabilities. More work is needed to ensure that proper control of false positives in empirical data settings.

3.7 APPENDIX



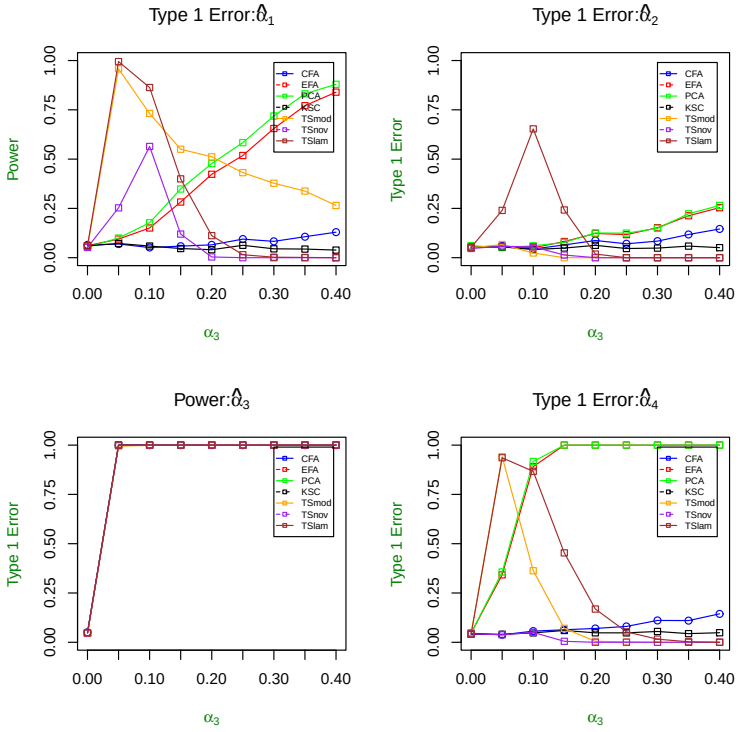
..

Figure 3.7: Overlaid Power and Type 1 Error Curves for $Y_t^{(1)}|\mathbf{S_t}$ and $\mathbf{\Psi_{given}}$: A(top left): The power for $\hat{\alpha}_1$ of each health effects model at the given initial value of α_1. B(top right): The type 1 error for $\hat{\alpha}_2$ of each health effects model at a given initial value of α_1. C(bottom left): The type 1 error for $\hat{\alpha}_3$ of each health effects model at a given initial value of α_1. D(bottom right): The type 1 error for $\hat{\alpha}_4$ of each health effects model at a given initial value of α_1.



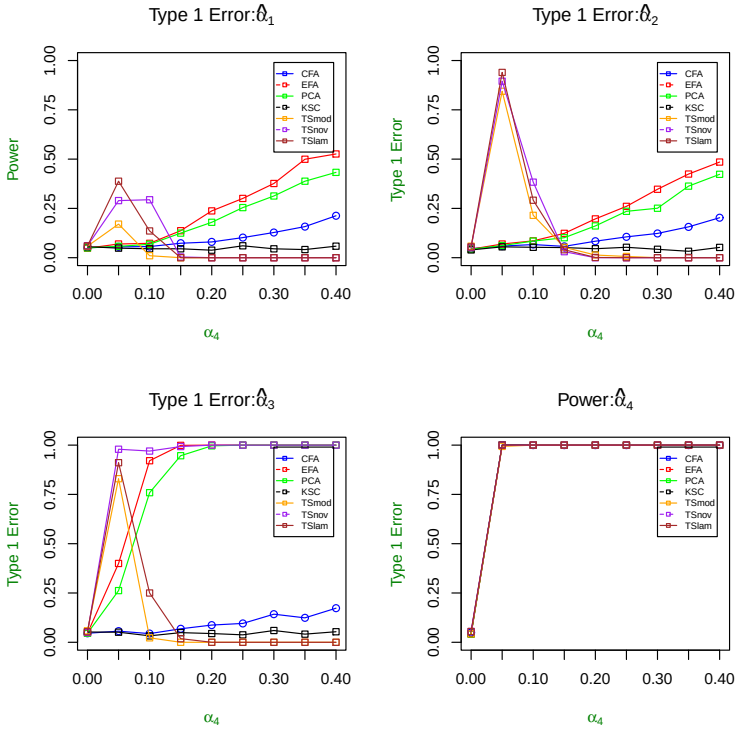
..

Figure 3.8: Overlaid Power and Type 1 Error Curves for $Y_t^{(2)}|\mathbf{S_t}$ and $\mathbf{\Psi_{given}}$: A(top left): The type 1 error for $\hat{\alpha}_1$ of each health effects model at the given initial value of α_2. B(top right): The power for $\hat{\alpha}_2$ of each health effects model at a given initial value of α_2. C(bottom left): The type 1 error for $\hat{\alpha}_3$ of each health effects model at a given initial value of α_2. D(bottom right): The type 1 error for $\hat{\alpha}_4$ of each health effects model at a given initial value of α_2.



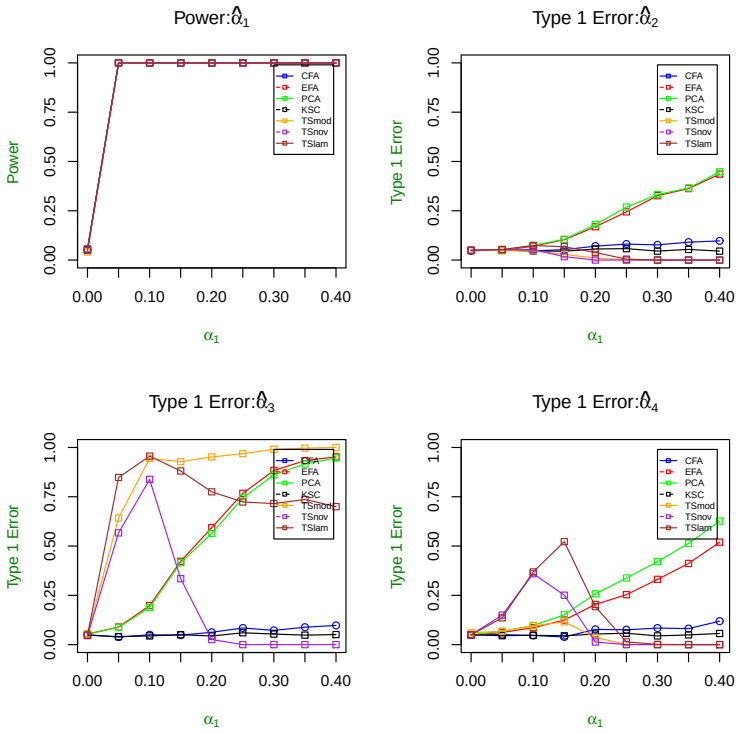
..

Figure 3.9: Overlaid Power and Type 1 Error Curves for $Y_t^{(3)}|\mathbf{S_t}$ and $\mathbf{\Psi_{given}}$: A(top left): The type 1 error for $\hat{\alpha}_1$ of each health effects model at the given initial value of α_3. B(top right): The type 1 error for $\hat{\alpha}_2$ of each health effects model at a given initial value of α_3. C(bottom left): The power for $\hat{\alpha}_3$ of each health effects model at a given initial value of α_3. D(bottom right): The type 1 error for $\hat{\alpha}_4$ of each health effects model at a given initial value of α_3.



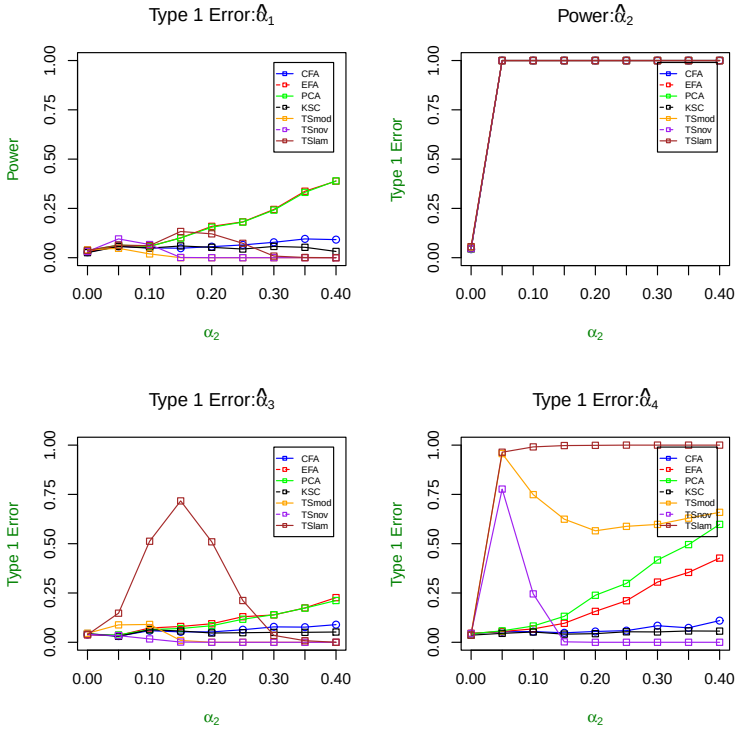
..

Figure 3.10: Overlaid Power and Type 1 Error Curves for $Y_t^{(4)}|\mathbf{S_t}$ and $\mathbf{\Psi_{given}}$: A(top left): The type 1 error for $\hat{\alpha}_1$ of each health effects model at the given initial value of α_4. B(top right): The type 1 error for $\hat{\alpha}_2$ of each health effects model at a given initial value of α_4. C(bottom left): The type 1 error for $\hat{\alpha}_3$ of each health effects model at a given initial value of α_4. D(bottom right): The power for $\hat{\alpha}_4$ of each health effects model at a given initial value of α_4.



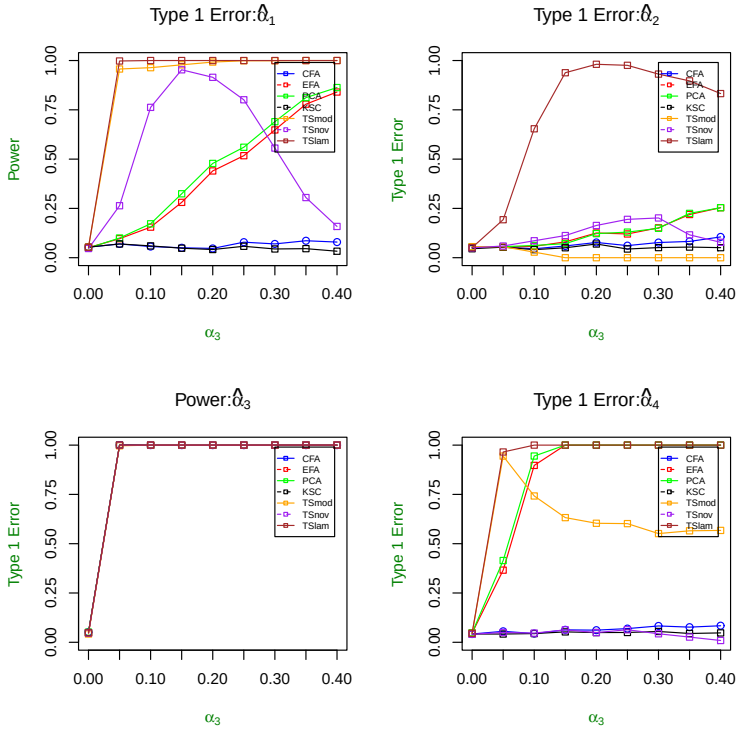
..

Figure 3.11: Overlaid Power and Type 1 Error Curves for $Y_t^{(1)}|\mathbf{S_t}$ and $\mathbf{\Psi_{13}}$: A(top left): The power for $\hat{\alpha}_1$ of each health effects model at the given initial value of α_1. B(top right): The type 1 error for $\hat{\alpha}_2$ of each health effects model at a given initial value of α_1. C(bottom left): The type 1 error for $\hat{\alpha}_3$ of each health effects model at a given initial value of α_1. D(bottom right): The type 1 error for $\hat{\alpha}_4$ of each health effects model at a given initial value of α_1.



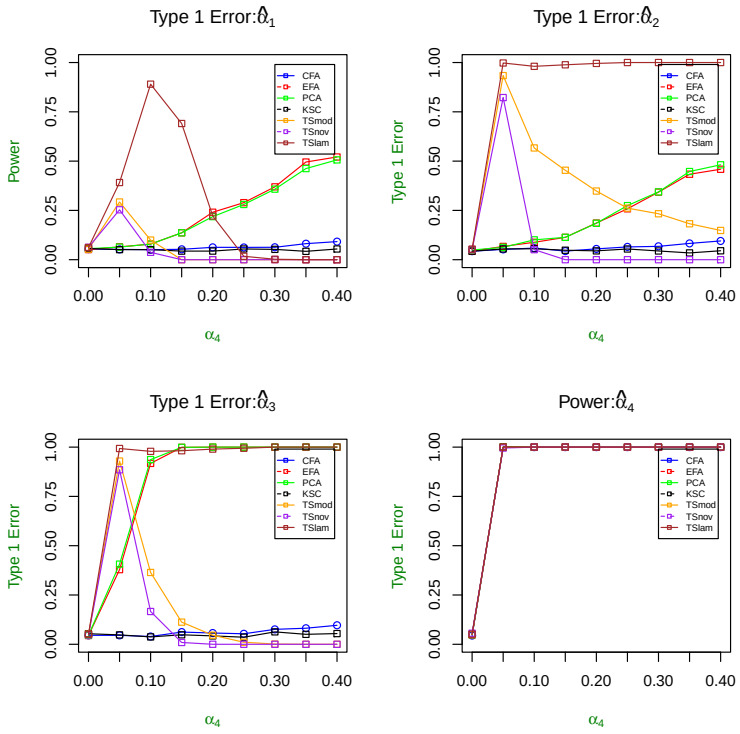
..

Figure 3.12: Overlaid Power and Type 1 Error Curves for $Y_t^{(2)}|\mathbf{S_t}$ and $\mathbf{\Psi_{13}}$: A(top left): The type 1 error for $\hat{\alpha}_1$ of each health effects model at the given initial value of α_2. B(top right): The power for $\hat{\alpha}_2$ of each health effects model at a given initial value of α_2. C(bottom left): The type 1 error for $\hat{\alpha}_3$ of each health effects model at a given initial value of α_2. D(bottom right): The type 1 error for $\hat{\alpha}_4$ of each health effects model at a given initial value of α_2.



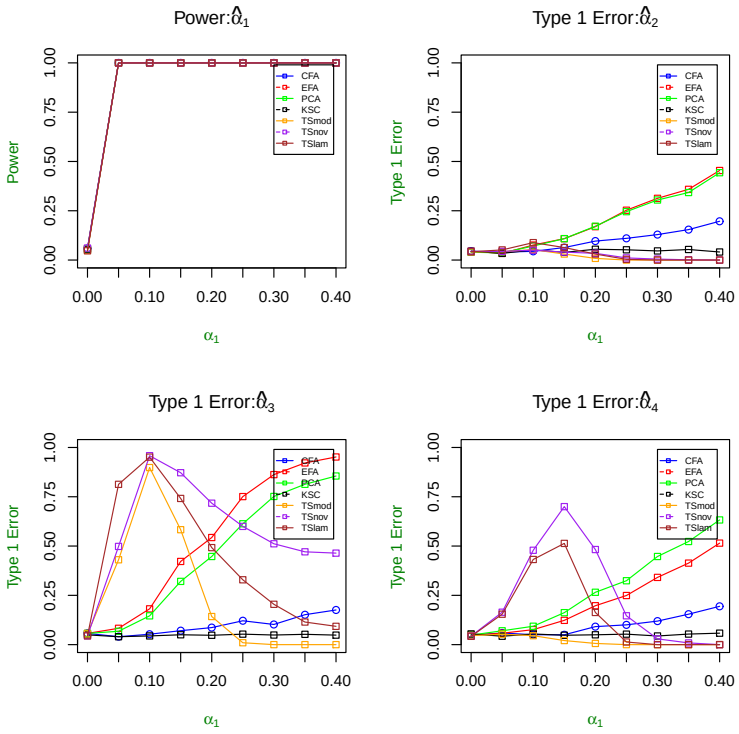
..

Figure 3.13: Overlaid Power and Type 1 Error Curves for $Y_t^{(3)}|\mathbf{S_t}$ and $\mathbf{\Psi_{13}}$: A(top left): The type 1 error for $\hat{\alpha}_1$ of each health effects model at the given initial value of α_3. B(top right): The type 1 error for $\hat{\alpha}_2$ of each health effects model at a given initial value of α_3. C(bottom left): The power for $\hat{\alpha}_3$ of each health effects model at a given initial value of α_3. D(bottom right): The type 1 error for $\hat{\alpha}_4$ of each health effects model at a given initial value of α_3.



..

Figure 3.14: Overlaid Power and Type 1 Error Curves for $Y_t^{(4)}|\mathbf{S_t}$ and $\mathbf{\Psi_{13}}$: A(top left): The type 1 error for $\hat{\alpha}_1$ of each health effects model at the given initial value of α_4. B(top right): The type 1 error for $\hat{\alpha}_2$ of each health effects model at a given initial value of α_4. C(bottom left): The type 1 error for $\hat{\alpha}_3$ of each health effects model at a given initial value of α_4. D(bottom right): The power for $\hat{\alpha}_4$ of each health effects model at a given initial value of α_4.



..

Figure 3.15: Overlaid Power and Type 1 Error Curves for $Y_t^{(1)}|\mathbf{S_t}$ and $\mathbf{\Psi_{31}}$: A(top left): The power for $\hat{\alpha}_1$ of each health effects model at the given initial value of α_1. B(top right): The type 1 error for $\hat{\alpha}_2$ of each health effects model at a given initial value of α_1. C(bottom left): The type 1 error for $\hat{\alpha}_3$ of each health effects model at a given initial value of α_1. D(bottom right): The type 1 error for $\hat{\alpha}_4$ of each health effects model at a given initial value of α_1.

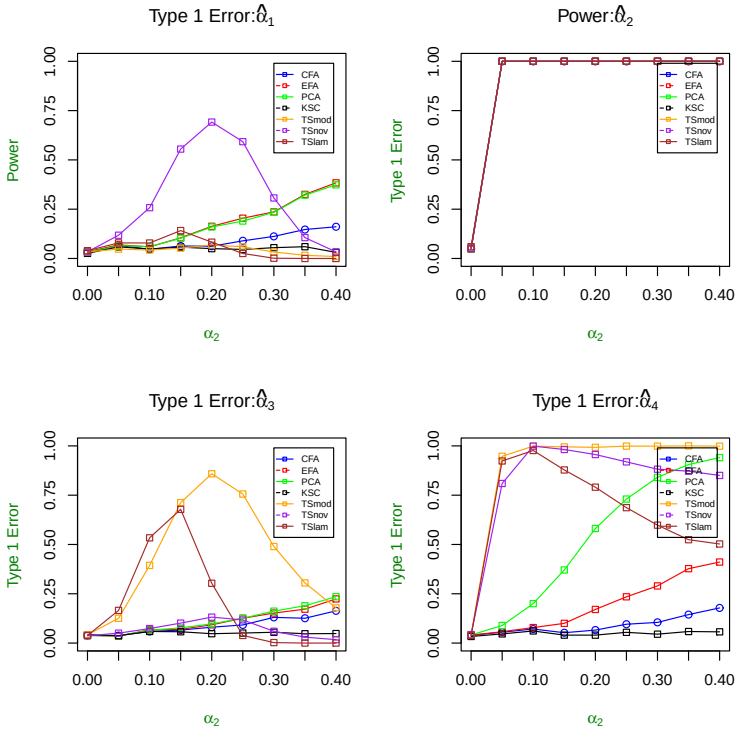


Figure 3.16: Overlaid Power and Type 1 Error Curves for $Y_t^{(2)}|\mathbf{S_t}$ and $\mathbf{\Psi_{31}}$: A(top left): The type 1 error for $\hat{\alpha}_1$ of each health effects model at the given initial value of α_2. B(top right): The power for $\hat{\alpha}_2$ of each health effects model at a given initial value of α_2. C(bottom left): The type 1 error for $\hat{\alpha}_3$ of each health effects model at a given initial value of α_2. D(bottom right): The type 1 error for $\hat{\alpha}_4$ of each health effects model at a given initial value of α_2.

$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.05, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = 0$

Coef.	TSov Std Err.	TSov Std Dev.
$\hat{\alpha_1}$	0.0040	0.0030
$\hat{\alpha_2}$	0.0049	0.0035
$\hat{\alpha^3}$	0.0058	0.0045
$\hat{\alpha^4}$	0.0085	0.0069

$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.0, \alpha_2 = 0.05, \alpha_3 = 0, \alpha_4 = 0$

Coef.	TSov Std Err.	TSov Std Dev.
$\hat{\alpha_1}$	0.0052	0.0028
$\hat{\alpha_2}$	0.0062	0.0038
$\hat{\alpha^3}$	0.0071	0.0046
$\hat{\alpha^4}$	0.0101	0.0068

$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.30, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = 0$

Coef.	TSov Std Err	TSov Std Dev.
$\hat{\alpha_1}$	0.01906	0.0039
$\hat{\alpha_2}$	0.0222	0.0050
$\hat{\alpha^3}$	0.0245	0.0050
$\hat{\alpha^4}$	0.0336	0.0073

$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.0, \alpha_2 = 0.30, \alpha_3 = 0, \alpha_4 = 0$

Coef.	TSov Std Err.	TSov Std Dev.
$\hat{\alpha_1}$	0.0228	0.0039
$\hat{\alpha_2}$	0.0266	0.0049
$\hat{\alpha^3}$	0.0293	0.0050
$\hat{\alpha^4}$	0.0402	0.0071

Table 3.6: This table represents the parameter estimates and errors for the 2-stage "overlap" models conducted on simulated data. Generalized linear models were conducted for 1000 iterations. The subsequent estimates were aggregated into medians, and 95% CI's. Each model included no covariates.

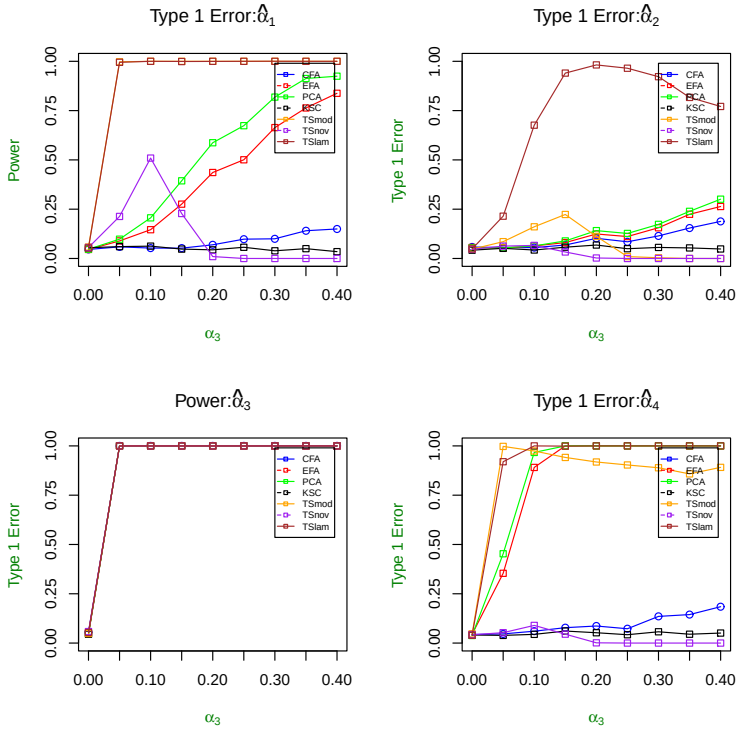


Figure 3.17: Overlaid Power and Type 1 Error Curves for $Y_t^{(3)}|\mathbf{S_t}$ and $\mathbf{\Psi_{31}}$: A(top left): The type 1 error for $\hat{\alpha}_1$ of each health effects model at the given initial value of α_3. B(top right): The type 1 error for $\hat{\alpha}_2$ of each health effects model at a given initial value of α_3. C(bottom left): The power for $\hat{\alpha}_3$ of each health effects model at a given initial value of α_3. D(bottom right): The type 1 error for $\hat{\alpha}_4$ of each health effects model at a given initial value of α_3.

$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.05, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = 0$

Coef.	TSnov Std Err.	TSnov Std Dev.
$\hat{\alpha_1}$	0.0008	0.0007
$\hat{\alpha_2}$	0.0011	0.0011
$\hat{\alpha^3}$	0.0017	0.0017
$\hat{\alpha^4}$	0.0014	0.0014

$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.0, \alpha_2 = 0.05, \alpha_3 = 0, \alpha_4 = 0$

Coef.	TSnov Std Err.	TSnov Std Dev.
$\hat{\alpha_1}$	0.0007	0.0008
$\hat{\alpha_2}$	0.0011	0.0012
$\hat{\alpha^3}$	0.0018	0.0017
$\hat{\alpha^4}$	0.0014	0.0016

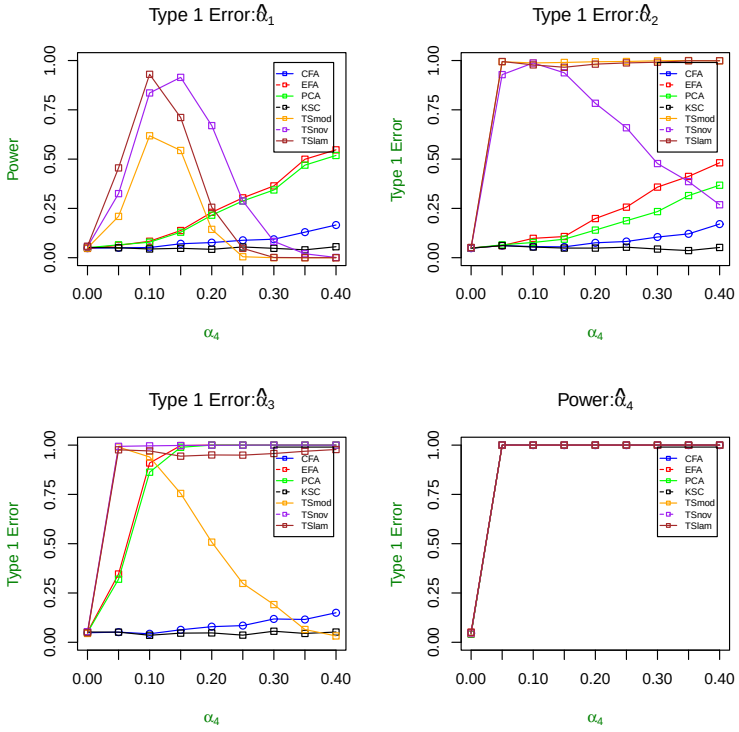
$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.30, \alpha_2 = 0, \alpha_3 = 0, \alpha_4 = 0$

Coef.	TSnov Std Err	TSnov Std Dev.
$\hat{\alpha_1}$	0.0163	0.0009
$\hat{\alpha_2}$	0.0163	0.0016
$\hat{\alpha^3}$	0.0232	0.0027
$\hat{\alpha^4}$	0.0189	0.0021

$\mathbf{Y}_t|\mathbf{S}_t, \alpha_1 = 0.0, \alpha_2 = 0.30, \alpha_3 = 0, \alpha_4 = 0$

Coef.	TSnov Std Err.	TSnov Std Dev.
$\hat{\alpha_1}$	0.0188	0.0010
$\hat{\alpha_2}$	0.0189	0.0017
$\hat{\alpha^3}$	0.0267	0.0026
$\hat{\alpha^4}$	0.0218	0.0022

Table 3.7: This table represents the parameter estimates and errors for the 2-stage "no-overlap" models conducted on simulated data. Generalized linear models were conducted for 1000 iterations. The subsequent estimates were aggregated into medians, and 95% CI's. Each model included no covariates.



..

Figure 3.18: Overlaid Power and Type 1 Error Curves for $Y_t^{(4)}|\mathbf{S_t}$ and $\mathbf{\Psi_{31}}$: A(top left): The type 1 error for $\hat{\alpha}_1$ of each health effects model at the given initial value of α_4. B(top right): The type 1 error for $\hat{\alpha}_2$ of each health effects model at a given initial value of α_4. C(bottom left): The type 1 error for $\hat{\alpha}_3$ of each health effects model at a given initial value of α_4. D(bottom right): The power for $\hat{\alpha}_4$ of each health effects model at a given initial value of α_4.

References

- Bell Ml, Davis DL., (2001). Reassessment of the Lethal London Fog of 1952: Novel Indicators of Acute and Chronic Consequences of Acute Exposure to Air Pollution. *Environmental Health Perspectives*. **109**(3): 389-394.

- Berger A, Zareba W, Schneider A, Ruckerl R, Ibald-Mulli A, Cyrys J, Wichmann HE, Peters A (2006). Runs of ventricular and supra ventricular tachycardia triggered by air pollution in patients with coronary heart disease (2006). *Journal of Occupational and Environmental Medicine*. **48**(11): 1149-1158.

- Brook RD, Franklin B, Cascio W, Hong Y, Howard G, Lipsett M, Luepker R, Mittleman M, Samet J, Smith SC Jr, Tager I (2004). Air pollution and cardiovascular disease: a statement for healthcare professionals from the Expert Panel on Population and Prevention Science of the American Heart Association. *Circulation*. 2004; **109**: 26552671.

- Carroll RJ , Ruppert D, Stefanski LA (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall/CRC. 1995: 40-78,141-162.

- Cascio WE (2005). Cardiopulmonary Health Effects of Air Pollution: Is a Mechanism Emerging? *American Journal of Respiratory and Critical Care Medicine*. 2005; **172** (12): 1482-1484.

- Chatfield C (2004). *The Analysis of Time Series Analysis; An Introduction*. Chapman and Hall/CRC: 33-50, 202-211.

- Chuang KJ, Coull BA, Zanobetti A, Suh H, Schwartz J, Stone PH, Litonjua A, Speizer FE, Gold DR (2008). Particulate Air Pollution as a Risk Factor for ST-segment Depression. *Circulation*. in press.

- Crainiceanu C, David Ruppert, M. P. Wand(2005). Bayesian Analysis for Penalized Spline Regression Using WinBUGS. *JSS Journal of Statistical Software*. October 2005 **14**(14). http://www.jstatsoft.org/

- Dockery DW (2001). Epidemiologic Evidence of Cardiovascular Effects of Particulate Air Pollution. *Environmental Health Perspectives*. **109**(4):483-486.

- Dockery DW, Pope CA, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BG, Speizer FE (1993). An Association between Air Pollution and Mortality in Six U.S. Cities. *New England Journal of Medicine*. **329** (24): 1753-1759.

- Dockery DW, Luttmann-Gibson H, Rich DQ, Link MS, Schwartz JD, Gold DR, Koutrakis P, Verrier RL, Mittleman MA (2005). Particulate air pollution and nonfatal cardiac events. Part II. Association of air pollution with confirmed arrhythmias recorded by implanted defibrillators.*Research report (Health Effects Institute)*. **124**: 83-126, discussion 127-148.

- Dockery DW, Luttmann-Gibson H, Rich DQ, Link MS, Mittleman MA, Gold DR, Koutrakis P, Schwartz JD, Verrier RL (2005). Association of air pollution with increased incidence of ventricular tachyarrhythmias recorded by implanted cardioverter defibrillators. *Environmental Health Perspectives*. 2005;**113**(6):670-674.

- Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties (with discussion) (1996). *Statistical Science*. 1996; **11**: 89-121.

- Fox J (2006). Structural Equation Modeling with the SEM Package in R. *Structural Equation Modeling*. **13**(3): 465-486.

- Gold DR, Litonjua A, Schwartz J, Lovett E, Larson A, Nearing B. (2000). Ambient pollution and heart rate variability. *Circulation*. **101**: 1267-1273.

- Gold DR, Litonjua AA, Zanobetti A, Coull BA, Schwartz J, MacCallum G, Verrier RL, Nearing BD, Canner MJ, Suh H, Stone PH (2005). Air pollution and ST-segment depression in elderly subjects. *Environmental Health Perspectives*. **113**: 883 887.

- Greenland S (1993). Methods for Epidemiologic Analysis if Multiple Exposures: A Review and Comparative Study of Maximum Likelihood, Preliminary-Testing, and Empirical-Bayes Regression. *Statistics in Medicine*. **12**: 717-736.

- Guttikunda S (2009). Urban Particulate Pollution Source Apportionment: Part I Definition, Methodology, and Resources. *Simple Interactive Models for Better Air Quality*: SIM-air Working paper Series,**16**.

- Guttikunda S (2009). Urban Particulate Pollution Source Apportionment: Part II Applications, Results, and Policy Implications. *Simple Interactive Models for Better Air Quality*: SIM-air Working paper Series, **23**.

- Hardin J, Schmiediche H, Carroll RJ (2003). The Regression-Calibration Method for Fitting Generalized Linear Models with Additive Measurement Error.*The Stata Journal* . **3**(4): 361372.

- Hatcher L (1994). A Step by Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling. *SAS Institute Inc*.: 141-225.

- Hopke P, Ito K, Mar T, Christensen W, Eatough D, Henry R, Kim E, Laden F, Lall R, Larson T, Liu H, Neas L, Pinto J, Stlozel M, Suh H, Paatero P, Thurston G (2006) PM source apportionment and health effects: 1. Intercomparison of

source apportionment results. *Journal of Exposure Science and Environmental Epidemiology*. **16**: 275 - 286.

- Ito K, Christensen W, Eatough D, Henry R, Kim E, Laden F, Lall R, Larson T, Neas L, Hopke P, Thurston G (2006). PM source apportionment and health effects: 2. An investigation of intermethod variability in associations between source-apportioned fine particle mass and daily mortality in Washington, DC. *Journal of Exposure Science and Environmental Epidemiology*. **16**: 300 - 310.

- Klepeis N, Nelson W, Ott W. The National Human Activity Pattern Survey (NHAPS) A Resource for Assessing Exposure to Environmental Pollutants. *Journal of Exposure Analysis and Environmental Epidemiology*. **11**: 231252.

- Laden F, Neas LM, Dockery DW (2000). Association of fine particulate matter from different sources with daily mortality in six US cities. *Environmental Health Perspectives*. **108**(10): 941-947.

- Laden F, Schwartz J, Speizer F, Dockery D (2006). Reduction in fine particulate air pollution and mortality: extended follow-up of the Harvard Six Cities Study. *American Journal of Respiratory and Critical Care Medicine*. **173**: 667-672.

- Lall R, Ito K, Thurston G (2011). Distributed Lag Analyses of Daily Hospital Admissions and Source-Apportioned Fine Particle Air Pollution. *Environmental Health Perspectives* . **119**: 455 - 460.

- Mar T, Ito K, Koenig J, Larson T, Eatough D, Henry R, Kim E, Laden F, Lall R, Neas L, Stlozel M, Paatero P, Hopke P, Thurston G (2006). PM source apportionment and health effects. 3. Investigation of inter-method variations in associations between estimated source contributions of $PM_{2.5}$ and daily

mortality in Phoenix, AZ. *Journal of Exposure Science and Environmental Epidemiology*. **16**: 311 - 320.

- Maynard D, Coull BA, Gryparis A, Schwartz J (2007). Mortality risk associated with short-term exposure to traffic particles and sulfates. *Environmental Health Perspectives*. **115**(5): 751-755.

- Mittleman MA, Verrier RL (2003). Air pollution: small particles, big problems? *Epidemiology*. **14**: 512513.

- Momoli F, Abrahammowicz M, Parent M, Krewski D, Siemiatyckt J (2010) Analysis of Multiple Exposures: An Empirical Comparison of Results From Conventional ans Semi-Bayes Modeling Strategies. *Epidemiology*. **21**(1).

- National Research Council (1997). Use of the Gray Literature and Other Data in Environmental Epidemiology, Committee on Environmental Epidemiology and Commission on Life Sciences,, NATIONAL ACADEMY PRESS, Washington, D.C., Environmental Epidemiology,Volume 2.

- Nearing BD, Verrier RL (2002). Modified moving average analysis of T-wave alternans to predict ventricular fibrillation with high accuracy.*Jornal of Applied Physiology*. **92**(2): 541-549.

- Nieminen T, Lehtimaki T, Viik J, Lehtinen R, Nikus K, Koobi T, Niemela K, Turjanmaa V, Kaiser W, Huhtala H, Verrier RL, Huikuri H, Kahonen M (2007). T-wave alternans predicts mortality in a population undergoing a clinically indicated exercise test. *European Heart Jornal*. **28**(19): 2332-2337.

- Nikolov M, Coull B, Catalano P, Godleski J (2007). An Informative Bayesian Structural Equation Model to Assess Source-Specific Health Effects of Air Pollution.*Biostatistics*. **8**(3): 609-624.

- Peters A, Liu E, Verrier RL, Schwartz J, Gold DR, Mittleman M, Baliff J, Oh JA, Allen G, Monahan K, Dockery DW (2000). Air pollution and incidence of cardiac arrhythmia. *Epidemiology*. **11**(1): 11-17.

- Peters A, von Klot S, Heier M, Trentinaglia I, Hormann A, Wichmann HE, Lowel H (2004). Exposure to traffic and the onset of myocardial infarction. *New England Journal of Medicine*. **351**(17): 1721-1730.

- Pope CA, 3rd, Dockery DW (1999). Epidemiology of particle effects. In: Air pollution and health (Holgate ST, Koren HS, Samet JM, Maynard RL, eds). London: *Academic Press*: 673-705.

- Pope CA, 3rd, Dockery DW (1991). Respiratory Health and PM_{10} Pollution: A Daily Time Series Analysis. *American Review of Respiratory Disease*. **144**: 668-674.

- Pope CA, 3rd, Hansen ML, Long RW, Nielsen KR, Eatough NL, Wilson WE, et al. (2004). Ambient particulate air pollution, heart rate variability, and blood markers of inflammation in a panel of elderly subjects. *Environmental health perspectives*. **112**: 339-345.

- Pope CA, 3rd, Burnett R, Thurston GD, Thun MJ, Calle EE, Krewski D, Godleski JJ (2004). Cardiovascular Mortality and Long-Term Exposure to Particulate Air Pollution. *Circulation*. **109**: 71-77.

- Rich DQ, Schwartz J, Mittleman MA, Link M, Luttmann-Gibson H, Catalano PJ, Speizer FE, Dockery DW (2005). Association of short-term ambient air pollution concentrations and ventricular arrhythmias. *American Journal of Epidemiology*. **161**(12): 1123-1132.

- Ruppert D, Wand MP, Carroll RJ (2003).Semiparametric Regression. *Cambridge University Press*. 2003: 57-185.

- Schwartz J (1994). What are people dying of on high air pollution days? *Environmental Research*. **64**(1): 26-35.

- Schwartz J, Litonjua A, Suh H, Verrier M, Zanobetti A, Syring M, Nearing B, Verrier R, Stone P, MacCallum G, Speizer FE, Gold DR (2005). Traffic related pollution and heart rate variability in a panel of elderly subjects. *Thorax*. **60**: 455 461.

- Skrondal A, Rabe-Hasketh S, (2004). Generalized Latent variable Modeling; Multilevel, Longitudinal, and Structural Equation Models. *Chapman and Hall CRC*: 49-93, 221-249.

- Speigelman D, McDermott A, Rosner B, (1997). Regression Calibration Method for Correcting Measurement-Error Bias In Nutritional Epidemiology. *American Journal of Clinical Nutrition*. **65** (suppl): 1179S-86S.

- Stein PK, Sanghavi D, Domitrovich PP, Mackey RA, Deedwania P (2008). Ambulatory ECG-Based T-Wave Alternans Predicts Sudden Cardiac Death in High-Risk Post-MI Patients with Left Ventricular Dysfunction in the EPH-ESUS Study. *Journal of Cardiovascular Electrophysiology*.

- Suh H, Zanobetti A, Schwartz J, Coull B (2011). Associations Between the Chemical properties of Air Pollution and Cause-Specific Hospital Admissions in Atlanta, GA.

- Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology (1996). Heart rate variability: Standards of

measurement, physiological interpretation, and clinical use. *Circulation*: 1043 - 1065.

- Thomas D, Witte J, Greenland S (2007). Dissecting Effects of Complex Mixtures: Who's Afraid of Complex Mixtures? *Epidemiology*. **18**(2).

- Thurston G, Ito K, Lall R (2011). A source apportionment of U.S. fine particulate matter air pollution. *Atmospheric Environment* . **45**: 3924-3936.

- Van der Leeuw J (1993). The Covariance Matrix of ARMA Errors in Closed Form, *Journal of Econometrics*. **63**: 397-405.

- Witte J, Greenland S, Kim L, Arab L (2000). Multilevel Modeling in Epidemiology with Glimmix. *Epidemiology*. **11**(6).

- Witte J, Greenland S, Kim L (1998). Software for Hierarchical Modeling of Epidemiologic Data. *Epidemiology*. **9**(5).

- Witte J, Greenland S (1996). Simulation Study of Hierarchical Regression. *Statistics In Medicine*. **15**: 1161-1170.

- Young J, Glass T, Bernasconi E, Rickenbach M, Furrer H, Hirschel B, Tarr P, Vernazza P, Battegay M, Bucher H (2009). Hierarchical Modeling Gave Plausible Estimates of Associations Between Metabolic Syndrome and Components of Antiretroviral Therapy. *Journal of Clinical Epidemiology*. **62**: 632-641.

- Zanobetti A, Gold D, Stone P, Suh H, Schwartz J, Coull B, Speizer F (2009). Reduction in Heart Rate Variability with Traffic and Air Pollution in Patients with Coronary Artery Disease. *Environmental Health Perspectives*. **118** (3).

- Zanobetti A, Stone P, Speizer F, Schwartz J, Coull B, Suh H, Nearing B, Mittleman M, Verrier R, Gold D (2009). T-Wave Alternans, Air Pollution and Traffic

in High-Risk Subjects. *American Journal of Cardiology*. **104**: 665 - 670.

- Zanobetti A, Wand MP, Schwartz J, Ryan LM (2000) . Generalized Additive Distributed Lag Models: Quantifying Mortality Displacement. *Biostatistics*. **1**(3): 279-292.

- Zeger S, Thomas D, Dominici, Samet F, Schwartz J, Dockery D, Cohen A (2000). Exposure Measurement Error in Time-Series Studies of Air Pollution:Concepts and Consequences. *Environ Healt Persptives*. **108**: 419-426.

- Zhou J, Ito K, Lall R, Lippman M, Thurston G (2011). Time-Series Analysis of Mortality Effects of Fine Particulate Matter Components in Detroit and Seattle. *Environmental Health Perspectives* **119**: 461-466 (2011).