DISSERTATION

BAYESIAN TREE BASED METHODS FOR LONGITUDINALLY ASSESSED ENVIRONMENTAL MIXTURES

Submitted by

Seongwon Im

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2024

Doctoral Committee:

Advisor: Ander Wilson

Kayleigh Keller
Matt Koslovsky
Andreas Neophytou

Copyright by Seongwon Im 2024

All Rights Reserved

ABSTRACT

BAYESIAN TREE BASED METHODS FOR
LONGITUDINALLY ASSESSED ENVIRONMENTAL MIXTURES

In various fields, there is interest in estimating the lagged association between an exposure and an outcome. This is particularly common in environmental health studies, where exposure to an environmental chemical is measured repeatedly during gestation for the assessment of its lagged effects on a birth outcome. The relationship between longitudinally assessed environmental mixtures and a health outcome is also of greater interest. For a single exposure, a distributed lag model (DLM) is a widely used method that provides an appropriate temporal structure for estimating the time-varying effects. For mixture exposures, a distributed lag mixture model is used to address the main effect of each exposure and lagged interactions among exposures. The main inferential goals include estimating the lag-specific effects and identifying a window of susceptibility, during which a fetus is particularly vulnerable.

In this dissertation, we propose novel statistical methods for estimating exposure effects of longitudinally assessed environmental mixtures in various scenarios. First, we propose a method that can estimate a linear exposure-time-response function between mixture exposures and a count outcome that may be zero-inflated and overdispersed. To achieve this, we employ a Bayesian Pólya-Gamma data augmentation with a treed distributed lag mixture model framework. We apply the method to estimate the relationship between weekly average fine particulate matter ($PM_{2.5}$) and temperature and pregnancy loss with live-birth identified conception time series design with administrative data from Colorado. Second, we propose a tree triplet structure to allow for heterogeneity in exposure effects in an environmental mixture exposure setting. Our method accommodates modifier and exposure selection, which allows for personalized and subgroup-specific effect estimation and windows of susceptibility identification. We apply the method to Colorado administrative

birth data to examine the heterogeneous relationship between $PM_{2.5}$ and temperature and birth weight. Finally, we introduce an R package **dlmtree** that integrates tree structured DLM methods into convenient software. We provide an overview of the embedded tree structured DLMs and use simulated data to demonstrate a model fitting process, statistical inference, and visualization.

ACKNOWLEDGEMENTS

It has been a long journey since I decided to pursue statistics as my area of expertise, and there have definitely been ups and downs in this graduate program. Many things were as expected when I began, while some were completely different from what I anticipated. Reflecting on the twists and turns of my experience, I am grateful for the support and guidance of those who have contributed to my growth and success. Here, I would like to acknowledge and thank the individuals who have made my journey toward the completion of my doctoral degree truly fruitful.

Beyond just my time in the graduate program, my family's support has been crucial in shaping my entire life up to this point. Without their unwavering encouragement and love, I would not be where I am today, both personally and academically. My parents, Younghee Baek and Jongtae Im, always believed in me, supporting me all the way from South Korea, and my brother, Hyowon "Joey" Im, in Seattle, was an incredible companion who made sure I never felt alone during my time in the States. I cannot express enough gratitude for their consistent love and support.

My PhD journey began with my advisor's interview phone call and concluded with him hooding me. I am deeply thankful to my advisor, Dr. Ander Wilson, for his exceptional guidance and support. His expertise has taught me how to think critically, ask insightful questions, and solve problems as a statistician. I truly appreciate his patience and dedication to his mentorship, which has greatly contributed to my growth and development in the field of statistics.

I wish to extend my thanks to the amazing mentors I had during the graduate program. I am grateful to Dr. Zach Weller for my first valuable experience in statistical consulting and publishing a research paper, as well as serving as a connecting bridge to my advisor. Additionally, I would like to thank Dr. Daniel Mork for his role as a wonderful friend, mentor, and collaborator during my research. Collaborating with him, given his positive morale, made my research experience enjoyable, and his forward-thinking advice was always greatly helpful. I also express my gratitude to my committee members, Dr. Kayleigh Keller, Dr. Matthew "Matt" Koslovsky, and Dr. Andreas Neophytou for constructive discussions and suggestions for improvement on my research work.

Their teaching and insights have significantly improved my statistical expertise for more in-depth research.

Miraculously, I met such amazing people, 6,110 miles away from home, exactly during this period. Had it not been for my wonderful, extraordinary, and charming friends: Austin Ellingworth, Lane Drew, Liz Lawler, Dr. Connor Gibbs, Julia Campbell, Simon Weller, Dr. Mantautas Rimkus, Vaida Petraviciute, Dr. Nathan Ryder, Joy Ryder, Dr. Ian Taylor, Ginny Taylor, and Karissa Palmer, my experience at CSU would not have been as adventurous, invigorating, or heartwarming. I cannot thank them enough for the precious and unforgettable memories they have created for me.

Finally, I would like to express my special thanks to Dr. Patti Frazer Lock for igniting my passion for statistics and its philosophy through her exciting introductory statistics course at St. Lawrence University. Thanks to Patti, I was privileged to discover a field to which I could dedicate myself and gained the motivation to reach this point.

This work was supported by National Institutes of Health grants ES029943 and ES028811. This work would not have been possible without the support of the National Institutes of Health and the Colorado Department of Public Health and Environment. These data were supplied by the Center for Health and Environmental Data Vital Statistics Program of the Colorado Department of Public Health and Environment, which specifically disclaims responsibility for any analyses, interpretations, or conclusions it has not provided.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Individuals are exposed to a complex mixture of chemicals. One common source of these chemicals is air pollution, which contains many harmful pollutants, including particulate matter with a diameter of less than 2.5 micrometers ($PM_{2.5}$) or 10 micrometers (PM_{10}), nitrogen dioxide (NO_2), and ozone (O_3). Exposure to these chemicals and other ambient exposures has been shown to be detrimental to a variety of health endpoints (Crouse et al., 2015; Di et al., 2017; Liu et al., 2019; Dearborn et al., 2023). Notably, there has been rising concern with ambient $PM_{2.5}$ and extreme heat due to climate change and increased frequency of wildfire smoke (Vose et al., 2017; Burke et al., 2023). Increased exposure to $PM_{2.5}$ and the extreme heat has been associated with their damage to human health such as increased hospitalization and mortality (Medina-Ramón et al., 2006; Karlsson and Ziebarth, 2018; Bell et al., 2024).

1.1 Prenatal exposure and windows of susceptibility

The effects of exposure to air pollution begin as early as conception. Fetal exposure to environmental chemicals occurs through maternal exposure as toxins can penetrate the mother's placenta, which may lead to oxidative stress that negatively affects the developing fetus (Wright, 2017). This early exposure can result in adverse effects on the birth outcome, children's health, and potentially health into adulthood (Salam et al., 2005; Bekkar et al., 2020). For example, maternal exposure to ambient $PM_{2.5}$ can lead to negative birth outcomes such as lower birth weight and preterm birth (Zeka et al., 2008; Rosa et al., 2019b). Another example is the impact of $PM_{2.5}$ on an increased risk of developing respiratory conditions such as asthma or on introducing threats to the neurodevelopment of children (Hsu et al., 2015; Chiu et al., 2016). Moreover, a long duration of exposure to heat waves could increase the risk of preterm and early-term birth (Darrow et al., 2024).

Starting at conception, a fetus goes through a carefully orchestrated sequence of stages of biological development. As a result, the exposure effect of environmental chemicals on fetal devel-

opment and programming is sensitive to exposure timing, with equal levels of exposure resulting in different outcomes depending on the developmental stage at which exposure occurred. A period when a fetus is particularly vulnerable is often referred to as a window of susceptibility (or a critical window). Two major inferential goals when studying the effect of maternal exposure on birth and children's health outcomes are to estimate the magnitude of exposure effects of air pollution on fetal health and to identify windows of susceptibility to pollutants across gestational weeks.

Traditionally, the impact of air pollution on fetal growth was evaluated by examining an association between a birth outcome and average exposure measurements during gestation. Another approach was to regress a birth outcome on each trimester's average exposure measurements. For example, Dadvand et al. (2013) examined the negative effects of $PM_{2.5}$ and PM_{10} exposure during trimesters on birth weight and Mortimer et al. (2008) found that prenatal exposure to PM_{10} in the first trimester and NO_2 in the second trimester could lead to impaired pulmonary function of children. Although both approaches are easy to model and implement, assuming constant effects across all gestational weeks or within each trimester can result in biased estimates of exposure effects. Furthermore, clinically defined trimesters may not align with developmental stages and the underlying windows of susceptibility (Wilson et al., 2017b).

1.2 Distributed lag models

Researchers commonly employ a distributed lag model (DLM) as a data-driven approach to identify windows of susceptibility. The DLM regresses a scalar outcome on repeated exposure measurements during a pre-specified time span. Consider a vector of continuous outcomes denoted $\boldsymbol{y} = (y_1, \ldots, y_n)$ for observation $i = 1, \ldots, n$. With a total of T time lags, we denote the exposure measurements observed at equally spaced time points for observation i as $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iT})$. In our context, T represents the total number of gestational weeks and $\boldsymbol{x}_i$ is a vector of exposure measurements of either $PM_{2.5}$ or temperature. We also observe a set of covariates, denoted $\boldsymbol{z}_i$,

which includes the intercept. The DLM for a single exposure is

$$y_i = f(\boldsymbol{x}_i) + \boldsymbol{z}_i'\boldsymbol{\gamma} + \varepsilon_i, \quad f(\boldsymbol{x}_i) = \sum_{t=1}^{T} x_{it}\theta_t, \tag{1.1}$$

where f is the exposure-time-response function parameterized with linear lag effect θ_t at time t, $\boldsymbol{\gamma}$ is a vector of regression coefficients for the fixed effect, and ε_i is an independent error following Normal$(0, \sigma^2)$. The DLM in (1.1) suffers from multicollinearity arising from autocorrelation in the repeated exposure measurements. Temporal structure is often applied to $\theta_1, \ldots, \theta_T$ as a remedy to regularize and smooth the exposure-time-response function. The smoothing constraints for regularization include polynomial, splines, Gaussian process, and regression tree structure (Zanobetti et al., 2000; Gasparrini et al., 2010; Warren et al., 2020; Mork and Wilson, 2023). Alternatively, this can be formulated as a functional regression model with a smooth functional predictor (Wilson et al., 2017a; Gao and Kowal, 2024). The DLM framework is advantageous as it provides a temporal structure to the association between a single exposure and an outcome, allowing for the data-driven identification of the windows of susceptibility.

The DLM has been extended to a distributed lag mixture model (DLMM) to incorporate mixture exposures (i.e. simultaneous exposure to multiple components of air pollution) and lagged interactions among components (Muggeo, 2007; Chen et al., 2019; Mork and Wilson, 2023). We assume that we are interested in the time-varying association between $M \geq 2$ components and a scalar outcome. We let $\boldsymbol{x}_{im} = (x_{im1}, \ldots, x_{imT})$ denote the exposure measurements of component m for time lag $t = 1, \ldots, T$ for observation i, where $m = 1, \ldots, M$. The function f in (1.1) can be rewritten to incorporate mixture exposures as

$$f(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iM}) = \sum_{m=1}^{M}\sum_{t=1}^{T} x_{imt}\theta_{mt} + \sum_{m_1=1}^{M}\sum_{m_2=m_1}^{M}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T} x_{im_1t_1}x_{im_2t_2}\theta_{m_1m_2t_1t_2}, \tag{1.2}$$

where θ_{mt} is the main effect of exposure m at time t and $\theta_{m_1m_2t_1t_2}$ is a pairwise interaction effect between exposure m_1 at time t_1 and exposure m_2 at time t_2. Estimating the mixture-exposure-time-response function in (1.2) is challenging because of the correlation between components of

the mixture exposures at the same time, autocorrelation within each exposure over time, and high-dimensional parameter space of $MT + \binom{M+1}{2}T^2$ parameters. Recent statistical advancements in DLMMs include additive models without interaction (Bello et al., 2017), models with interaction (Warren et al., 2022; Wang et al., 2023), and models allowing for a large number of exposures with pairwise lagged interactions (Mork and Wilson, 2023; Antonelli et al., 2024). The methods and software developed in this dissertation build on the treed distributed lag model (TDLM) framework, introduced by Mork and Wilson (2023). Here, we briefly recapitulate previous work on the TDLM framework.

1.3 Regression tree structure for DLMs

The TDLM framework introduces a Bayesian regression tree structure, referred to as a DLM tree, as a smoothing constraint for a DLM. A DLM tree is shown in Figure 1.1. A DLM tree

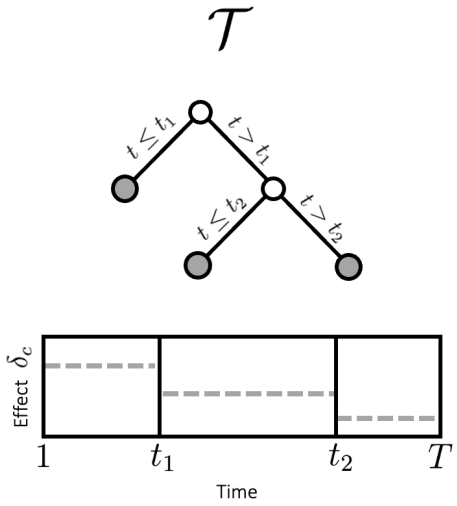


Figure 1.1: A diagram of a DLM tree. The tree splits the exposure time span into multiple time intervals, here resulting in three terminal nodes (gray nodes). The terminal nodes are each assigned a scalar effect (gray dashed lines).

denoted $\mathcal{T}$ is a binary tree with internal nodes that are assigned a time lag as a split point and splits the exposure time span into multiple non-overlapping segments. The terminal nodes of the DLM tree, denoted λ_c for $c = 1, \ldots, C$, represent the time intervals of the exposure. Each terminal node

is assigned a scalar parameter, denoted δ_c, that represents a constant exposure effect for that time interval. The DLM tree defines a function of time,

$$g(t|\mathcal{T}, \mathcal{D}) = \delta_c, \quad t \in \lambda_c, \tag{1.3}$$

where $\mathcal{D} = \{\delta_1, \ldots, \delta_C\}$ is a set of scalar parameters linked to the terminal nodes. The TDLM comprises A DLM trees, each indexed as $\mathcal{T}_a$ with a corresponding set of scalar effect, $\mathcal{D}_a$ for $a = 1, \ldots, A$. The TDLM uses (1.3) to formulate the lag-specific effect parameters in (1.1) such that

$$\theta_t = \sum_{a=1}^{A} g(t|\mathcal{T}_a, \mathcal{D}_a). \tag{1.4}$$

The tree structured DLM framework is similar to the Bayesian additive regression tree (BART; Chipman et al., 2010) framework in that both are additive models with a tree ensemble representation but differ as a DLM tree is modified to split the time span of exposure, rather than the predictor space. Using an ensemble of DLM trees has several advantages. First, the piecewise constant representation of the DLM tree allows for an abrupt change in the exposure effect, outperforming other methods such as spline-based DLMs in identifying windows of susceptibility. Second, the constant effect parameters allow for effect shrinkage, which can effectively shrink the poor-fitting DLM trees for accurate effect estimation. Additionally, the ensemble representation allows for smoothing in the exposure effect as each DLM tree in the ensemble will split the time span differently. Lastly, the split points and the constant effects are all learned from data, enabling a data-driven approach to estimating the exposure-time-response relationship.

In addition to the features offered by the TDLM framework, Mork and Wilson (2023) introduced a treed distributed lag mixture model (TDLMM) with a DLM tree pair to incorporate mixture exposures. The DLM tree pair allows for the estimation of pairwise lagged interactions between exposures through its unique interaction surface structure. The DLM tree pair also accommodates multiple exposures with exposure selection for identifying those most correlated with an outcome. We discuss this framework in further detail in Chapter 2.

1.4 Outline

In this dissertation, we address specific gaps in the DLM framework by developing DLMMs for count outcome data, heterogeneous DLMMs, and user-friendly software and tutorials. First, health outcomes in environmental health applications, such as daily mortality, are often count values with potential zero-inflation and overdispersion. However, there is a lack of DLM framework for estimating exposure effects of mixture exposures for count data. In Chapter 2, motivated by a time series design of live birth identified conceptions (LBIC), we propose a method that accommodates an overdispersed count outcome containing a zero mass within the TDLMM framework, using a Bayesian Pólya-Gamma (PG) data augmentation. The proposed method can estimate the mixture-exposure-time-response relationship between mixture exposures and a count outcome while accounting for the pairwise lagged interactions and performing exposure selection. In simulation studies, we show that our method outperforms the spline-based method for a single exposure in terms of estimating exposure effects and identifying windows of susceptibility. We also demonstrate that the proposed method can perform exposure selection and incorporate lagged interaction in a mixture setting. We apply our method to Colorado birth administrative data to examine the relationship between $PM_{2.5}$ and temperature and pregnancy loss.

Second, the exposure effect of pollutants on each individual may be different due to multiple factors. For example, studies have found that the exposure effect of air pollution on a fetus may be sex-specific or vary across different demographics (Lee et al., 2018; Chiu et al., 2022). Traditionally, the heterogeneity in exposure effects has been addressed by stratifying the dataset with a modifier chosen prior to analysis and applying an appropriate statistical model on each data stratum. Some statistical advancements have allowed the DLM to incorporate modification or heterogeneity in the time-varying exposure effects. Namely for a single exposure, Wilson et al. (2017a) introduced Bayesian distributed lag interaction models (BDLIM) that allow for heterogeneity in both exposure effect and windows of susceptibility. Similarly, Demateis et al. (2024) introduced a penalized distributed lag interaction model (DLIM) for estimating the personalized exposure effect with a single continuous modifier. However, these methods are restricted to a sin-

gle exposure and the sole modifier must be chosen prior to the analysis, which may lead to missing out on true modifiers and potential interactions between the modifiers. In Chapter 3, motivated by precision environmental health objectives, we propose a heterogeneous distributed lag mixture model (HDLMM) for estimating heterogeneous exposure-time-response relationships in a mixture exposure setting. We introduce a tree triplet structure that can incorporate modifier selection, exposure selection, and pairwise interaction between exposures. Via simulation studies, we show that the proposed method can determine modifiers that contribute to heterogeneity and exposures that are most associated with an outcome. We also demonstrate that this leads to an accurate estimation of exposure effects and precise identification of windows of susceptibility at personalized and subgroup levels. We apply our method to the Colorado birth registry data to estimate the heterogeneous association between $PM_{2.5}$ and temperature and birth weight.

In order for novel statistical methodologies to be applied in practice, it is essential to have user-friendly software with extensive documentation and examples. In Chapter 4, we introduce a user-friendly R package **dlmtree** that incorporates tree structured DLMs into one comprehensive software package. We present an overview of extensions of tree structured DLM framework with different model assumptions. With a simulated dataset, we demonstrate the model fitting process for each tree structured DLM in detail. We present useful functions for summarizing the fitted models and visualizing the estimated exposure effects. We further highlight the flexibility of the package by illustrating the extraction of the information within the context of a DLM framework, such as marginalized exposure effects and cumulative effects.

Lastly, in Chapter 5, we provide a summary of our proposed methods and findings. We suggest potential extensions of the proposed models and future research directions. Additionally, we discuss the anticipated impact of this work in environmental health science and related fields.

Chapter 2

Treed Distributed Lag Mixture Model With Zero-Inflated Count Data to Investigate the Association Between Air Pollution and Pregnancy Loss

2.1 Introduction

Parents who experience a miscarriage suffer serious psychological effects, with women reporting long-term consequences of anxiety, major depressive disorder, loss of security, and post-traumatic stress disorder (Schwerdtfeger and Shreffler, 2009). Early pregnancy loss, defined as a loss in the first trimester, is estimated to occur in about one in three pregnancies (Walter, 2023). It is challenging to quantify the rate of pregnancy losses because many pregnancies are lost before individuals are aware that they are pregnant. Because pregnancy loss is hard to observe, its causes including environmental causes, are difficult to study. Most evidence comes from either small cohorts that can be carefully monitored (Wilcox et al., 1988) or cohorts that enroll expectant parents in the first trimester and are limited to observing the subset of losses that occur later in pregnancy. Relying on small, closely monitored cohorts is expensive and results in insufficient power to achieve some inferential goals such as estimating gestational week-specific associations, identifying critical windows of sensitivity when the association between component and health outcome is greatest, or estimating the mixture effect of multiple simultaneous exposures.

To circumvent the limitation of needing to identify all losses and the corresponding reliance on small closely monitored cohorts, Kioumourtzoglou et al. (2019) proposed a novel time-series design using live birth-identified conceptions (LBIC). The LBIC time series design regresses the count of identified conceptions in a given week that result in a live birth, referred to as LBIC, on

exposures during the weeks following conception. This study design relies on two facts. First, the total number of conceptions for a given week is the sum of LBICs and those that end in loss. Second, exposure post-conception cannot affect the number of conceptions. Based on these two factors, the number of pregnancy losses can be identified from the time series study by regressing LBIC on post-conception exposure. This allows us to use a large administrative dataset with readily available data such as birth date and gestational age, and answer more complex epidemiological questions about air pollution and the pregnancy loss relationship, including estimating mixture-exposure-time-response relationships and identifying windows of susceptibility. In this chapter, we use the LBIC time series design for pregnancy loss to estimate the association between pregnancy loss and weekly average exposure to fine particulate matter air pollution smaller than 2.5 micrometers in diameter ($PM_{2.5}$) and temperature in a large administrative birth cohort in the state of Colorado.

A distributed lag model (DLM) is widely used to estimate the relationship between health outcomes and longitudinally assessed environmental exposure, considering each of the days or weeks preceding the health endpoint (Schwartz, 2000; Zanobetti et al., 2000; Gasparrini et al., 2010). In contrast to regressing on a single measurement of exposure, such as pregnancy average exposure, or average exposure over a small number of pre-specified time windows such as trimesters, the DLM has shown lower bias in effect estimation (Wilson et al., 2017b). The DLM also allows for identification of windows of susceptibility, a specific period during development when the developing fetus is particularly sensitive to an environmental exposure (Wright, 2017).

An environmental mixture is a group of environmental chemicals, weather, or other components to which individuals are simultaneously exposed. Recent statistical innovations have allowed for estimating a mixture-exposure-time-response function and identifying windows of susceptibility with mixture exposures. These methods include additive models that consider the additive effect of mixture components assessed longitudinally but do not allow for interactions among components (Bello et al., 2017; Warren et al., 2022; Wang et al., 2023), models that include linear main effects and interactions (Mork and Wilson, 2023; Antonelli et al., 2024), and models for

high-dimensional nonlinear exposure-response function estimation (Wilson et al., 2022). None of these methods are applicable to count data while accounting for lagged interaction among mixture exposures. Furthermore, preliminary investigation shows a large number of zeros in the weekly counts of LBIC in Colorado suggesting a need for a model that can handle both zero-inflation and overdispersion simultaneously (see Figure 2.1 and Appendix Figure A.1). For count data, Chen et al. (2019) developed Poisson distributed lag interaction models, but this can only handle two mixture components and does not apply to zero-inflated and overdispersed count data. To our knowledge, there are no models currently available for longitudinally assessed mixture exposures that accommodate zero-inflated and overdispersed count outcomes.

We present a novel nonparametric Bayesian approach for estimating the mixture-exposure-time-response function on count data that may be zero-inflated and overdispersed. Of particular relevance to our approach is the treed distributed lag mixture model (TDLMM) framework (Mork and Wilson, 2023). TDLMM introduces an ensemble of tree pairs to add structure to main effects and pairwise interactions between lagged components and an outcome, characterizing the linear exposure-time-response function. TDLMM is particularly well suited for the time series design used with LBIC because the tree formulation of the distributed lag function performs well in time-series studies of extended lag periods with low bias amplification compared to spline-based alternatives (Leung et al., 2023). With TDLMM, our model assumes the distributional combination of zero mass and negative binomial (NB) distribution of the count outcome, making it robust to zero-inflation and overdispersion for a more accurate estimation of the exposure-time-response function. Furthermore in the mixture exposure setting, our model can perform component selection and identify windows of susceptibility while effectively mitigating issues of autocorrelation among repeated measures of exposure, correlation between mixture components, and high-dimensional parameter space.

We apply the model to a large administrative birth registry in Colorado to estimate the association between weekly exposure to $PM_{2.5}$ and temperature using the LBIC study design. Our study is novel in several ways and only possible with the new statistical methods developed herein. First,

it is the largest study to use the LBIC design to study pregnancy loss. Second, it is the first study that seeks to identify windows of sensitivity to a mixture exposure for pregnancy loss.

2.2 Colorado birth registry data and LBIC study design

We obtained birth records from the Colorado Department of Public Health and Environment. The dataset includes all registered births from 2007 to 2018 in all 64 counties of Colorado. The dataset also contains the date of birth, estimated gestational age at birth, and the maternal residence categorized by county.

We created weekly counts of the number of conceptions that resulted in a live birth at the county level as described in Kioumourtzoglou et al. (2019). For each live birth, we constructed the estimated date of conception using the date of birth and estimated gestational age. We then aggregated the data to the week of conception and county of maternal residence. Conditioning on live birth in pregnancy cohorts can cause fixed cohort bias as shorter pregnancies at the beginning and longer pregnancies at the end may be missing (Strand et al., 2011; Leung et al., 2021). To avoid fixed cohort bias, we trimmed the dataset to conceptions between 2006-11-11 and 2017-08-11. As a result, we have a time series study design where each county has 561 weeks resulting in a total number of county-weeks of 35,904 with 7,638 weeks (21.3%) with no estimated LBICs. The total number of LBICs was 713,078. Figure 2.1, Appendix Figure A.1, A.2, and A.3 illustrate the data.

We acquired the maximal daily temperature data from Abatzoglou (2013) and census-tract level exposure data for $PM_{2.5}$ from the United States Environmental Protection Agency (https://www.epa.gov/hesc/rsig-related-downloadable-data-files). To calculate daily county-level exposures, we used a population-weighted average for both components. Specifically, with the temperature data on a 4-kilometer grid, we created census tract average exposure by averaging the grid cells with centroids in each census tract. We then created county-level average exposures by averaging the daily census tract-level exposure weighted by 2010 census population data for each census tract. For $PM_{2.5}$, we used the same procedure starting from the census tract-level estimate. We then constructed weekly averages for each county corresponding to the LBIC weekly counts. We regress

LBIC on the weekly average exposures in the 40 weeks following the week of conception. This study has been approved by the Colorado State University Institutional Review Board.

2.3 Methods

2.3.1 Zero-inflated negative binomial framework

Let y_{ij} be count outcome for location i and study week $j = 1, \ldots, n_i$. In our analysis, y_{ij} is the LBIC count, where i represents the county of maternal residence and j represents the calendar week of conceptions. Our proposed methodology is designed to handle lagged exposures. Typically lagged exposures are exposure measurements observed in the T time points preceding the outcome in location i and study week j. For the LBIC study design, we study forward lags or leads, however, our proposed approach is suitable for both exposure lags and leads. For consistency in terminology across the DLM literature, we refer to these as time lags (or lags) throughout. We consider M components of a mixture. For mixture components $m = 1, \ldots, M$, we denote $\boldsymbol{x}_{ijm} = (x_{ijm1}, \ldots, x_{ijmT})$ as a vector of observed exposure during time lag $t = 1, \ldots, T$ following the outcome. In our analysis, $\boldsymbol{x}_{ij1}$ and $\boldsymbol{x}_{ij2}$ describe $PM_{2.5}$ and temperature, observed during $T = 40$ gestational weeks post conception at county i in calendar week j, respectively.

We assume our count outcome is potentially zero-inflated (ZI) and overdispersed. We model y_{ij} as

$$y_{ij} \sim \pi_{ij}\delta_0 + (1 - \pi_{ij})\text{NB}(\mu_{ij}, r), \tag{2.1}$$

where π_{ij} is the probability that an observation belongs to a zero mass δ_0 represented with a Dirac delta function, and μ_{ij} and r represent the mean and dispersion parameter of the NB distribution, respectively. Our proposed model largely follows the approach of Neelon (2019) for zero-inflated negative binomial (ZINB) regression. In the ZINB model in (2.1), the count outcome comprises two pieces: a zero mass and NB distribution. We can assume that non-zero values originate from the NB distribution whereas zero-valued outcomes can come from either component of the mixture distribution. We refer to zeros from the zero mass as ZI zeros and NB zeros if otherwise. We

introduce an auxiliary ZI zero indicator variable w_{ij}. If $\mathrm{w}_{ij} = 1$, y_{ij} belongs to ZI zeros. If $\mathrm{w}_{ij} = 0$, y_{ij} belongs to the NB distribution. We model w_{ij} with the logistic regression,

$$\begin{aligned} \mathrm{w}_{ij} &\sim \text{Bernoulli}(\pi_{ij}) \\ \text{Logit}(\pi_{ij}) &= \boldsymbol{z}'_{1ij}\boldsymbol{\gamma}_1, \end{aligned} \tag{2.2}$$

where $\boldsymbol{z}_{1ij}$ is a vector of covariates and $\boldsymbol{\gamma}_1$ denotes a vector of regression coefficients. We refer to this model as the ZI regression model. We parameterize the NB distribution in (2.1) as

$$\begin{aligned} \mathbb{P}(y_{ij}|\mathrm{w}_{ij} = 0, r, \psi_{ij}) &= \frac{\Gamma(y_{ij} + r)}{\Gamma(r)y_{ij}!}(1 - \psi_{ij})^r \psi_{ij}^{y_{ij}} \\ \text{Logit}(\psi_{ij}) &= \boldsymbol{z}'_{2ij}\boldsymbol{\gamma}_2 + f(\boldsymbol{x}_{ij1}, \ldots, \boldsymbol{x}_{ijM}), \end{aligned} \tag{2.3}$$

where ψ_{ij} is a success probability parameter, $\boldsymbol{z}_{2ij}$ is a vector of covariates, $\boldsymbol{\gamma}_2$ is a vector of regression coefficients, and $f(\boldsymbol{x}_{ij1}, \ldots, \boldsymbol{x}_{ijM})$ is a mixture-exposure-time-response function of all M components throughout time lag $t = 1, \ldots, T$. The parameterization in (2.3) leads to the mean of NB distribution in (2.1) as $\mu_{ij} = r\left(\frac{\psi_{ij}}{1-\psi_{ij}}\right)$. We refer to this model as the NB regression model.

We consider various forms of $f(\boldsymbol{x}_{ij1}, \ldots, \boldsymbol{x}_{ijM})$. For a single component model, i.e., $M = 1$, we omit the subscript m and write $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijT})$ for the simplicity of presentation. When $M = 1$, the exposure-time-response function is

$$f(\boldsymbol{x}_{ij}) = \sum_{t=1}^{T} x_{ijt}\theta_t, \tag{2.4}$$

where x_{ijt} is observed exposure for ijth observation at time lag t and θ_t is a linear exposure effect at time lag t. For a mixture of M components, we reparameterize $f(\boldsymbol{x}_{ij1}, \ldots, \boldsymbol{x}_{ijM})$ to a mixture-exposure-time-response function with additional parameters to account for the main exposure effect of each component and the pairwise interaction among each distinct pair of components.

For mixture exposures where $M > 1$, the mixture-exposure-time-response function is

$$f(\boldsymbol{x}_{ij1}, \ldots, \boldsymbol{x}_{ijM}) = \sum_{m=1}^{M} \sum_{t=1}^{T} x_{ijmt}\theta_{mt} + \sum_{m_1=1}^{M-1} \sum_{m_2>m_1}^{M} \sum_{t_1=1}^{T} \sum_{t_2=1}^{T} x_{ijm_1t_1} x_{ijm_2t_2} \theta_{m_1m_2t_1t_2}, \tag{2.5}$$

where θ_{mt} is the main exposure effect of component m at time lag t and $\theta_{m_1m_2t_1t_2}$ is the interaction effect between component m_1 and m_2 at time lags t_1 and t_2. The mixture model in (2.5) incorporates time-sensitive interactions implying exposure to a particular component at one time point can influence susceptibility to a different component at a different time point.

There are three primary challenges when estimating the mixture-exposure-time-response function. First, there is high autocorrelation between repeated measurements of components. Second, there is a need to account for the correlation between mixture components at each time point. Lastly, the mixture model in (2.5) is high dimensional for even a moderate number of mixture components. With M components measured across T time points, estimating the mixture-exposure-time-response function in (2.5) requires $MT + \binom{M}{2}T^2$ parameters. Hence a parameterization and estimation approach that regularizes the temporally ordered parameters and, when M is large, performs component selection is needed. We use the tree-based approach proposed by Mork and Wilson (2023).

2.3.2 Treed distributed lag model

In this section, we recapitulate treed distributed lag model (TDLM) and TDLMM and parameterize $f(\boldsymbol{x}_{ij1}, \ldots, \boldsymbol{x}_{ijM})$ in the ZINB framework with the regression tree formulation (Mork and Wilson, 2023). The single component model TDLM introduces an additive model with an ensemble of A trees to estimate θ_t for time lag $t = 1, \ldots, T$ in (2.4). Each binary tree $\mathcal{T}_a$ in the ensemble partitions the total time span T to B_a non-overlapping intervals each corresponding to a terminal node of the tree for $a = 1, \ldots, A$. As a result, tree $\mathcal{T}_a$ has B_a terminal nodes denoted η_{ab} with a corresponding terminal node parameter λ_{ab} for $b = 1, \ldots, B_a$. On its own, each tree defines a function of time lag $g(t)$ as

$$g(t|\mathcal{T}_a, \Lambda_a) = \lambda_{ab} \quad \text{if } t \in \eta_{ab}, \tag{2.6}$$

where $\Lambda_a = \{\lambda_{a1}, \ldots, \lambda_{aB_a}\}$ is a set of terminal node parameters of tree $\mathcal{T}_a$. TDLM employs an additive ensemble of A regression trees and formally defines the exposure effect at time lag t as a sum of the binary trees,

$$\theta_t = \sum_{a=1}^{A} g(t|\mathcal{T}_a, \Lambda_a). \tag{2.7}$$

For the mixture model in (2.5), we follow the structured tree pair approach of Mork and Wilson (2023). The approach uses an ensemble of tree pairs denoted $\mathcal{T}_a = \{\mathcal{T}_{a1}, \mathcal{T}_{a2}\}$ for $a = 1, \ldots, A$. Each tree pair jointly defines the main effects and pairwise interaction in the time lags for two components. Tree $\mathcal{T}_{a1}$ and $\mathcal{T}_{a2}$ are each associated with mixture components. We denote $S_{ap} = m$ if component m is assigned to tree $\mathcal{T}_{ap}$ for $p = 1, 2$. Similar to (4.3), tree pair $\mathcal{T}_a$ defines main effects through a function

$$g(t|\mathcal{T}_{ap}, \Lambda_{ap}) = \lambda_{apb} \quad \text{if } t \in \eta_{apb}, \quad p = 1, 2, \tag{2.8}$$

where η_{apb} denote the terminal nodes of tree $\mathcal{T}_{ap}$ and $\Lambda_{ap} = \{\lambda_{ap1}, \ldots, \lambda_{apB_{ap}}\}$ is a set of corresponding terminal node parameters of tree $\mathcal{T}_{ap}$ for $b_p = 1, \ldots, B_{ap}$. The tree pair also jointly defines the interaction effect for two components through a function

$$g_I(t_1, t_2|\mathcal{T}_{a1} \times \mathcal{T}_{a2}, \mathcal{I}_a) = \rho_{ab_1b_2} \quad \text{if } t_1 \in \eta_{ab_1}, t_2 \in \eta_{ab_2}, \tag{2.9}$$

where $\mathcal{T}_{a1} \times \mathcal{T}_{a2}$ denotes a pairwise interaction surface and $\mathcal{I}_a$ is the set of corresponding interaction effects for a tree pair $\mathcal{T}_a$. Using the full ensemble of A tree pairs, the main effect of the mixture-exposure-time-response function for component m at time lag t is

$$\theta_{mt} = \sum_{a=1}^{A} \sum_{p=1}^{2} g(t|\mathcal{T}_{ap}, \Lambda_{ap}) \mathbb{I}(S_{ap} = m), \tag{2.10}$$

where $\mathbb{I}(\cdot)$ is an indicator function. Similarly, the interaction effect for components m_1 and m_2 at time lags t_1 and t_2 is

$$\theta_{m_1 m_2 t_1 t_2} = \sum_{a=1}^{A} g_I(t_1, t_2 | \mathcal{T}_{a1} \times \mathcal{T}_{a2}, \mathcal{I}_a) \mathbb{I}(S_{a1} = m_1, S_{a2} = m_2). \tag{2.11}$$

For both the single component and mixture models, the tree structures and terminal node parameters are learned from the data. In the mixture model, the components associated with each tree pair are also learned from the data allowing for model-based selection of the mixture components and interactions included in the model.

There are two alternative forms of the mixture model TDLMM. First is a purely additive model which assumes no interaction and reduces (2.5) to a purely additive model, i.e., $\theta_{m_1 m_2 t_1 t_2} = 0$ for $(m_1, m_2) \in \{1, \ldots, M-1\} \times \{m_1, \ldots, M\}$ for all t_1 and t_2 (TDLMMadd). The second alternative form allows for self-interactions by changing the bounds of the summations for the interaction terms in (2.5) to allow $m_1 = m_2$ (TDLMMall). This form of the model allows for quadratic main effects and for exposure to one component to modify susceptibility to that same component at a later time.

2.3.3 Prior specification

The prior specification of the TDLM framework for a single component contains two parts: A prior on the structure of $\mathcal{T}_a$ and a prior on the terminal node parameters λ_{ab}. First, the tree structure of $\mathcal{T}_a$ is defined with a node splitting probability and splitting rule assignment. We follow the approach of Chipman et al. (2010) to define the priors. Specifically, the splitting probability of a tree node η is $\mathbb{P}_{split}(\eta) = \alpha(1+d)^{-\beta}$ where d is a depth of the node and $\alpha \in (0,1), \beta \in [0, \infty)$ control the shape and the number of terminal nodes. The splitting probability quickly decays to zero as a tree grows bigger controlling for the tree size. We fix the parameters as $\alpha = 0.95$ and $\beta = 2$. When a node is split, a new splitting rule is assigned by selecting a time lag randomly and uniformly, conditional on the previous splits of the parental nodes. Secondly, we specify a prior on the terminal node parameters as $\lambda_{ab} \sim \text{Normal}(0, \tau_a^2 \nu^2)$ where both hyperparameters are

set as half-Cauchy$(0, 1)$. This prior specification for the terminal node parameters benefits from the conditional conjugacy and allows for global-local shrinkage where ν imposes global shrinkage overall and τ_a imposes local shrinkage at the tree level due to the hierarchical relationship between the horseshoe prior and half-Cauchy distribution (Carvalho et al., 2010; Makalic and Schmidt, 2016). Controlling the trees to be small imposes temporal smoothing on the exposure effect across the time span, adjusting for the high autocorrelation from the repeated measurements. The prior on the tree structure also prevents a few large trees from overwhelming the tree ensemble limiting the advantage of additive model representation.

For TDLMM with a mixture of M components, the tree pair structure requires a prior specification for the trees, the interaction surface, and component selection. For each tree in a tree pair, we use the same prior for the tree structure as specified in the TDLM framework. For the terminal node parameters of a tree with exposure m, we introduce component-specific shrinkage hyperparameter υ_m in addition to the global-local shrinkage. The prior on the terminal node parameters of a tree splitting exposure m in a tree pair is set to $\lambda_{apb}|S_{ap} = m \sim \text{Normal}(0, \tau_a^2\nu^2\upsilon_m^2)$ for $p = 1, 2$. All hyperparameters are set as half-Cauchy$(0, 1)$. For the interaction surface, we do not specify the prior for its structure as it is fully determined by the prior on the tree structure. For the pairwise interaction effects on the interaction surface, we specify $\rho_{ab_1b_2} \sim \text{Normal}(0, \tau_a^2\nu^2\upsilon_{m_1m_2}^2)$ where $\upsilon_{m_1m_2}$ is the local shrinkage hyperparameter for interaction effect between component m_1 and m_2, set as half-Cauchy$(0, 1)$. The mixture model additionally requires a prior on which components are included in each tree pair. For component selection, we assign categorical-Dirichlet prior on S_{ap} such that

$$S_{ap}|\mathcal{E} \sim \text{Categorical}(\mathcal{E}), \quad \mathcal{E} \sim \text{Dirichlet}(\kappa, \ldots, \kappa), \tag{2.12}$$

where $\mathcal{E} = \{E_1, \ldots, E_M\}$ is a vector of probabilities that a tree is assigned component m for $m = 1, \ldots, M$ and κ is a sparsity hyperparameter of the components (Linero, 2018). With this formulation, component m is selected out of the model if it is not included in any tree pair. This results in hierarchical variable selection such that the interaction between two components is only included if both main effects are included in the model as well. We fix $\kappa = 0.639$ based on the

ensemble of 20 trees with two components using the Bayes factor method suggested by Mork and Wilson (2023).

2.3.4 Other parameters and computation

We assign Normal$(0, 100)$ prior to the coefficient of both ZI regression $\boldsymbol{\gamma}_1$ and the NB regression $\boldsymbol{\gamma}_2$. Additionally, we specify Discrete-Uniform$(0, 10)$ prior for the dispersion parameter r. The posterior distribution of r did not approach the upper boundary of its prior range for simulation and data analysis. We implement the algorithm with a Markov Chain Monte Carlo (MCMC) and Bayesian backfitting within the Gibbs sampler technique (Hastie and Tibshirani, 2000b). We consider four transition steps to update trees in the Metropolis-Hasting algorithm: grow, prune, change, and switch-component, and apply these rules with equal probability. Further details on the transition steps are described in Mork and Wilson (2023). The algorithm roughly combines the approach of Bayesian ZINB regression with the Pólya-Gamma variable proposed by Neelon (2019) and TDLMM algorithm of Mork and Wilson (2023). We provide full details of the algorithm with full conditional posterior distributions in Appendix A.

2.3.5 Marginalized effect

In the mixture exposure setting, the exposure effects of one component are dependent on the level of other co-exposures due to the interaction effects across lags. To obtain the marginalized exposure effect of component m, we marginalize across the other components at fixed levels. If we fix the co-exposures to be at $\tilde{\boldsymbol{X}}_{(-m)} = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{m-1}, \tilde{\boldsymbol{x}}_{m+1}, \ldots, \tilde{\boldsymbol{x}}_M\}$, the marginalized exposure effect of component m at time lag t is

$$\tilde{\theta}_{mt}(\tilde{\boldsymbol{X}}_{(-m)}) = \theta_{mt} + \sum_{m'=1}^{m-1}\sum_{t'=1}^{T} \tilde{x}_{m't'}\theta_{m'mt't} + \sum_{m'=m+1}^{M}\sum_{t'=1}^{T} \tilde{x}_{m't'}\theta_{mm'tt'}. \tag{2.13}$$

We fix $\tilde{\boldsymbol{X}}_{(-m)}$ to empirical means of co-exposures since it has been shown that evaluating (2.13) with empirical means is equivalent to integrating the co-exposures out (Mork and Wilson, 2023).

2.4 Simulation

The goal of the simulation study is to assess the operating characteristics of the proposed model for a single component and for the mixture components with a count outcome. Specifically, we are interested in the marginal exposure effect estimation and windows of susceptibility identification for each component, and component selection for mixture components. We consider two scenarios that mimic our data analysis of county-level time series design: 1) a single component scenario and 2) a two component scenario that includes a time-sensitive interaction effect. For both scenarios, we use real exposure data from the Colorado birth data. All simulations can be replicated with R package **dlmtree**.

For both scenarios, we fit models with 20 trees and perform the MCMC algorithm for 10,000 iterations thinning every two iterations after discarding 2,500 iterations as burn-in. As a competing method for the first scenario, we consider natural splines with a generalized linear model. We let the degrees of freedom range from 4 to 19 and chose the best-fitting model. We fit a ZINB generalized linear model with natural splines with 10 degrees of freedom using **pscl** package in R (Zeileis et al., 2008; R Core Team, 2021). Further information on the degrees of freedom selection is provided in Appendix A. We did not find a suitable competing method for the second scenario that had available code for multiple lagged components and zero-inflated count outcome.

For model performance comparison, we quantify models' marginalized exposure effect estimation performance with root mean-squared error (RMSE) and 95% confidence/credible interval (CI/CrI) coverage and width. On the model scale, we calculate the RMSE of the marginal effect as

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\tilde{\theta}_t - \hat{\tilde{\theta}}_t)^2} \tag{2.14}$$

where $\tilde{\theta}_t$ and $\hat{\tilde{\theta}}_t$ denote the true and estimated marginal exposure effect at time lag t, respectively. The CI/CrI coverage indicates the proportion of time lags for which the model's CI/CrI successfully captures the true marginal exposure effect. We obtain CI/CrI width by taking the average of CI/CrI widths across all time lags.

We also evaluate each model's ability to identify windows of susceptibility with true positive (TP), false positive (FP), and precision. TP measures the number of successfully identified lags among the lags of the true window of susceptibility. Conversely, FP measures the number of lags falsely identified to be the window of susceptibility. We calculate the precision as TP/(TP + FP).

2.4.1 Scenario 1: Single component

For the first scenario, we simulate a count outcome with a single component. We generated 100 independent datasets where we randomly sample 20 of the 64 counties from Colorado birth data for each dataset. Each county contains 561 weeks resulting in 11,220 county-weeks. The data generating model for scenario 1 is

$$\begin{aligned} y_{ij} &\sim \pi_{ij}\delta_0 + (1-\pi_{ij})\text{NB}(\mu_{ij}, r) \\ \mu_{ij} &= r\left(\frac{\psi_{ij}}{1-\psi_{ij}}\right) \\ \text{Logit}(\pi_{ij}) &= \boldsymbol{z}'_{1ij}\boldsymbol{\gamma}_1 \\ \text{Logit}(\psi_{ij}) &= \boldsymbol{z}'_{2ij}\boldsymbol{\gamma}_2 + f_1(\boldsymbol{x}_{ij1}). \end{aligned} \tag{2.15}$$

We specified a single categorical covariate of the county of maternal residence for $\boldsymbol{z}_{1ij}$ to generate realistic data of county-level time-series design. We sampled $\boldsymbol{\gamma}_1$ from Normal$(0, 1)$ to generate the zero-inflated outcome. On average across the datasets, 50.3% of the observations were ZI zeros (IQR: $47.8 - 53.1\%$), 7.7% were NB zeros (IQR: $6.8 - 8.3\%$), and 42.0% were non-zeros (IQR: $39.6 - 44.3\%$). For $\boldsymbol{z}_{2ij}$, we included the month and year of conception in addition to the county of maternal residence. We sampled $\boldsymbol{\gamma}_2$ from Normal$(0, 0.1^2)$ and fixed the dispersion parameter r to 3. This setting controls the size of count values of the outcome making them more representative of the actual number of births and generates data resembling overdispersed count outcome with a reasonable amount of NB zeros for both scenarios.

We set the exposure data $\boldsymbol{x}_{ij1}$ as the real $PM_{2.5}$ data corresponding to the randomly chosen counties. We set the true exposure-time-response function $f_1(\boldsymbol{x}_{ij1})$ to be non-zero for a randomly selected 8-week interval representing a window of susceptibility, and let the exposure effect for

those weeks be $\theta_t = 0.1$. The true marginal exposure effect $\tilde{\theta}_t$ is 0.1 during the weeks of the window of susceptibility as we only consider a single component for scenario 1.

We fit the natural splines method and TDLM assuming only $PM_{2.5}$ data were available while we include temperature data along with $PM_{2.5}$ for TDLMM methods. All components are centered and scaled prior to data generation and fitting.

Table 2.1: Model performance measures for scenario 1: The first three columns state the measures for the marginalized exposure effect estimation: RMSE, CI/CrI coverage, and CI/CrI width. The next columns show the measures for the window of susceptibility identification: TP, FP, and precision.

	Exposure Effect Estimation			Window Identification		
Model	RMSE$\times 100$	Coverage	Width	TP	FP	Precision
Splines	1.470	0.712	0.022	1.00	0.189	0.845
TDLM	0.406	0.994	0.023	1.00	0.005	0.995
TDLMMadd	0.344	0.992	0.020	1.00	0.006	0.995
TDLMMns	0.358	0.992	0.021	1.00	0.006	0.995
TDLMMall	0.378	0.991	0.022	1.00	0.008	0.993

Table 3.1 presents the simulation results for scenario 1. Both TDLM and TDLMM outperform natural splines. TDLM and TDLMM have significantly lower RMSEs with a high coverage rate above 99% compared to the natural splines with 71.2%. While natural splines can achieve lower RMSE and a higher coverage rate with greater degrees of freedom, we find that this leads to wider CIs as a trade-off. The lowest RMSE for the natural splines was 1.30 with 19 degrees of freedom which is still higher compared to TDLM and TDLMM. The higher RMSE of natural splines can be explained by the method failing to capture an abrupt change in a true marginal exposure effect and by the curvature of the splines around zero in the areas of null effect. These results reflect the advantage of employing the ensemble of trees model. The TDLM and TDLMM's coverage rates for lags of null effect are slightly higher than those of windows of susceptibility because the shrinkage priors give TDLM and TDLMM a very low probability of identifying lags of null effect as a window of susceptibility and shrink the exposure-time-response function to zero over those periods. TDLMM has a slightly lower RMSE than TDLM as a result of including a component-specific shrinkage hyperparameter.

For window of susceptibility identification, TDLM and TDLMM both reach a high precision near 99% compared to only 84.5% of natural splines. While natural splines achieve the same level of identifying the true window of susceptibility with a true positive rate of 1, the curvature of the natural splines in the lags of null effect results in misidentifying the lags as windows of susceptibility, resulting in a higher false positive.

2.4.2 Scenario 2: Mixture components with interaction

For the second scenario, we similarly sampled 20 of the 64 counties from Colorado birth data for each dataset and generated 100 datasets of 11,220 county-weeks. Each dataset includes the corresponding real exposure data of $PM_{2.5}$ and temperature. We generated the covariates, regression coefficients of the ZI and NB regression, and dispersion parameter as described in scenario 1.

We generated count outcome with the main exposure effect of $PM_{2.5}$ in addition to its time-sensitive interaction effect with temperature. The data generating model for NB regression with the updated exposure effect is

$$\begin{aligned} \text{Logit}(\psi_{ij}) &= \boldsymbol{z}'_{2ij}\boldsymbol{\gamma}_2 + f_2(\boldsymbol{x}_{ij1}, \boldsymbol{x}_{ij2}) \\ f_2(\boldsymbol{x}_{ij1}, \boldsymbol{x}_{ij2}) &= 0.1 \sum_{t=s_1}^{s_1+7} x_{ij1t} + 0.025 \sum_{t_1=s_1}^{s_1+7} \sum_{t_2=s_2}^{s_2+7} x_{ij1t_1} x_{ij2t_2}, \end{aligned} \tag{2.16}$$

where $\boldsymbol{x}_{ij2}$ represents temperature data. For the exposure effect $f_2(\boldsymbol{x}_{ij1}, \boldsymbol{x}_{ij2})$, we uniformly sampled independent starting points s_1 and s_2 from $\{1, \ldots, T-7\}$ for $PM_{2.5}$ and temperature. We then fixed the main effect of $PM_{2.5}$ from time point s_1 to $s_1 + 7$ to 0.1 ($\theta_{1t} = 0.1$) and its interaction effect with temperature from time point s_2 to $s_2 + 7$ as 0.025 ($\theta_{12t_1t_2}$ = 0.025). The resulting true marginal main effect $\tilde{\theta}_{1t}$ is 0.1 as all components are centered and scaled prior to data generation and fitting.

Table 3.2 presents results demonstrating that all versions of TDLMM perform well under the mixture components setting. Notably, overall TDLMMns and TDLMMall perform better than TDLMMadd with lower RMSEs and coverage rates above the nominal level. Specifically for esti-

Table 2.2: Model performance measures for scenario 2: The first four columns state RMSE and coverage rate for $PM_{2.5}$ and temperature. The next three columns show TP, FP, and precision. In contrast to scenario 1, we calculate TP, FP, and precision with $PM_{2.5}$ and temperature combined since the temperature is only included in the interaction effect, i.e., the marginalized main effect of temperature has no window of susceptibility resulting in a zero in the denominator of TP and FP.

	Exposure Effect Estimation						
	RMSE$\times$100		Coverage		Window Identification		
Model	$PM_{2.5}$	Temp	$PM_{2.5}$	Temp	TP	FP	Precision
TDLMMadd	0.623	0.182	0.894	0.999	0.764	0.046	0.943
TDLMMns	0.375	0.116	0.990	1.000	0.765	0.004	0.994
TDLMMall	0.382	0.125	0.986	1.000	0.765	0.007	0.990

mating the main effect of $PM_{2.5}$, TDLMMadd had higher RMSE and lower coverage rate compared to TDLMMns and TDLMMall. TDLMMadd without interaction effect assumption estimates the interaction effect as the main effect of $PM_{2.5}$ and temperature resulting in a higher RMSE and a lower coverage rate for both components. TDLMMns performed better than TDLMMall with the lowest RMSE and the highest coverage rate for both components as TDLMMall allows for self-interactions which induce non-linear exposure effects.

With respect to identifying the window of susceptibility, all three versions of TDLMM reach combined precision of at least 94%. While TDLMMadd shares a similar level of TP with TDLMMns and TDLMMall, the misspecified TDLMMadd falsely identifies lags with no true effect as windows of susceptibility. This results in lower precision.

We designed the scenarios to mimic our time-series approach with two components and to demonstrate component selection in TDLMM methods by including the temperature of the null effect. The TDLMM method is scalable for higher dimensional exposure spaces as illustrated for Gaussian data by Mork and Wilson (2023).

2.5 Analysis of LBIC in the Colorado birth data

We fit the proposed model to the Colorado birth data described in Section 2.2. We estimated the main effects and the interaction effect of two components, $PM_{2.5}$ and temperature, over a gestational period of 40 weeks on LBIC. We included the county information of maternal residence as a

single covariate for the ZI model. In the NB model, we additionally included categorical variables for the month and year of conception as recommended by Leung et al. (2023). We fit our model with 20 trees with the prior distributions described in Section 3.3.2. Sensitivity analyses indicate few changes for the different number of trees and options of priors. We imposed no-self interaction for linear interpretability of the result. We ran the model for 50,000 MCMC iterations thinning every two iterations after 10,000 burn-in. Running multiple chains with different initial values did not change the result and the MCMC diagnostics indicated good mixing and convergence.

2.5.1 Zero-inflation per county and dispersion parameter

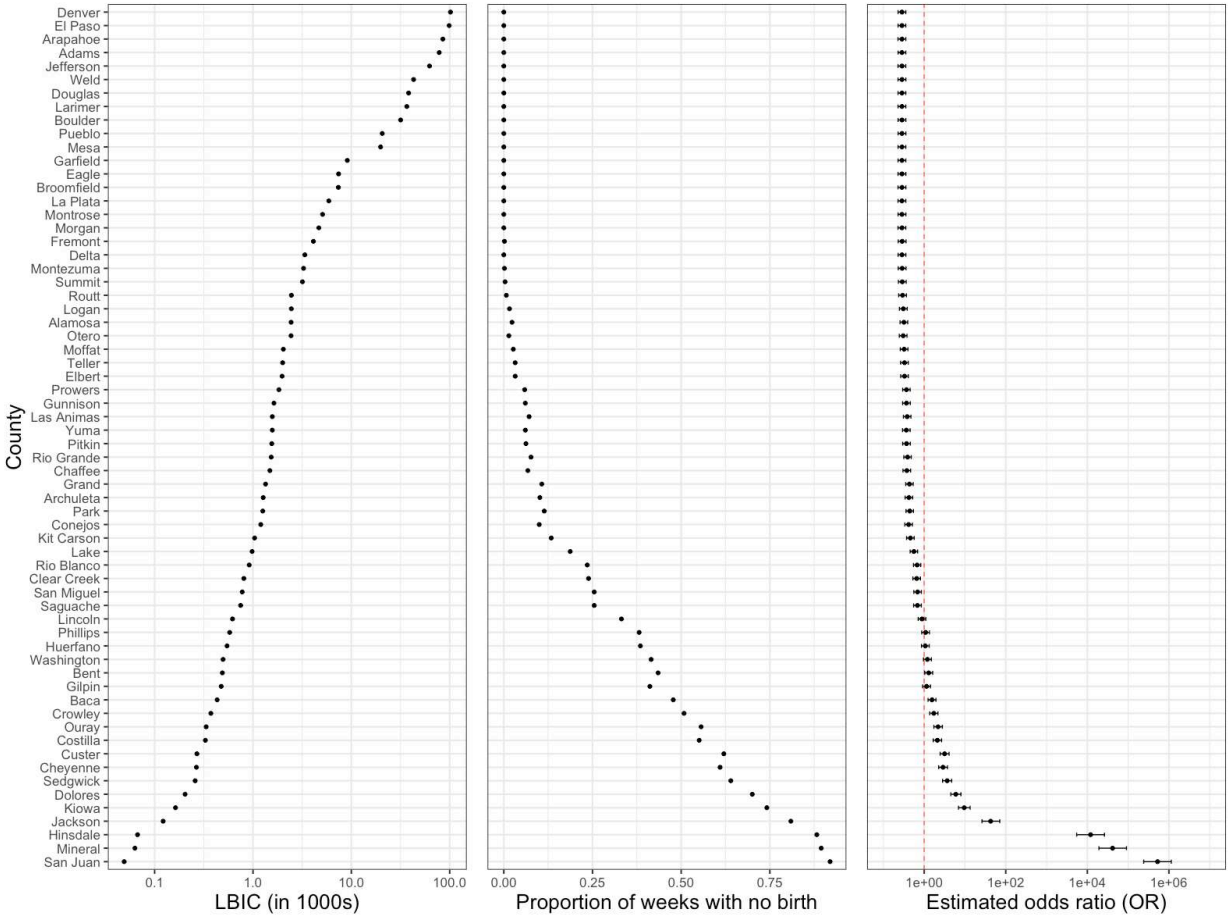


Figure 2.1: Total number of LBIC, the proportion of weeks with no birth, and estimated odds ratio of each county in Colorado. A higher odds ratio indicates a higher chance of zero-inflation. The horizontal bars of the right panel indicate 95% credible intervals of the odds ratios.

Figure 2.1 shows the total number of LBIC by county, the proportion of weeks with no birth by county, and the estimated odds ratio (OR) of zero-inflation by county in Colorado. The model provides strong evidence that some counties are heavily zero-inflated while other counties are not. Counties on the Front Range and around the Grand Junction in the western part of the state all had ORs near zero, indicating the data had little or no zero-inflation. Many of the rural counties had high ORs indicating a large amount of zero-inflation. This is not surprising given the small populations and very low number of births in these counties. The posterior mean of the dispersion parameter was 6.53 (95% CrI: [6, 7]).

2.5.2 Marginalized exposure effect

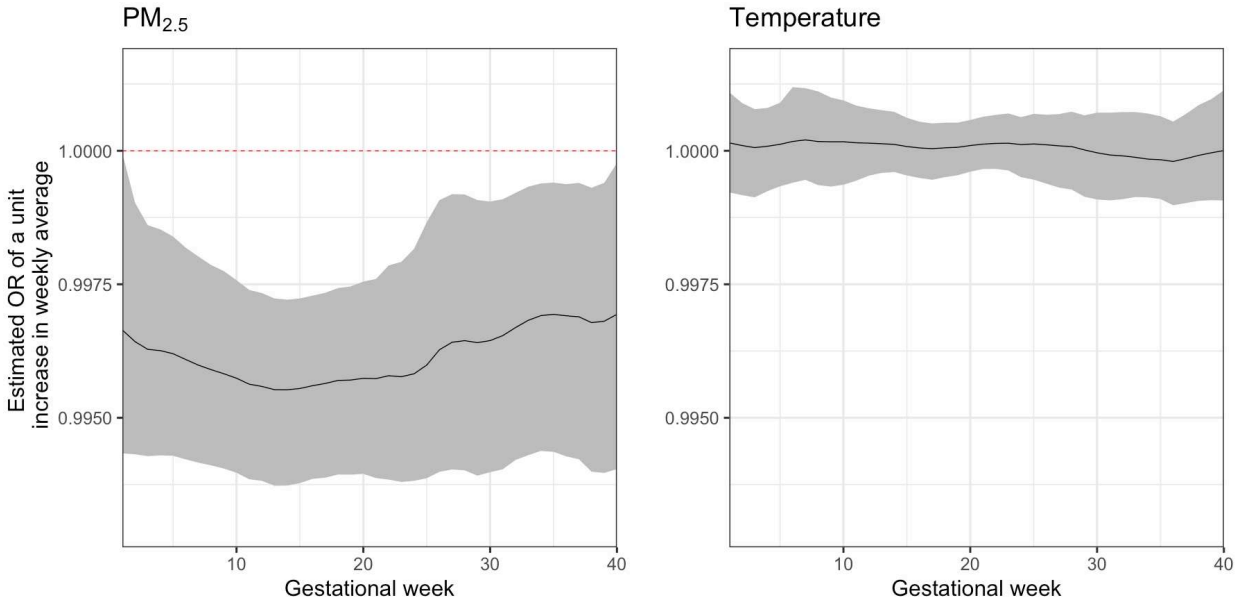


Figure 2.2: Estimated marginal distributed lag function: Posterior distributed lag function for each component while fixing the other component at its empirical mean. The gray area indicates the 95% credible interval of the effect.

The posterior inclusion probabilities of $PM_{2.5}$ and temperature with component selection were both 1.00, which is not surprising considering that we only considered two components. Figure 2.2 shows the estimated distributed lag effect of $PM_{2.5}$ and temperature for 40 gestational weeks

marginalized at the empirical mean of the co-exposure. The left panel of Figure 2.2 shows that a window of susceptibility of $PM_{2.5}$ spans throughout the gestational period. The estimated ORs were comparatively lower in weeks 10 - 24. The lowest OR of $PM_{2.5}$ was 0.996 (95% CrI: [0.994, 0.997]) at week 14. The cumulative OR of $PM_{2.5}$ on the change of LBIC was 0.858 (95% CrI: [0.833, 0.883]). These results suggest that increased exposure to $PM_{2.5}$ is associated with decreased LBIC. This implies that maternal exposure to $PM_{2.5}$ is positively associated with pregnancy loss. The model did not identify any window of susceptibility for temperature and the marginalized posterior means of its OR hover close to 1 indicating no direct exposure effect from temperature at all weeks.

2.5.3 Interaction with co-exposure

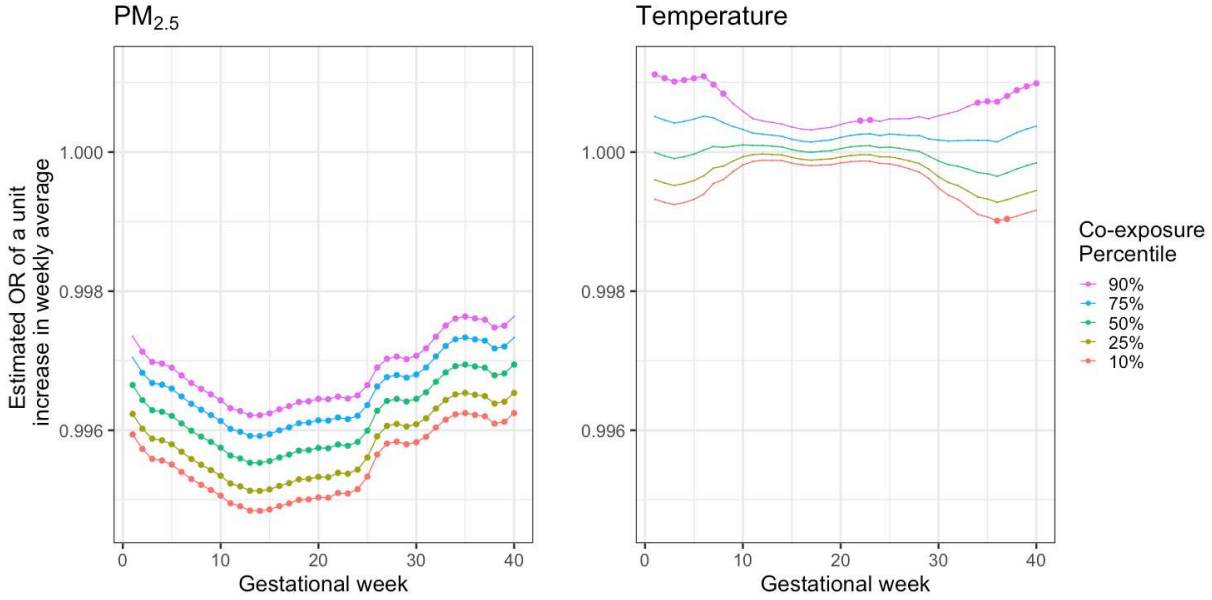


Figure 2.3: Estimated distributed lag effect conditional on the different percentiles of co-exposure. Lines are the distributed lag effects of the component and dots for each lag indicate the zero inclusion of 95% credible intervals.

Figure 2.3 shows the change in the estimated distributed lag effect of $PM_{2.5}$ and temperature at different percentiles of co-exposure. The left panel of Figure 2.3 indicates that the effect of $PM_{2.5}$

is mitigated at higher temperatures. This implies that the weekly time-resolved interaction effect of $PM_{2.5}$ and temperature is mostly positive. However, despite its positive interaction with temperature, the negative effect of $PM_{2.5}$ persists with nearly identical windows of susceptibility. The right panel of Figure 2.3 shows that the effect of temperature increases when the $PM_{2.5}$ increases. The interaction with $PM_{2.5}$ amplifies the temperature effect during earlier and later weeks compared to those in the middle. Notably, the first 8 weeks and the last 7 weeks emerge as windows of susceptibility conditional on the 90th percentile of $PM_{2.5}$.

2.5.4 Expected pregnancy losses

To provide more interpretable estimates, we use Bayesian g-computation to estimate the difference in the number of pregnancy losses that would have been observed under different levels of exposure to $PM_{2.5}$. This approach provides a more complete picture of the association and uncertainty as the results presented in Figures 2 and 3 are based only on the NB model, whereas the g-computation results here account for the zero-inflation and its uncertainty. In addition, this approach accounts for modification of the $PM_{2.5}$ by temperature. The estimand of interest, expected pregnancy loss difference (EPLD) for $PM_{2.5}$, is

$$\text{EPLD}_{1t} = \sum_i \left[E\left(Y_i | X_{i1t} = x_1^{(0.25)}\right) - E\left(Y_i | X_{i1t} = x_1^{(0.75)}\right) \right], \tag{2.17}$$

where $x_1^{(0.25)}$ and $x_1^{(0.75)}$ denote the 25th and 75th percentile of all measurements of $PM_{2.5}$ across all county-weeks, respectively. This is the marginal difference in expected pregnancy losses under two counterfactual exposures scenarios for one component at one time point across all counties in the state. The first assumes that $PM_{2.5}$ exposure during week t is at the 25th percentile of $PM_{2.5}$ for all county-weeks, while the second assumes exposure for that same week is set to the 75th percentile of $PM_{2.5}$ for all county-weeks. We estimate the difference in expected LBIC where the negative difference in LBIC is the difference in the number of pregnancy losses. Using the observed data

and fitted model, we estimate the EPLD for each gestational week t as

$$\widehat{\text{EPLD}}_{1t} = \sum_i n_i^{-1} \sum_{j=1}^{n_i} \Big[\widehat{E}\left(Y | X_{1t} = x_1^{(0.25)}, \mathbf{X}_{1[t]} = \mathbf{x}_{ij1[t]}, \mathbf{X}_2 = \mathbf{x}_{ij2}, \mathbf{Z}_1 = \mathbf{z}_{1ij}, \mathbf{Z}_2 = \mathbf{z}_{2ij}\right) - \widehat{E}\left(Y | X_{1t} = x_1^{(0.75)}, \mathbf{X}_{1[t]} = \mathbf{x}_{ij1[t]}, \mathbf{X}_2 = \mathbf{x}_{ij2}, \mathbf{Z}_1 = \mathbf{z}_{1ij}, \mathbf{Z}_2 = \mathbf{z}_{2ij}\right) \Big], \tag{2.18}$$

where $\mathbf{X}_{1[t]} = \mathbf{x}_{ij1[t]}$ denote fixing measurements of $PM_{2.5}$ at its observed level of location i and study week j, excluding gestational week t. For ease of interpretation, we scale estimated EPLD to be the total expected losses per month.

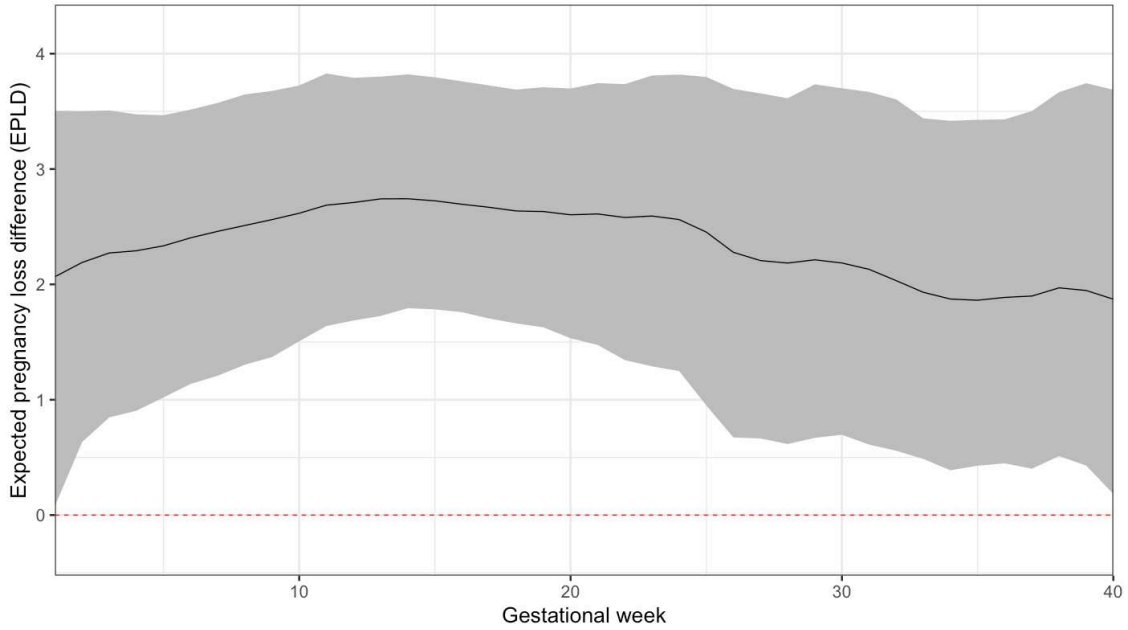


Figure 2.4: Estimated EPLD per month for $PM_{2.5}$ for each week post conception. The gray area indicates the 95% credible interval of the estimate.

Figure 2.4 shows the estimated EPLD for each gestational week. The EPLD is scaled by the total number of months, 130. Across all weeks, the increment of $PM_{2.5}$ exposure from the 25th (4.19 $\mu g/m^3$) to the 75th (6.40 $\mu g/m^3$) percentile was associated with 2.35 EPLD per month. Week 14 had the largest average of 2.74 losses, which translates to 2.2 EPLDs per 1000 LBICs per month across all counties in Colorado. The window of susceptibility of $PM_{2.5}$ spanning the entire

gestational week shown in Figure 2.2 is reflected in the credible interval of the EPLD estimate. We conducted Bayesian g-computation exclusively on $PM_{2.5}$ as the null result of the estimated distributed lag effect of temperature will translate to null changes in the EPLD.

2.6 Discussion

We have proposed a new method to estimate the association between an environmental mixture component that is assessed longitudinally and a count-valued health outcome. Previous literature has suggested various approaches to examine the relationship between longitudinally assessed environmental mixtures and outcomes. However, none of these methods are available for zero-inflation or overdispersed count outcome data. To overcome these limitations, we developed a model by combining the ZINB framework and a tree ensemble approach of TDLMM (Neelon, 2019; Mork and Wilson, 2023). Our flexible model offers several advantages. First, it can account for zero-inflation and overdispersion in the count data. Furthermore, our model can estimate the main exposure effect of each component and simultaneously accommodate time-resolved interaction effects in a mixture component setting. Lastly, it performs component selection and improves windows of susceptibility identification with the shrinkage prior.

Through a simulation study, we demonstrated that our model outperforms a natural spline approach in terms of estimating the exposure effects with lower RMSE, improved credible interval coverage rate, and identifying windows of susceptibility with higher precision. These results suggest that employing a tree structure to model a distributed lag function leads to more accurate estimation.

In data analysis, we examined the association between exposure to environmental components and pregnancy loss by regressing $PM_{2.5}$ and temperature on the number of births in the counties of Colorado with the LBIC framework. Our findings showed that the window of susceptibility of $PM_{2.5}$ spans across the entire 40 gestational weeks, notably weeks 10 - 24. Additionally, we found that the negative exposure effect of $PM_{2.5}$ throughout gestation was persistent even after accounting for its positive lagged interaction with temperature. Lastly, through Bayesian g-computation,

we estimated a monthly average of 2.35 EPLD when maternal exposure to $PM_{2.5}$ increased by approximately 2 $\mu g/m^3$ in any gestational week. Across all gestational weeks, the estimate was equivalent to 1.8 EPLDs per 1000 LBICs on average per month across all counties in Colorado.

Our results add to existing literature providing evidence of an association between air pollution and pregnancy loss (Kioumourtzoglou et al., 2019). However, due to the new methodology including the TDLMM model presented here, we are able to more precisely estimate these models by eliminating the possible effect of variance inflation with spline-based models identified by Leung et al. (2023). Furthermore, the methods enabled our study to be the largest study of air pollution and pregnancy loss and the first to use the LBIC design to consider mixtures. Finally, the methods presented here will enable not only additional studies using the LBIC design for pregnancy loss but also studies for lagged mixtures for count outcomes such as mortality and hospitalization, which have been elusive due to a lack of appropriate statistical methodology.

Chapter 3

Heterogeneous Distributed Lag Mixture Model for Precision Environmental Health With Longitudinally Assessed Mixture Exposures

3.1 Introduction

Precision environmental health focuses on individualized risk assessment for the targeted prevention of disease. This paradigm is motivated by the precision medicine movement but seeks to consider a broad set of environmental factors including chemical exposures, weather, built environment, and social context. Precision environmental health is supported by substantial evidence that the effects of environmental exposures are heterogeneous among individuals with different genetic (Baccarelli et al., 2023), demographic (Lee et al., 2018; Chiu et al., 2022), socioeconomic, social and other neighborhood factors (Casey et al., 2016; Brunst et al., 2018). By identifying factors and subgroups that define sensitive populations or increased vulnerability, practitioners can intervene directly either by modifying factors that are found to increase sensitivity to exposures or target interventions to sensitive sub-populations that may be defined by multiple modifiable and non-modifiable factors.

In addition to accounting for heterogeneity across a potentially high dimensional set of candidate modifying factors, precision environmental health analyses are further complicated by exposures to environmental mixtures and exposure timing. There is ample research on the health effects of exposure to environmental mixtures, defined as simultaneous exposure to multiple environmental factors (e.g. multiple air pollutants or air pollution and temperature) (Billionnet et al., 2012; Park et al., 2014; Anenberg et al., 2020). There is also evidence that the effect of exposure may vary across the life course. An important and special case of this is the impact of maternal exposure during pregnancy on birth and childrens health outcomes. In this context, a core concept

is windows of susceptibility, a time period during the fetal developmental process when a fetus is particularly vulnerable (Wright, 2017). Studies have separately addressed who is most susceptible, what mixtures or mixture component individuals are most sensitive to, or when during the life course individuals are most susceptible. However, due to a lack of sufficient statistical methodology, no epidemiological studies have simultaneously addressed all three aspects in precision environmental health studies.

There are existing statistical methods for longitudinally assessed exposures and mixture exposures to study the homogeneous effect of environmental exposures. For a single longitudinally assessed exposure, a popular choice of statistical method is a distributed lag model (DLM) (Schwartz, 2000; Gasparrini et al., 2010). The DLM regresses an outcome observed at a single time point on exposure measurements taken during the preceding time period. The DLM framework using daily or weekly average exposure results in lower bias compared to using exposure averaged over a pre-specified period, such as the trimesters of pregnancy (Wilson et al., 2017b). DLMs typically involve a smoothing constraint to regularize the time-varying exposure effect in the presence of autocorrelation among repeated measures of exposure, such as polynomials, splines, Gaussian processes, and regression trees. The DLM framework has been extended to mixture exposures (e.g. air pollutants and temperature) that are all assessed longitudinally. Existing work to accommodate mixture-exposure-time-response relationship includes additive models (Bello et al., 2017; Wang et al., 2023), models with lagged interaction effect among components (Warren et al., 2022; Mork and Wilson, 2023), and high-dimensional functional approaches (Liu et al., 2018; Wilson et al., 2022; Antonelli et al., 2024).

The majority of environmental health studies that address heterogeneity or modification, including studies of mixtures or longitudinally assessed exposures, rely on stratification or effect modification with a single or a select few modifiers chosen prior to the analyses (Wilson et al., 2017a; Hsu et al., 2023; Dematteis et al., 2024). Conducting multiple sequential analyses for modification by different variables presents challenges related to multiple testing, potentially overlooking modification by correlated candidate modifiers, and the failure to identify sub-populations defined

by a combination of multiple modifiers. Notably for multiple modifiers, Odden et al. (2020) proposed a modified random forest for a single component observed at one time point, and Mork et al. (2023) introduced a nested tree structure model to incorporate multiple modifiers for a single longitudinally assessed exposure. However, there is no statistical method available for heterogeneous analysis of longitudinally assessed mixture exposures.

We propose a framework to estimate the heterogeneous effect of longitudinally assessed mixture exposures. We refer to our model as a heterogeneous distributed lag mixture model (HDLMM). Specifically, we parameterize our HDLMM with a Bayesian additive model consisting of an ensemble of tree triplets. A tree triplet consists, first, of a regression tree that partitions the sample into mutually exclusive subgroups based on a set of candidate modifying variables. Each tree triplet is completed with two additional regression trees that define the time-resolved main effects and time-sensitive pairwise interactions for two longitudinally assessed mixture components. The HDLMM framework with the tree triplet structure accounts for multi-level interactions of modifiers and exposure timing, which enables identifying a sensitive subgroup that is only defined by the intersection of multiple modifying variables and the corresponding susceptible time period of exposure. The model regularizes the time-varying effects to account for high autocorrelation between repeated measures of exposure and performs variable selection on mixture components. In contrast to a stratified analysis with pre-specified modifiers, the proposed method performs modifier selection and takes a data-driven approach to determine the modifiers that contribute to heterogeneity. This leads to more effective identification of vulnerable individuals and subgroups and estimation of individualized exposure-time-response relationships based on a potentially large number of modifying factors.

Our model holds significance in three key aspects. First, this is the first model to simultaneously account for the who, what, and when of environmental exposures. That is, the proposed HDLMM can simultaneously identify who is most susceptible, when they are most susceptible, and to what in a mixture they are most susceptible. Further with an emerging need for a precision environmental health analysis, the proposed method pertains to tailored risk assessment for more

effective intervention. Second, we introduce a novel tree triplet structure to estimate the HDLMM in a coherent Bayesian framework that provides the appropriate parameterization and regularization to each component of the model. Finally, this is the first study to employ a fully data-driven approach to examine the heterogeneous lagged effects of air pollution mixture exposures on birth weight.

3.2 Colorado administrative cohort data

We analyze administrative data on Colorado, USA birth records obtained from the Colorado Department of Public Health and Environment. The data includes all registered births in the state of Colorado between 2007 and 2018, inclusive.

We included mother-child dyads of full-term ($\geq$ 37 weeks of gestation) singleton births with dates of conception between January 1, 2007, and December 31, 2017. We excluded the dyads whose residences were above 6,000 feet to account for the confounding effect from altitude and to reduce the variability in exposure measurement error resulting from both altitude and terrain. Finally, we restricted our analysis to the Front Range area. The total number of dyads included in the analysis is 402,853. Descriptive statistics of the covariates are shown in Appendix B.1.

We considered an association between exposure to fine particulate matter air pollution ($PM_{2.5}$) and maximal daily temperature during the first 37 weeks of pregnancy and birth weight. $PM_{2.5}$ and temperature are of particular interest due to frequent co-occurrence due to wildfires, particularly in the study region of Colorado. For each dyad, we considered birth weight for gestational age z-score (BWGAZ) as the outcome. For the exposure data, we obtained $PM_{2.5}$ exposure measurements from the United States Environmental Protection Agency (EPA) community multi-scale air quality modeling system using down-scaled data (Berrocal et al., 2010). For temperature data, we obtained the maximal daily temperature from the gridded surface meteorological dataset which contained the centroids of a 4-kilometer grid (Abatzoglou, 2013). For both $PM_{2.5}$ and temperature, we calculated the weekly average levels for each census tract.

3.3 Methods

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a vector of continuous outcomes for a sample $i = 1, \ldots, n$. In our data analysis, y_i represents BWGAZ for dyad i. We are interested in lagged exposure to an environmental mixture of Q components, where $Q \geq 2$. We assume that exposure to each component of the mixture is observed at T equally spaced time points, which we denote as $\boldsymbol{x}_{iq} = (x_{iq1}, \ldots, x_{iqT})$ for $q = 1, \ldots, Q$. In our context, we have $T = 37$ gestational weeks and $Q = 2$ components with $\boldsymbol{x}_{i1}$ and $\boldsymbol{x}_{i2}$ denoting $PM_{2.5}$ and temperature, respectively. We also observe a vector of covariates $\boldsymbol{z}_i$, which we will assume includes a 1 for the intercept in the model, of which a subset $\boldsymbol{m}_i \subseteq \boldsymbol{z}_i$ is a set of candidate modifying variables that may govern the heterogeneity in the exposure-time-response relationship.

The proposed HDLMM is

$$y_i = f(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iQ}, \boldsymbol{m}_i) + \boldsymbol{z}_i'\boldsymbol{\gamma} + \varepsilon_i,$$

where f is an unknown exposure-time-response function dependent on the set of candidate modifiers and the longitudinally assessed mixture exposures, $\boldsymbol{\gamma}$ is a vector of regression coefficients, and ε_i is independent error distributed Normal$(0, \sigma^2)$. We parameterize f as

$$f(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iQ}, \boldsymbol{m}_i) = \sum_{q=1}^{Q}\sum_{t=1}^{T} x_{iqt}\theta_{qt}(\boldsymbol{m}_i) + \sum_{q_1=1}^{Q-1}\sum_{q_2=q_1+1}^{Q}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T} x_{iq_1t_1}x_{iq_2t_2}\theta_{q_1q_2t_1t_2}(\boldsymbol{m}_i), \quad (3.1)$$

where $\theta_{qt}(\boldsymbol{m}_i)$ is the main exposure effect of component q at time t and $\theta_{q_1q_2t_1t_2}(\boldsymbol{m}_i)$ is the interaction effect between component q_1 at time t_1 and component q_2 at time t_2 specific to individuals with modifier levels $\boldsymbol{m}_i$. The HDLMM in (3.1) takes the same general functional form as the distributed lag mixture model used by Mork and Wilson (2023) and Antonelli et al. (2024) but allows for the effects to vary as a function of the modifiers. The HDLMM in (3.1) is highly parameterized and the exposure measurements are likely correlated both over time and between components at each

time. Therefore, regularization of the exposure-time-response function is necessary. We propose a Bayesian tree approach to add structure that regularizes this model and allows for heterogeneity.

3.3.1 HDLMM framework with tree triplet structure

We introduce a tree triplet structure to estimate the parameters of the HDLMM in (3.1). We employ an additive approach by creating an ensemble of A tree triplets indexed by $a = 1, \ldots, A$. For the sake of clarity, we focus our model presentation on a single tree triplet. Figure 3.1 illustrates a single tree triplet. A tree triplet denoted $\{\mathcal{M}_a, \mathcal{T}_{a1}, \mathcal{T}_{a2}\}$ is composed of two types of tree: a modifier tree denoted $\mathcal{M}_a$ and DLM trees denoted $\mathcal{T}_{a1}$ and $\mathcal{T}_{a2}$.

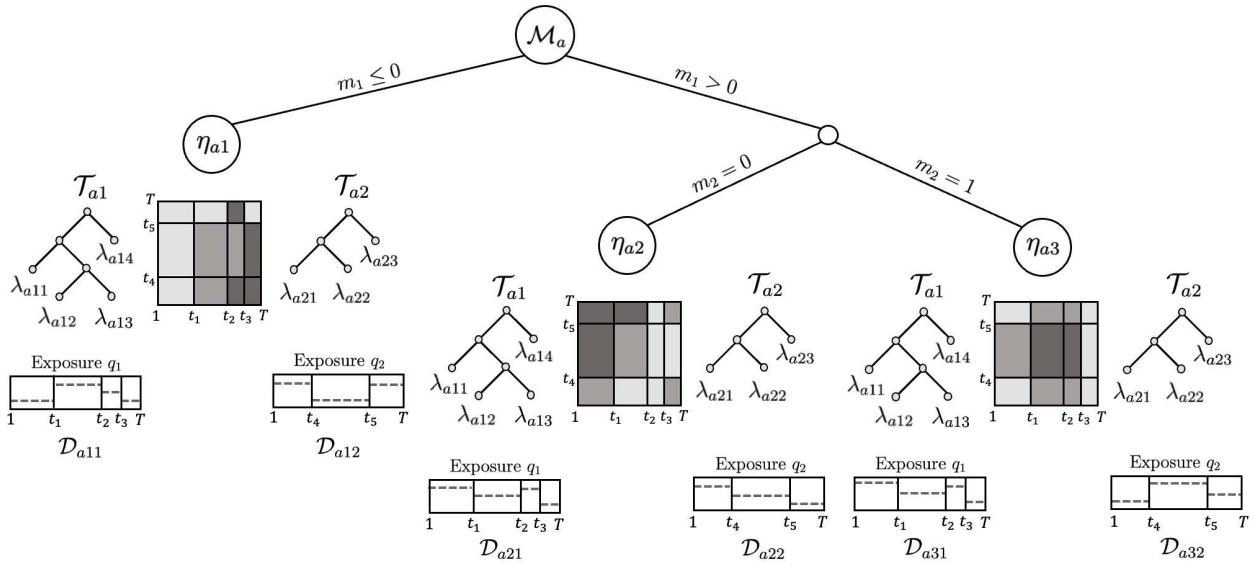


Figure 3.1: A diagram of a tree triplet. A modifier tree $\mathcal{M}_a$ splits the modifier space, which equates to partitioning the population into subgroups. Each terminal node of the modifier tree η_{ab} is linked to a pair of DLM trees denoted $\mathcal{T}_{a1}$ and $\mathcal{T}_{a2}$. The DLM trees are each assigned mixture components q_1 or q_2. The DLM tree structures and assigned mixture components are assumed to be shared across η_{ab}. For each terminal node of a DLM tree, the corresponding time lag interval is assigned a unique constant effect (dashed lines). The interaction surface is jointly defined with the structures of the two DLM trees and similarly assigned a constant effect (shaded boxes).

A modifier tree $\mathcal{M}_a$ is a binary regression tree applied to the candidate modifiers $\boldsymbol{m}_i$. The terminal nodes of the modifier tree are denoted η_{ab} for $b = 1, \ldots, B_a$ where B_a is a total number

of terminal nodes of modifier tree $\mathcal{M}_a$. Each terminal node defines a non-overlapping subgroup of the sample partitioned based on modifiers included in that modifier tree.

Each tree triplet is completed by a pair of trees referred to as DLM trees (Mork and Wilson, 2023). A DLM tree provides structure to the exposure-time-response relationship for two components of the mixture and their interaction surface for all subgroups defined by the terminal nodes of the modifier tree. Each DLM tree is a binary tree that partitions the exposure time span, $t = 1, \ldots, T$, into non-overlapping segments. The terminal nodes of the resulting DLM tree are denoted λ_{apc} for $c = 1, \ldots, C_{ap}$ where C_{ap} is a total number of terminal nodes of $\mathcal{T}_{ap}$ for $p = 1, 2$. Each DLM tree is assigned a single component of the mixture exposures. We let $S_{ap} = q$ denote component q is assigned to DLM tree $\mathcal{T}_{ap}$.

When put together, a tree triplet partitions the parameter space into subgroups based on the modifiers and exposure time for each of the two components in mixture exposures. Each combination of terminal nodes from the modifier tree and each of the DLM trees is assigned a scalar terminal node parameter. The scalar parameter contributes to the linear effect of exposure at all exposure times contained in the terminal node of the DLM tree for individuals in that terminal node of the modifier tree. The scalar parameter is denoted δ_{abpc} for triplet $a = 1, \ldots, A$, modifier tree terminal node $b = 1, \ldots, B_a$, DLM tree $p = 1, 2$, and DLM tree terminal node $c = 1, \ldots, C_{ap}$. A set of scalar parameters of a DLM tree $\mathcal{T}_{ap}$ linked to η_{ab} is denoted $\mathcal{D}_{abp} = \{\delta_{abp1}, \ldots, \delta_{abpC_{ap}}\}$. Each combination of $\mathcal{M}_a$ and $\mathcal{T}_{ap}$ linked to η_{ab} represents the main exposure effect as a function $g(\boldsymbol{m}_i, t|\mathcal{M}_a, \mathcal{T}_{ap}, \mathcal{D}_{abp}) = \delta_{abpc}\mathbb{I}(\boldsymbol{m}_i \in \eta_{ab}, t \in \lambda_{apc})$. In addition to the terminal node parameters that define main effects, a tree triplet allows for interaction effects between mixture components. The interaction surface of a tree triplet denoted $\mathcal{T}_{a1} \times \mathcal{T}_{a2}$ is constructed fully based on the time segments of the two DLM trees. As shown in Figure 3.1, each combination of terminal nodes from the modifier tree and two DLM trees is assigned an additional scalar parameter for the interaction surface. The parameter $\omega_{abc_1c_2}$ represents the interaction effect between two components assigned to two DLM trees at all exposure times contained in a terminal node c_1 of the first DLM tree and at all exposure times contained in a terminal node c_2 of the second DLM tree, specific to a termi-

nal node of the modifier tree, η_{ab}. We define the interaction effect of $\mathcal{T}_{a1} \times \mathcal{T}_{a2}$ linked to η_{ab} as a function $g_I(\boldsymbol{m}_i, t_1, t_2 | \mathcal{M}_a, \mathcal{T}_{a1} \times \mathcal{T}_{a2}, \Omega_{ab}) = \omega_{abc_1c_2} \mathbb{I}(\boldsymbol{m}_i \in \eta_{ab}, t_1 \in \lambda_{a1c_1}, t_2 \in \lambda_{a2c_2})$, where Ω_{ab} is the set of interaction effects for $\mathcal{T}_{a1} \times \mathcal{T}_{a2}$ linked to η_{ab}.

The parameters in (3.1) can be reconstructed from an ensemble of tree triplets. The main effect of component q at time t conditional on $\boldsymbol{m}_i$ with the ensemble of A tree triplets is

$$\theta_{qt}(\boldsymbol{m}_i) = \sum_{a=1}^{A} \sum_{b=1}^{B_a} \sum_{p=1}^{2} \sum_{c=1}^{C_{ap}} g(\boldsymbol{m}_i, t | \mathcal{M}_a, \mathcal{T}_{ap}, \mathcal{D}_{abp}) \mathbb{I}(S_{ap} = q), \tag{3.2}$$

where c is an index for a set $\mathcal{D}_{abp}$. Similarly, the interaction effect between component q_1 at time t_1 and component q_2 at time t_2 conditional on $\boldsymbol{m}_i$ is

$$\theta_{q_1q_2t_1t_2}(\boldsymbol{m}_i) = \sum_{a=1}^{A} \sum_{b=1}^{B_a} g_I(\boldsymbol{m}_i, t_1, t_2 | \mathcal{M}_a, \mathcal{T}_{a1} \times \mathcal{T}_{a2}, \Omega_{ab}) \mathbb{I}(S_{a1} = q_1, S_{a2} = q_2).$$

The HDLMM framework can be simplified by assuming no interaction effects between components in (3.1) to obtain a purely additive model (HDLMMadd).

3.3.2 Prior specification

A tree triplet for the HDLMM framework requires a prior specification on the structure of the trees and the terminal node-specific parameters for the main and interaction effects. The model additionally requires priors to incorporate variable selection on candidate modifiers and components of the mixture exposures.

For a modifier tree structure, we follow the prior specification of Chipman et al. (2010). For a DLM tree structure, we follow Mork and Wilson (2023). Full details of priors on modifier and DLM tree structures are provided in Appendix B.2. For the terminal node parameters, we set $\delta_{abpc} \sim \text{Normal}(0, \tau_a^2 \nu^2 \sigma^2)$ where all hyperparameters follow half-Cauchy$(0, 1)$. We similarly set the prior of the interaction effect as $\omega_{abc_1c_2} \sim \text{Normal}(0, \tau_a^2 \nu^2 \sigma^2)$. The hyperparameters impose global-local shrinkage where ν addresses the global shrinkage across all tree triplets and the tree triplet-specific hyperparameter τ_a shrinks the effect of poor fitting tree triplets (Carvalho et al.,

2010; Mork and Wilson, 2022). Finally for regression coefficients, the prior on $\boldsymbol{\gamma}$ is MVN$(0, k\sigma^2\boldsymbol{I})$ where k is fixed at a large number.

For modifier selection, we consider the L-vector of probabilities denoted $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_L)$ for selecting a modifier from the L candidate modifiers to include in the internal nodes of a modifier tree. For the prior of $\boldsymbol{\psi}$, we follow Linero (2018) for modifier selection and specify $\boldsymbol{\psi} \sim$ Dirichlet$(\kappa/L, \ldots, \kappa/L)$, $\kappa/(\kappa + L) \sim$ Beta$(\zeta, 1)$, where $\zeta \in (0.5, 1)$. We fixed $\zeta = 0.5$ for simulation and data analysis for greater sparsity of modifiers. For component selection, we impose categorical-Dirichlet prior on S_{ap} such that $S_{ap}|\mathcal{E} \sim$ Categorical$(\mathcal{E})$ and $\mathcal{E} \sim$ Dirichlet$(\phi, \ldots, \phi)$, where $\mathcal{E} = \{E_1, \ldots, E_Q\}$ is a vector of probabilities that a DLM tree is assigned component q and ϕ is a sparsity hyperparameter. Using this formulation, the model chooses component q only if it is included in the DLM trees. This also leads to hierarchical component selection as the time-resolved interaction between two mixture components can only be included if the main effects of both components are included in the model. To ensure that the probability of each component included at least once in 20 tree triplets is 0.9, we follow Mork and Wilson (2023) and fixed ϕ to 0.638 for $Q = 2$ and 0.786 for $Q = 3$.

3.3.3 Computation

We use Markov chain Monte Carlo (MCMC) to update the ensemble of tree triplets. To update each tree triplet, we sequentially update the modifier tree and each DLM tree with the Metropolis-Hastings algorithm. We propose and update the modifier tree with the four transition steps used in Chipman et al. (2010): grow, prune, change, and swap. For the proposal for DLM trees, we grow a new tree structure from the root node with a specified splitting probability prior and assign a randomly sampled component from the component selection prior. In our simulation and data analysis, growing a DLM tree from root proved to be more effective than the transition steps used in Chipman et al. (2010).

For both simulation and data analysis, we verified the good mixing and convergence of the MCMC algorithm by examining traceplots and comparing the results from different MCMC tuning

parameters. Changes in the model parameter specification including the number of tree triplets and hyperparameters did not significantly affect the results. A detailed outline of the algorithm and convergence diagnostics are provided in Appendix B.3 and B.4.

3.4 Simulation

We evaluated the operating characteristics of the HDLMM for estimating heterogeneous exposure effects and identifying a window of susceptibility in a mixture exposure setting. We considered two scenarios. In scenario 1, we tested the model with three subgroups, each susceptible to a different component of a mixture. In scenario 2, we applied the model to a more complex structure which included individual scaled effects and interactions among components of the mixture exposures. All scenarios can be replicated with R package **dlmtree**.

We fit two versions of our proposed model: HDLMM and its simplified version HDLMMadd, which assumes no interaction among components of the mixture exposures. We compared our proposed models with two different models. First, we fit the treed distributed lag mixture model (TDLMM; Mork and Wilson, 2023) which assumes that exposure effects and window of susceptibility are homogeneous across all observations. We let TDLMM account for lagged interaction effects among components. Second, specifically for scenario 1, we stratified the dataset with the true subgroups used to generate the data and fit TDLMM separately on each stratum. We refer to this model as TDLMM with fixed subgroups. This model is correctly specified but unrealistic in practice. For both scenarios, we fit an ensemble of 20 tree triplets or tree pairs for all models. We performed the MCMC algorithm for 10,000 iterations thinning every ten iterations after discarding 2,500 iterations as burn-in.

We assessed the performance of HDLMM in terms of modifier and component selection, exposure effect estimation, and window of susceptibility identification. We evaluated the model performance with the following measures: modifier posterior inclusion probability (mPIP) and component posterior inclusion probability (cPIP) for modifier and component selection, root mean-squared error (RMSE) and credible interval (CrI) coverage for estimating exposure effects, and true

positive (TP), false positive (FP), and precision for identifying a window of susceptibility. For the model comparison, we used a model selection criterion of mean-squared prediction error (MSPE). Details on these metrics are in Appendix B.5.

3.4.1 Scenario 1: Three subgroups with no interaction

For scenario 1, we generated a continuous response variable with three subgroups and corresponding exposure effects from three components. We generated 100 independent datasets each containing a sample size of 5,000. We considered 10 covariates for $\boldsymbol{z}_i$ where five variables are sampled from Normal(0, 1) and the other five are sampled from Bernoulli(0.5). The covariates do not affect the heterogeneity but we included these covariates as candidate modifiers to evaluate the performance of modifier selection. We drew $\boldsymbol{\gamma}$ and ε_i from Normal$(0, 1)$.

The subgroups are determined by a continuous and a binary modifier randomly generated from Normal$(0, 1)$ and Bernoulli(0.5), each denoted as m_c and m_b. As shown in (3.3), we partitioned the first subgroup and second subgroup with m_b and separated the third subgroup with m_c from the other groups. The heterogeneous effect of the data generating model is

$$
\begin{aligned}
f(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iQ}, \boldsymbol{m}_i) &= \sum_{q=1}^{Q} \sum_{t=1}^{T} x_{iqt} \theta_{qt}(\boldsymbol{m}_i) \\
\theta_{qt}(\boldsymbol{m}_i) &= \begin{cases} \mathbb{I}(q = 1, t \in [t_1, t_1 + 7]) & \text{if } m_{ci} > 0 \text{ and } m_{bi} = 1 \\ \mathbb{I}(q = 2, t \in [t_2, t_2 + 7]) & \text{if } m_{ci} > 0 \text{ and } m_{bi} = 0 \\ \mathbb{I}(q = 3, t \in [t_3, t_3 + 7]) & \text{if } m_{ci} \leq 0, \end{cases}
\end{aligned} \tag{3.3}
$$

where t_1, t_2, and t_3 denote starting points of windows of susceptibility. All starting points were randomly sampled from weeks among $\{1, \ldots, T-7\}$. We obtained $PM_{2.5}$, temperature, and ozone (O_3) exposure data in the Front Range area of Colorado between January 1, 2000, and December 31, 2020, from the EPA to ensure realistic correlations in the exposure data. We calculated the weekly averages for each component and randomly selected dates from those for which complete exposure data for all 37 weeks preceding the chosen date were available. $PM_{2.5}$, temperature, and

O_3 are indicated with $q = 1, 2,$ and 3, respectively. All exposure data were scaled and continuous covariates were standardized prior to fitting.

Table 3.1: Model performance measures of HDLMM, HDLMMadd, TDLMM, and TDLMM with fixed subgroups (FS-TDLMM) for scenario 1: The first two columns show the average RMSE and CrI coverage rate across the three components. The next three columns show the TP, FP, and precision for identifying windows of susceptibility (WOS). The last column shows a 10-fold cross-validation mean-squared prediction error (MSPE).

	Effect Estimation		WOS			
Model	RMSE$\times 100$	Coverage	TP	FP	Precision	MSPE
HDLMM	1.176	0.964	0.965	0.020	0.980	1.041
HDLMMadd	1.148	0.960	0.962	0.023	0.978	1.041
TDLMM	3.033	0.721	0.865	0.170	0.839	28.95
FS-TDLMM	0.582	0.995	0.975	0.003	0.997	

Table 3.1 presents the simulation results for scenario 1. Overall, TDLMM with fixed subgroups performed the best on the effect estimation, which is not surprising as the model was provided with true subgroups. The RMSEs of HDLMM and HDLMMadd were acceptable compared to that of TDLMM with fixed subgroups. Both methods also reached high coverage rates of 96%, approximately the target level. TDLMM without heterogeneity assumption performed the worst because the model assumed a combined exposure effect from three components across all observations, rather than to estimate the exposure effect specific to each subgroup. This misspecification further led to a low coverage rate.

For windows of susceptibility identification, TDLMM with fixed subgroups performed the best, yet both HDLMM and HDLMMadd reached a level of performance as high as TDLMM with fixed subgroups with high precision above 0.978. TDLMM performed the worst with high FP rates due to the misspecification of the homogeneous window of susceptibility across all observations. Given three different windows of susceptibility, TDLMM tended to identify most lags as a single window of susceptibility combining the three windows into a single extended window. This led to high TP but resulted in poor precision with high FP.

In terms of model comparison, the MSPE suggested that both HDLMM and HDLMMadd are suitable for this scenario. For both versions of HDLMM, the mPIPs for m_b and m_c were both 1.00, followed by the third highest mPIPs of 0.55 on average. This indicates that the models can identify the modifiers that contribute the most to the effect heterogeneity. The cPIPs for three mixture components were 1.00.

3.4.2 Scenario 2: Scaled effect with interaction

In the second scenario, we add modification by scaling of effects and subgroup-specific interactions. For two subgroups, defined by $m_{ci} > 0$ and $m_{ci} \leq 0$, we assume a main effect of $PM_{2.5}$ that is scaled by a modifier m_{si}; however, the timing of the window varies by subgroup. In addition, for the subgroup with $m_{ci} > 0$, there is an interaction effect between $PM_{2.5}$ and temperature, but no interaction for the other subgroup. The specific exposure-time-response function for this data generating mechanism is

$$f(\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{iQ},\boldsymbol{m}_i) = \sum_{q=1}^{Q}\sum_{t=1}^{T} x_{iqt}\theta_{qt}(\boldsymbol{m}_i) + \sum_{q_1=1}^{Q-1}\sum_{q_2=q_1+1}^{Q}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T} x_{iq_1t_1}x_{iq_2t_2}\theta_{q_1q_2t_1t_2}(\boldsymbol{m}_i)$$

$$\theta_{qt}(\boldsymbol{m}_i) = \begin{cases} m_{si}\mathbb{I}(q=1, t\in[t_4,t_4+7]) & \text{if } m_{ci}>0 \\ m_{si}\mathbb{I}(q=1, t\in[t_5,t_5+7]) & \text{if } m_{ci}\leq 0 \end{cases}$$

$$\theta_{q_1q_2t_1t_2}(\boldsymbol{m}_i) = \begin{cases} 0.025\cdot\mathbb{I}(q_1=1,q_2=2,t_1\times t_2\in[t_4,t_4+7]\times[t_6,t_6+7]) & \text{if } m_{ci}>0 \\ 0 & \text{if } m_{ci}\leq 0, \end{cases}$$

where t_4, t_5, and t_6 denote the starting weeks randomly selected among $\{1,\ldots,T-7\}$ and m_c is a scaling modifier from Uniform$(0,1)$. We generated 100 independent datasets with a sample size of 5,000. The covariates, modifiers, and coefficients were generated similarly to scenario 1. There are two levels of heterogeneity in scenario 2. At a subgroup level with m_c, each subgroup has a different window of susceptibility and associated components. At an individual level, the main effects, which are dependent on m_s, vary uniquely for each individual. TDLMM with fixed

subgroups was removed for scenario 2, as it cannot account for the true scaled effect for each individual. Table 3.2 shows the simulation results for scenario 2. For exposure effect estimation,

Table 3.2: Model performance measures of HDLMM, HDLMMadd, and TDLMM for scenario 2: The first four columns show the RMSE and coverage for exposure effect estimation of $PM_{2.5}$ and temperature (Temp). The next three columns state the TP, FP, and precision for windows of susceptibility (WOS) identification. The last column shows 10-fold MSPE.

	Effect Estimation							
	RMSE$\times$100		Coverage		WOS			
Model	$PM_{2.5}$	Temp	$PM_{2.5}$	Temp	TP	FP	Precision	MSPE
HDLMM	2.000	0.075	0.937	0.998	0.900	0.013	0.986	1.075
HDLMMadd	2.070	0.316	0.925	0.945	0.899	0.045	0.954	1.080
TDLMM	6.170	0.075	0.511	0.999	0.980	0.093	0.915	29.59

both versions of HDLMM outperformed TDLMM. For $PM_{2.5}$, HDLMM had the lowest RMSE and the highest coverage rate. HDLMMadd performed slightly worse than HDLMM with higher RMSEs and lower coverage rates for both components because it did not allow interactions among mixture components. The homogeneous TDLMM performed the worst as the model again does not account for the scaled exposure effect and the heterogeneous window of susceptibility of the two subgroups. For a window of susceptibility identification, HDLMM outperformed the homogeneous TDLMM. This was, again, due to TDLMM assuming the same window of susceptibility across all observations. HDLMM reached a high precision of 98.6%.

The MSPE criterion suggested using the HDLMM for scenario 2. HDLMM had the lowest MSPE of 1.075 followed by HDLMMadd with 1.080. HDLMMadd was subject to more prediction error due to its model misspecification on the interaction among components. For both versions of HDLMM, the mPIPs for m_c and m_s were 1.00 and 0.99, respectively. The cPIPs for two mixture components were 1.00.

3.5 Estimands for inference

The HDLMM estimates distributed lag effects while integrating three key aspects: identifying the most susceptible subgroups, determining the impact of different mixture components, and iden-

tifying a window when individuals are vulnerable. However, the complexity of the problem and model requires guidance on how to interpret the results and make inferences to address specific questions in precision environmental health. In particular, the parameters in (3.1) are difficult to interpret because of the possibility of interactions between mixture components over time. The parameters in (3.1) for distributed lag effect of component q at time t for an individual with a set of modifiers $\mathbf{m}$ can be written as

$$\begin{aligned}\theta_{qt}(\mathbf{m}) = \mathbb{E}\left(Y|X_{qt} = x+1, \mathbf{X}_{q[t]} = 0, \mathbf{X}_{[q]} = 0, M = \mathbf{m}\right) \\ - \mathbb{E}\left(Y|X_{qt} = x, \mathbf{X}_{q[t]} = 0, \mathbf{X}_{[q]} = 0, M = \mathbf{m}\right),\end{aligned} \tag{3.4}$$

where $\mathbf{X}_{q[t]} = 0$ denotes fixing measurements of component q at zero except for time t, $\mathbf{X}_{[q]} = 0$ denotes fixing all exposure measurements at zero across all time lags except for component q, and $M = \mathbf{m}$ specifies that the expectation is specific to $\mathbf{m}$. The effect in (3.4) is a "raw" estimate that represents a change in the outcome for a unit increase in exposure to component q at time t when all other exposure measurements are fixed at zero, hence not fully incorporating lagged interactions in a realistic way. To address this limitation, we highlight two estimands better suited for HDLMM: conditional average treatment effect (CATE) and group-specific average treatment effect (GATE).

The HDLMM framework allows us to estimate the expected outcomes conditioned on different exposure levels of components at one or more time points and different levels of the modifiers. In the context of precision environmental health, CATE is the individualized effect of exposure and can inform on individualized effects of interventions. The most basic CATE for our model is

$$\text{CATE}_{qt}(\mathbf{m}) = \mathbb{E}\left(Y|X_{qt} = x+1, M = \mathbf{m}\right) - \mathbb{E}(Y|X_{qt} = x, M = \mathbf{m}). \tag{3.5}$$

$\text{CATE}_{qt}(\mathbf{m})$ in (3.5) results from marginalizing over levels of exposure to other mixture components and at other times for component q. $\text{CATE}_{qt}(\mathbf{m})$ is the expected difference in outcome for a unit increase in exposure to component q at time t for an individual with a set of modifiers $\mathbf{m}$, while

accounting for its interaction with the co-exposures at observed levels. By definition, $\text{CATE}_{qt}(\mathbf{m})$ is equal to (3.4) for HDLMMadd, which assumes no interaction between mixture components. The CATE can also be interpreted as the average effect among individuals who share the same set of modifiers. For estimating how differences in modifier levels impact the exposure effect, one can use CATE to construct an estimand, $\text{CATE}_{qt}(\mathbf{m}) - \text{CATE}_{qt}(\mathbf{m}^*)$, where $\mathbf{m}^*$ is a set of modifiers reflecting potential differences in susceptibility.

Another goal of precision environmental health is to estimate subgroup-specific effects and to identify subgroups that are most susceptible to environmental exposures. GATE is defined

$$\text{GATE}_{qt}(G) = \mathbb{E}\left(Y|X_{qt} = x+1, M \in G\right) - \mathbb{E}(Y|X_{qt} = x, M \in G)\,, \tag{3.6}$$

where G is a subgroup of interest. In (3.6), $\text{GATE}_{qt}(G)$ is the difference in the expected outcome for a unit increase in exposure to component q at time t for subgroup G, accounting for its lagged interactions with other mixture components and while marginalizing over the joint distribution of other mixture components and modifiers not specified in G. The sample average treatment effect, which is an overall estimate of the entire sample, can be obtained by setting G as the entire sample.

Obtaining the estimates for the CATE and GATE from the proposed model requires marginalization of co-exposures. The GATE further requires marginalizing over modifiers that are irrelevant for defining a subgroup. We use Bayesian g-computation to calculate the estimates. Estimating the marginalized effect of exposure to component q at time t with Bayesian g-computation involves creating two counterfactual datasets. For time t, we create the first counterfactual dataset where the measurement of component q at time t is fixed at the numerical value of one while fixing co-exposures and covariates at their observed levels. The second counterfactual dataset can be similarly generated but fixed at zero. We then obtain the predictive posterior distribution of the outcome with the two datasets and compare the estimates for the desired estimand. Specifically in

our case, $\text{CATE}_{qt}(\mathbf{m})$ is estimated with

$$n^{-1}\sum_{i=1}^{n}\Big\{\widehat{\mathbb{E}}\left(Y|X_{qt}=1,\mathbf{X}_{q[t]}=\mathbf{x}_{iq[t]},\mathbf{X}_{[q]}=\mathbf{x}_{i[q]},\mathbf{Z}=\mathbf{z}_i,M=\mathbf{m}\right) \\ -\widehat{\mathbb{E}}\left(Y|X_{qt}=0,\mathbf{X}_{q[t]}=\mathbf{x}_{iq[t]},\mathbf{X}_{[q]}=\mathbf{x}_{i[q]},\mathbf{Z}=\mathbf{z}_i,M=\mathbf{m}\right)\Big\}.$$

Similarly, $\text{GATE}_{qt}(G)$ can be estimated with

$$|G|^{-1}n^{-1}\sum_{\substack{j\text{ s.t.}\\ \mathbf{m}_j\in G}}\sum_{i=1}^{n}\Big\{\widehat{\mathbb{E}}\left(Y|X_{qt}=1,\mathbf{X}_{q[t]}=\mathbf{x}_{iq[t]},\mathbf{X}_{[q]}=\mathbf{x}_{i[q]},\mathbf{Z}=\mathbf{z}_i,M=\mathbf{m}_j\right) \\ -\widehat{\mathbb{E}}\left(Y|X_{qt}=0,\mathbf{X}_{q[t]}=\mathbf{x}_{iq[t]},\mathbf{X}_{[q]}=\mathbf{x}_{i[q]},\mathbf{Z}=\mathbf{z}_i,M=\mathbf{m}_j\right)\Big\},$$

where $|G|$ is the number of observations of subgroup G. As estimating GATE can be computationally expensive, we randomly sample a reasonable number of observations from a subgroup of interest.

We framed inference around estimands based on a unit change in exposure to a single component at one time point. However, there are many potential variations on the estimand presented. For example, one can conduct inference based on a change for all components at a specific time point or a change in one or more mixture components at multiple time points. More complex and realistic potential exposures, such as simulated exposure levels under different climate scenarios, can be used for hypothetical or stochastic interventions.

3.6 Analysis of Colorado birth registry data

We applied our method to examine the distributed lag effect of $PM_{2.5}$ and the maximal daily temperature on BWGAZ over the gestation of 37 weeks, using the Colorado registry data described in Section 3.2. We selected the following variables as candidate modifiers: maternal age at conception, body mass index (BMI) prior to conception, yearly income, education attainment, marital status, prenatal care, smoking habits, race, Hispanic designation, and child sex. For covariates, we

included the mother's pre-pregnancy weight, height, elevation, maternal residence, and month and year of conception. We also considered all modifiers as covariates but excluded child sex since BWGAZ already accounted for child sex. The descriptions of variables are provided in Appendix B.6.1

We ran three models: HDLMM, HDLMMadd, and TDLMM for 10,000 MCMC iterations thinning every 10th iterations after 2,500 burn-in. We included 20 tree triplets for the HDLMM methods and 20 tree pairs for the TDLMM. We log-transformed $PM_{2.5}$ to account for the skewness, and centered and scaled both $PM_{2.5}$ and temperature prior to fitting. For the HDLMM methods, we followed the priors specified in Section 3.3.2. For modifier and component selection, we fixed the hyperparameters to maximize sparsity.

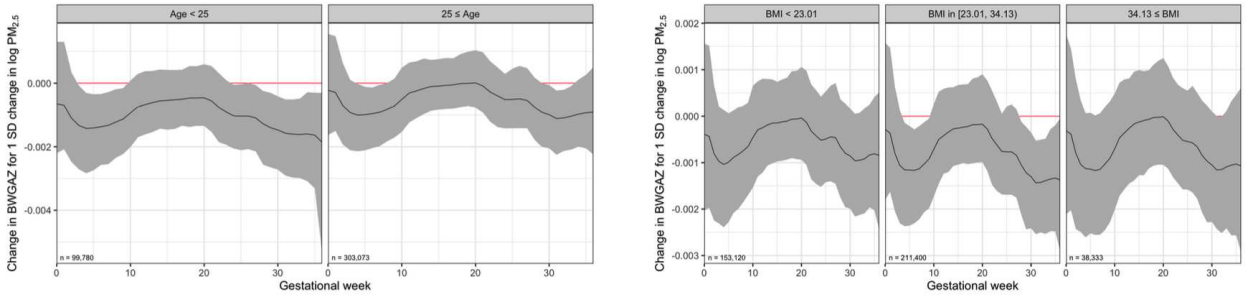
We performed model selection by randomly selecting 10,000 observations for out-of-sample validation and calculating MSPE for each model. We calculated the average MSPE from 20 independent holdout datasets. HDLMMadd showed the best performance with the lowest MSPE of 0.698 followed by TDLMM and HDLMM with 0.702 and 0.704, respectively. Based on the MSPE criterion, we present the results from HDLMMadd. Results from TDLMM of homogeneous effect for comparison are provided in Appendix B.6.2.

3.6.1 Modifier and component selection

The modifiers with the highest mPIPs were BMI (1.000), Hispanic designation (1.000), maternal age (0.998), and race (0.968). The pairs of modifiers with the highest two-way mPIPs, where two modifiers were included in the same modifier tree, were BMI–Hispanic (0.988), age–BMI (0.828), age–race (0.774), and race–BMI (0.588). We focus our discussion of the results on these modifiers. For component selection, cPIPs for $PM_{2.5}$ and temperature were 1.000 and 0.981, respectively.

3.6.2 Group average distributed lag effect estimates

Figure 3.2a and 3.2b show the estimated GATEs grouped by a single continuous modifier while marginalizing over the other modifiers. There are three ways of selecting split points: 1)



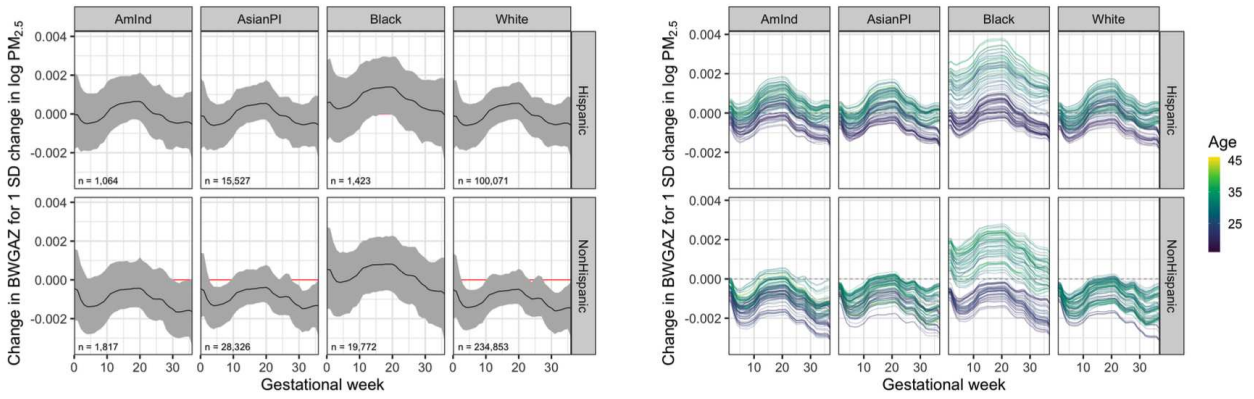
(a) Maternal age

(b) BMI

Figure 3.2: Estimated GATEs of $PM_{2.5}$ grouped by maternal age and BMI. The gray area shows the 95% credible interval for each effect. The sample size of each subgroup is indicated in the bottom left corner. The standard deviations (SD) of log $PM_{2.5}$ and temperature were 0.31 $\log(\mu g/m^3)$ and 9.44 °C, respectively.

choosing the splitting point modes in the modifier trees, 2) inspecting a noticeable change in the distributed lag effects, and 3) considering clinically relevant periods. We used the first approach and chose the split points of age 25 for maternal age and 23.01 and 34.13 for BMI (see Appendix B.6.3). Figure 3.2a presents the estimated GATEs for two subgroups grouped by maternal age. For mothers younger than 25, the negative effect of $PM_{2.5}$ on BWGAZ was more pronounced with the longer window of susceptibility of weeks 4–10, 28–37, compared to the other subgroup. The effect was smaller for mothers who are 25 or older with windows of susceptibility of weeks 5–8, 31–34. Figure 3.2b shows the estimated GATEs of $PM_{2.5}$ on BWGAZ grouped by BMI. $PM_{2.5}$ had the strongest effect on mothers with a BMI between 23.01 and 34.13 with windows of susceptibility of weeks 5–10, 29–37. For mothers with a BMI less than 23.01 or greater than 34.13, we found fewer windows of susceptibility with a smaller effect. Further analyses are provided in Appendix B.6.4.

Figure 3.3a shows the GATEs of $PM_{2.5}$, grouped by race and Hispanic designation. For Native American, Asian, Pacific Islander, and White mothers, $PM_{2.5}$ had a larger effect on BWGAZ among non-Hispanics. Notably for non-Hispanic mothers, the windows of susceptibility during the gestational weeks of Asian and Pacific Islander mothers were weeks 4–10, 29–36, and those of White mothers were weeks 4–11, 29–37. We did not identify any windows of susceptibility for Hispanic mothers of those races. For Black mothers, the GATEs of $PM_{2.5}$ showed a similar trend but with a decreased effect close to zero and no windows of susceptibility. The credible intervals



(a) Subgroup effect

(b) Individual effect

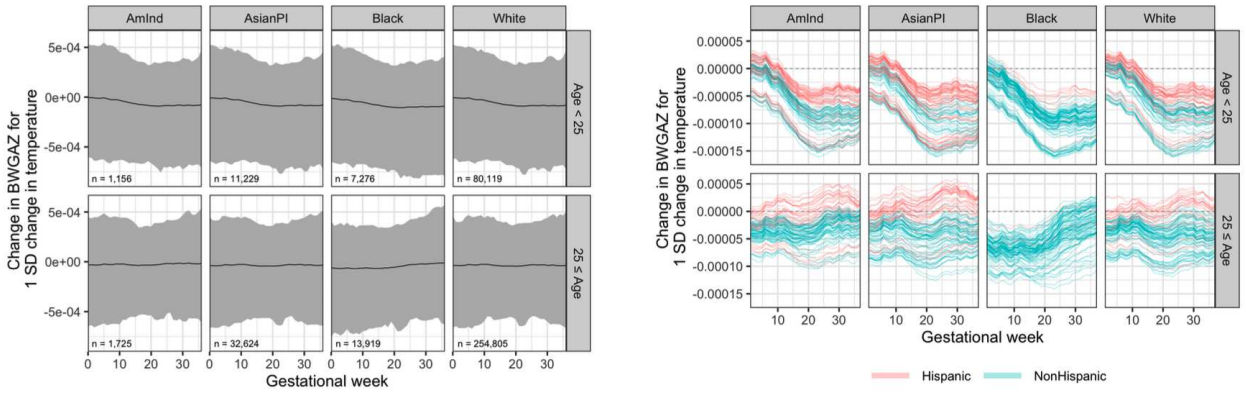
Figure 3.3: Estimated distributed lag effects of $PM_{2.5}$ grouped by race and the Hispanic designation. Panel (a) shows the GATEs where the gray area shows the 95% credible interval for each effect. The sample size of each subgroup is indicated in the bottom left corner. Panel (b) shows the CATEs where each line is colored with the maternal age. Each subgroup includes 100 randomly sampled mothers. This figure appears in color in the electronic version of this dissertation, and any mention of color refers to that version.

of the GATEs were wider for Black mothers. It is worth noting that the subgroups of non-Hispanic, Native American, Asian, Pacific Islander, and White mothers had a similar effect and windows of susceptibility to the marginalized effect of $PM_{2.5}$ from the preliminary analysis shown in Appendix B.6.2, whereas the effect was smaller with fewer windows of susceptibility for the other subgroups.

Figure 3.4a shows estimated GATEs of temperature grouped by race and maternal age. The GATEs hovered around zero with wide credible intervals across all subgroups, implying that the model did not find any noticeable heterogeneity in exposure effects between subgroups. Estimated GATEs of $PM_{2.5}$ and temperature with other subgroups are provided in Appendix B.6.4.

3.6.3 Individualized distributed lag effect estimates

While the subgroup analysis provides a tool for identifying the susceptible subgroups, the results only conveyed differences in exposure effect between subgroups defined by one or two modifiers with the other modifiers marginalized. When we are interested in CATE, all modifiers contribute to induce variability in exposure effects for each individual. We present CATEs for $PM_{2.5}$ and temperature to incorporate all modifiers into the exposure effects.



(a) Subgroup effect

(b) Individual effect

Figure 3.4: Estimated distributed lag effects of maximal daily temperature grouped by race and age. Panel (a) shows the GATEs where the gray area indicates the 95% credible interval. The sample size of each subgroup is indicated in the bottom left corner. Panel (b) shows the CATEs where each line is colored by the mother's Hispanic designation. Each subgroup includes 100 randomly sampled mothers. This figure appears in color in the electronic version of this dissertation, and any mention of color refers to that version.

Figure 3.3b shows estimated CATEs of $PM_{2.5}$ on BWGAZ, grouped by race, Hispanic designation, and maternal age. In contrast to the GATEs, a unique exposure effect is shown for each individual due to additional heterogeneity from all other modifiers. Across all races, $PM_{2.5}$ had a larger magnitude of negative effect on BWGAZ for younger mothers and the difference was more pronounced for Black mothers. The difference in estimated GATEs of temperature in Figure 3.4a was minuscule, but we highlight different trends of its CATEs on each subgroup. Figure 3.4b shows estimated CATEs of temperature on BWGAZ, grouped by race, age, and Hispanic designation. For mothers younger than 25, the CATEs began to decline in the earlier gestational weeks and then stabilized thereafter as a negative effect. For Native American, Asian, Pacific Islander, and White mothers who are 25 or older, the CATEs of temperature were relatively steady with those for most Hispanic mothers hovering relatively close to zero across gestational weeks. The temperature was more negatively associated with BWGAZ for non-Hispanic mothers who are 25 or older. For Black mothers who are 25 or older, the effect was reversed compared to Black mothers who are younger than 25, displaying a negative effect in the early gestational stage which progressively approached zero in the later stage. Estimated CATEs of $PM_{2.5}$ and temperature with other subgroups are provided in Appendix B.6.4.

3.7 Discussion

We introduced the HDLMM framework and proposed a tree triplet structure to estimate the heterogeneous distributed lag effects. Our model allows for multiple modifiers and performs modifier selection to identify those that contribute the most to the heterogeneity in exposure effects. The tree triplet structure uses temporal constraints to account for the time-varying effect of longitudinally assessed exposure. Additionally, our model can incorporate mixture exposures while performing component selection and assuming time-sensitive interactions between components.

We applied our model to Colorado birth registry data to examine the association between the birth weight and $PM_{2.5}$ and daily maximal temperature. We found that maternal age, BMI, Hispanic designation, and race contributed the most to heterogeneity in the exposure effects. We estimated that the exposure effect of $PM_{2.5}$ was larger for mothers younger than 25 and for those with a BMI between 23.01 and 34.13. $PM_{2.5}$ also had a more pronounced negative effect on non-Hispanic Asian, Pacific Islanders, and White mothers with longer windows of susceptibility in the early and late stages of pregnancy. We estimated the effects of temperature to be around zero but found different trends across subgroups. Our findings demonstrate that the presence of heterogeneity can lead to substantial differences in exposure effect estimates.

The HDLMM framework is a novel data-driven approach for precision environmental health that can simultaneously identify who is the most vulnerable, when they are most susceptible, and what components of environmental mixture they are most susceptible to. In contrast to the stratified approaches of previous studies, our model offers a more comprehensive understanding of interactions between modifiers and mixture exposures. Our framework allows researchers to formulate estimates to address a wide variety of research questions, such as subgroup-specific effects and the effects of changes in exposure and modifier levels. With the emergence of a precision environmental health framework, the proposed method will be a versatile statistical tool to elucidate the complex underlying mechanisms of environmental exposures and inform targeted interventions to improve health and prevent disease.

Chapter 4

Structured Bayesian Regression Tree Models for Estimating Distributed Lag Effects: The R Package dlmtree

4.1 Introduction

In many fields, there is interest in estimating the lagged relationship between an exposure (or treatment) and an outcome. In such cases, the length of the lag or how the exposure effect varies is often unknown. The lagged relationship can take either of two forms. The first form is the association between the exposure and an outcome concurrently and distributed across several subsequent times. For example, the impact of advertisement persists beyond that single time point of the investment (Koyck, 1954; Palda, 1965) or exposure to an environmental pollutant affects mortality on the same day and each of the following days (Schwartz, 2000). The second form is an exposure assessed longitudinally and an outcome assessed post-exposure. Examples of this form are maternal exposure to environmental chemicals during pregnancy on birth outcomes and children's health and development (Chiu et al., 2023; Hsu et al., 2023) and the effect of training and recovery activities over multiple days on athlete wellness (Schliep et al., 2021). A popular statistical method to estimate the time-varying association between exposure and outcome is a distributed lag model (DLM).

The DLM regresses a scalar outcome on the exposure measured at preceding time points (Schwartz, 2000; Gasparrini et al., 2010). The DLM framework is particularly useful as it allows for estimating time-resolved exposure effects and quantifies the temporal relationship between the exposure and the outcome. Because the repeated measurements of exposure are often correlated, DLMs typically include a smoothing constraint to add temporal structure to the estimated time-specific effects and to regularize the exposure effects in the presence of multicollinearity across

the measurements. Constraints include polynomials, splines, and Gaussian processes (Zanobetti et al., 2000; Warren et al., 2012). Notably, Mork and Wilson (2023) introduced an approach for constrained DLM using regression tree structures based on the Bayesian additive regression tree (BART) framework (Chipman et al., 2010). More recent methods have extended the tree structured DLM framework in several directions, including nonlinear exposure-response relationship (Mork and Wilson, 2022), models with lagged interaction among multiple exposures (Mork and Wilson, 2023), and models with heterogeneous lag effects (Mork et al., 2023).

Table 4.1: Available DLM related R packages with functionalities

Package	GLM	Nonlinearity	Mixture	Heterogeneity
dlnm	✓	✓		
bdlim	✓			✓
dlim	✓			✓
dlmtree	✓	✓	✓	✓

There are several related methods and packages for DLM implementation. Table 4.1 lists currently available R packages and their offered features. The **dlnm** package contains software for estimating a DLM with linear or nonlinear exposure-response relationships using splines for the smoothing constraint (Gasparrini and Armstrong, 2013). Mork and Wilson (2022, 2023) compare model performance between the spline-based and tree structured DLM implementations. The **dlnm** package does not consider heterogeneity or multiple exposures. There are several DLM methods for the linear effect of a single exposure with modification by a single covariate. The **dlim** package allows for modification by a single continuous factor (Dematteis et al., 2024) and the **bdlim** package allows for modification by a single categorical factor (Wilson et al., 2017a). However, these packages do not allow for multiple candidate modifying factors or multiple exposures. Hence, the **dlmtree** package fills several gaps including heterogeneous effects of multiple candidate modifiers and analysis of mixture or multivariate lagged exposures.

In this chapter, we introduce a comprehensive R package **dlmtree** which consolidates a wide range of tree structured DLMs. The package offers a user-friendly and computationally efficient

environment to fit tree structured DLMs and to address multiple potential research assumptions including nonlinearity, monotonicity and prior information, multiple simultaneous exposures, and heterogeneous lag effects. We first provide a conceptual review of a regression tree as a smoothing constraint in a DLM framework and an overview of the extensions of tree structured DLMs. We present a decision tree to help users select an appropriate model for their analysis. We illustrate the model fitting process with detailed descriptions and syntax of functions for implementing models, obtaining the summary output and inferential information, and plotting the fitted models for visualization. The package is available in the comprehensive R archive network (CRAN) and the installation instructions are provided at https://danielmork.github.io/dlmtree/.

4.2 Tree structured DLMs

4.2.1 DLM tree as a smoothing constraint

In this section, we review the regression tree approach to constrained DLM estimation, a key idea underlying the tree structured DLMs in the **dlmtree** package. Consider a vector of outcomes $\boldsymbol{y} = (y_1, \ldots, y_n)$ for a sample $i = 1, \ldots, n$. For clarity, we assume continuous outcomes when presenting the models. We discuss extensions to generalized linear models in Section 4.2.4. Suppose we are interested in the lagged association between the outcome and a single longitudinally assessed exposure $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iT})$ measured at equally spaced time points, $t = 1, \ldots, T$. The typical DLM is

$$y_i = f(\boldsymbol{x}_i) + \boldsymbol{z}_i'\boldsymbol{\gamma} + \varepsilon_i, \quad f(\boldsymbol{x}_i) = \sum_{t=1}^{T} x_{it}\theta_t, \tag{4.1}$$

where $\boldsymbol{z}_i$ is a vector of covariates including the intercept, $\boldsymbol{\gamma}$ is a vector of regression coefficients for covariates, and f is a distributed lag function parameterized by θ_t representing the linear effect of exposure at time t. The DLM framework in (4.1) assumes a linear association between the outcome and the exposure at multiple time points.

The estimation of θ_t for $t = 1, \ldots, T$ in (4.1) requires an appropriate temporal structure, such as a smooth or piecewise-smooth constraint on θ_t, to account for autocorrelation within the exposure

measurements. Mork and Wilson (2023) introduced a regression tree-structure, shown in Figure 4.1, referred to as a *DLM tree*. A DLM tree, denoted $\mathcal{T}$, is a binary tree that splits the exposure time

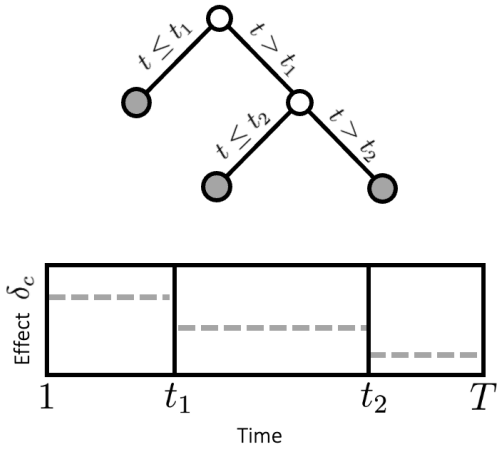


Figure 4.1: A DLM tree, $\mathcal{T}$. A binary tree splits the time span into non-overlapping intervals, here resulting in three terminal nodes representing three time segments (gray nodes). Each terminal node is assigned a constant effect (gray dashed lines).

span into non-overlapping segments. We denote the terminal nodes of $\mathcal{T}$ as λ_c for $c = 1, \ldots, C$ where C is the total number of terminal nodes. Each terminal node of the DLM tree is assigned a scalar parameter δ_c that represents the effect of the time lags contained in that terminal node. We denote a set of scalar parameters of the DLM tree $\mathcal{T}$ as $\mathcal{D} = \{\delta_1, \ldots, \delta_C\}$. A DLM tree defines a function of a time lag,

$$g(t|\mathcal{T}, \mathcal{D}) = \delta_c \quad \text{if } t \in \lambda_c. \tag{4.2}$$

The treed distributed lag model (TDLM) is a tree structured DLM in its simplest form and is a Bayesian additive model that consists of an ensemble of A DLM trees, indexed as $\mathcal{T}_a$, with a corresponding set of scalar parameters $\mathcal{D}_a$ for $a = 1, \ldots, A$. TDLM defines θ_t in (4.1) as

$$\theta_t = \sum_{a=1}^{A} g(t|\mathcal{T}_a, \mathcal{D}_a). \tag{4.3}$$

The representation of the DLM tree ensemble provides several advantages. Each DLM tree provides a temporal structure on the exposure-time-response function with data-driven learning of the

change points and time spans related to the outcome. The ensemble structure allows for flexibility to approximate smoothness in the exposure-time-response function as each DLM tree in the ensemble splits the time span differently.

4.2.2 DLM tree pair for lagged multivariate exposures

A distributed lag mixture model extends the DLM framework to a multivariate exposure (also referred to as *mixture exposures* in the environmental literature) assessed longitudinally. For mixture exposures with $M \geq 2$ exposures, we denote the vector of longitudinally assessed measurements of the m^{th} exposure, for $m = 1, \ldots, M$, as $\boldsymbol{x}_{im} = (x_{im1}, \ldots, x_{imT})$. The exposure-time-response function $f(\boldsymbol{x}_i)$ in (4.1) is replaced with a mixture-exposure-time-response function that incorporates the main effect of each exposure and the pairwise lagged interaction effects between exposures. The mixture-exposure-time-response function is

$$f(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iM}) = \sum_{m=1}^{M} \sum_{t=1}^{T} x_{imt}\theta_{mt} + \sum_{m_1=1}^{M} \sum_{m_2=m_1}^{M} \sum_{t_1=1}^{T} \sum_{t_2=1}^{T} x_{im_1t_1} x_{im_2t_2} \theta_{m_1m_2t_1t_2}, \tag{4.4}$$

where θ_{mt} is a main effect of exposure m at time t and $\theta_{m_1m_2t_1t_2}$ is an interaction effect of exposure m_1 at time t_1 and exposure m_2 at time t_2.

Estimating the function in (4.4) is challenging due to autocorrelation across repeated exposure measurements, correlation at the same time lag between exposures, and a high-dimensional parameter space. Mork and Wilson (2023) introduced the treed distributed lag mixture model (TDLMM) that structures the parameters in (4.4) using a *DLM tree pair*. Figure 4.2 illustrates a single DLM tree pair, denoted $\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_1 \times \mathcal{T}_2\}$ with corresponding parameters sets $\{\mathcal{D}_1, \mathcal{D}_2, \Omega\}$, where $\mathcal{D}_1$ and $\mathcal{D}_2$ represent the scalar main effects for $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively, and Ω contains the scalar interaction effects for interaction surface, denoted $\mathcal{T}_1 \times \mathcal{T}_2$. Each DLM tree in the pair is associated with one component of the mixture exposures. Similar to the DLM tree described above, each DLM tree in the pair partitions the time span of its assigned component. Each DLM tree in a tree pair is

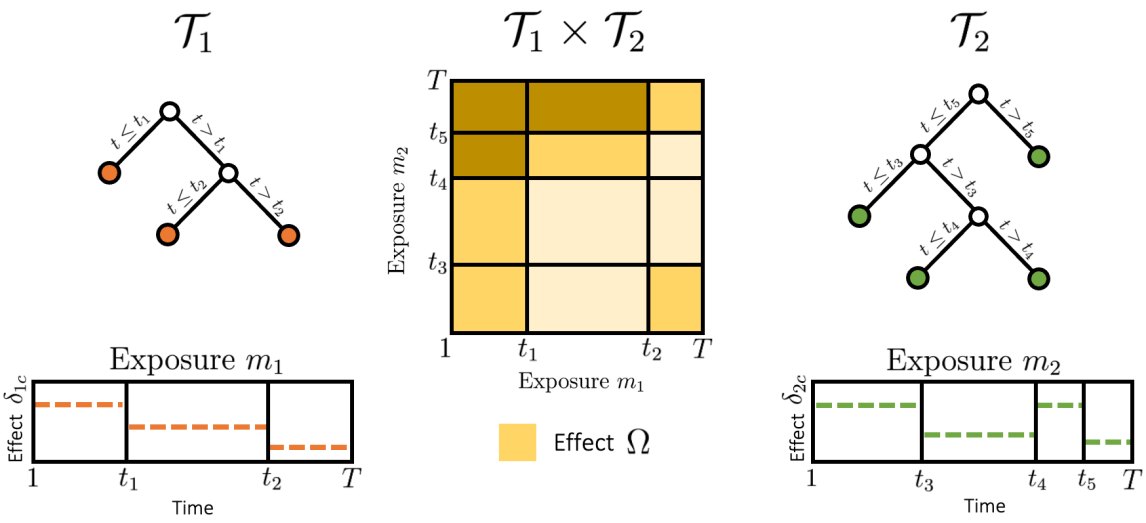


Figure 4.2: A DLM tree pair. Two DLM trees split the time span of the assigned exposure into non-overlapping intervals, here resulting in three time segments for exposure m_1 and four time segments for exposure m_2 (colored nodes). Each terminal node is assigned a scalar parameter that represents a constant effect of the assigned exposure (colored dashed lines). The interaction surface is fully defined by two DLM trees and each combination of time segments is assigned a scalar parameter (color-shaded boxes).

defined similarly to (4.2) such that

$$g(t|\mathcal{T}_p, \mathcal{D}_p) = \delta_{pc} \quad \text{if } t \in \lambda_{pc}, \quad p = 1, 2. \tag{4.5}$$

The DLM tree pair structure allows for an interaction surface to model lagged interactions between exposures. The interaction surface $\mathcal{T}_1 \times \mathcal{T}_2$ shown in the middle of Figure 4.2, is fully defined by the two DLM trees in a pair. Each time interval of the first DLM tree is paired with every interval of the second DLM tree. Each combination is assigned a scalar parameter $\omega_{c_1 c_2}$ that represents the lagged interaction effect where c_1 and c_2 are indices for terminal nodes of the first and second DLM tree, respectively. The interaction surface of a DLM tree pair defines a function

$$g_I(t_1, t_2|\mathcal{T}_1 \times \mathcal{T}_2, \Omega) = \omega_{c_1 c_2} \quad \text{if } t_1 \in \lambda_{c_1}, t_2 \in \lambda_{c_2}. \tag{4.6}$$

A DLM tree pair with an interaction surface provides a structure to regularize the exposure-time-response function for two components and their lagged interaction effects.

TDLMM employs an ensemble representation of A DLM tree pairs denoted $\{\mathcal{T}_{a1}, \mathcal{T}_{a2}, \mathcal{T}_{a1} \times \mathcal{T}_{a2}\}$ with the corresponding set of scalar parameters $\{\mathcal{D}_{a1}, \mathcal{D}_{a2}, \Omega_a\}$, using the formulation in (4.5) and (4.6). Each DLM tree $\mathcal{T}_{ap}$ is also associated with exposure m, indicated by $S_{ap} = m$. With TDLMM, the main effect of exposure m at time t in (4.4) is

$$\theta_{mt} = \sum_{a=1}^{A} \sum_{p=1}^{2} g(t|\mathcal{T}_{ap}, \mathcal{D}_{ap}) \mathbb{I}(S_{ap} = m). \tag{4.7}$$

The pairwise interaction effect between exposure m_1 at time t_1 and exposure m_2 at time t_2 in (4.4) is

$$\theta_{m_1 m_2 t_1 t_2} = \sum_{a=1}^{A} g_I(t_1, t_2|\mathcal{T}_{a1} \times \mathcal{T}_{a2}, \Omega_a) \mathbb{I}(S_{a1} = m_1, S_{a2} = m_2). \tag{4.8}$$

TDLMM has three representations depending on different assumptions of lagged interactions between exposures. The simplest form, TDLMMadd, assumes $\theta_{m_1 m_2 t_1 t_2} = 0$ in (4.4), implying no interaction between exposures resulting in an additive model of the main effects of exposures. The second form, TDLMMns, accounts for lagged interaction between exposures but not within exposures. The last, TDLMMall, allows for all lagged interactions between and within exposures, implying a nonlinear effect of exposure.

4.2.3 Extensions to nonlinear exposure-time-response functions

Mork and Wilson (2022) introduced the treed distributed lag nonlinear model (TDLNM) to estimate the nonlinear association between a single longitudinally assessed exposure and an outcome. The DLM tree, as shown in (4.2), assumes a linear association between exposure at each time point and the outcome. To relax the linearity assumption for a distributed lag nonlinear model framework, TDLNM modifies the DLM tree to additionally split exposure concentration levels along with the exposure time span, partitioning the bi-dimensional space of exposure concentration and time lags. Mork and Wilson (2024) further extended TDLNM to include a monotonicity assumption, where the exposure-response is constrained to be non-decreasing at each time point. See Mork and Wilson (2022, 2024) for details.

4.2.4 Extensions to generalized linear models

In various applications of DLMs, the response variables may be binary or counts. Examples of binary outcomes include the occurrence of conditions such as asthma or preterm birth and an example of a count-valued outcome is the daily number of deaths in a county. The linear representation of the DLM framework allows the tree structured DLMs to extend to the generalized linear model setting. TDLM, TDLNM, and TDLMM have been extended to incorporate binary response variables (Mork and Wilson, 2022, 2023). Further extensions on these models allow for count data that may be zero-inflated or overdispersed. The extensions to binary and count data rely on a framework based on the Pólya-Gamma data augmentation approach (Polson et al., 2013; Neelon, 2019).

4.2.5 Extensions to heterogeneous models

Another extension to the DLM framework is to assume heterogeneous exposure effects. The exposure effects may be heterogeneous due to a single modifying factor or a set of factors. For example, the impact of prenatal exposure to air pollution may be governed by genetic factors such as fetal sex (Rosa et al., 2019a). The set of factors, referred to as *modifiers*, may be continuous, categorical, or ordinal. The general approach is to introduce an additional tree, known as a *modifier tree*, that partitions the modifier space and has a DLM tree or a DLM tree pair affixed to each terminal node. Mork et al. (2023) extended TDLM to the heterogeneous distributed lag model (HDLM) by introducing a nested tree structure, as shown in Figure 4.3. In a nested tree, a modifier tree is applied to a set of candidate modifiers, defining mutually exclusive subgroups of the sample at each terminal node. A DLM tree is then attached to each terminal node of the modifier tree to define subgroup-specific effects with unique parameters for each subgroup. Other extensions include a heterogeneous distributed lag mixture model (HDLMM) that extends TDLMM to incorporate heterogeneity based on a set of multiple candidate modifiers using an ensemble of tree triplet structures.

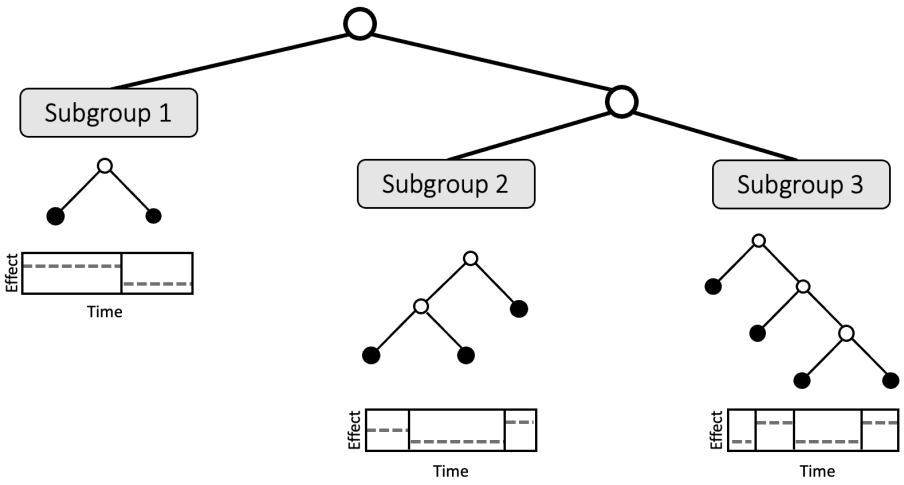


Figure 4.3: A nested tree structure for heterogeneous tree structured DLMs. The top tree is a modifier tree that is applied to candidate modifiers, here resulting in three subgroups. DLM trees are affixed to the terminal nodes of a modifier tree to estimate the exposure-time-response relationship specific to the subgroups.

4.3 Implementation

The tree structured DLMs are classified with three main criteria: 1) Linear or nonlinear exposure-time-response function, 2) One lagged component or a mixture of more than one lagged component, and 3) Homogeneous or heterogeneous exposure-time-response relationship. Figure 4.4 illustrates a decision tree as a guide to choosing an appropriate tree structured DLM.

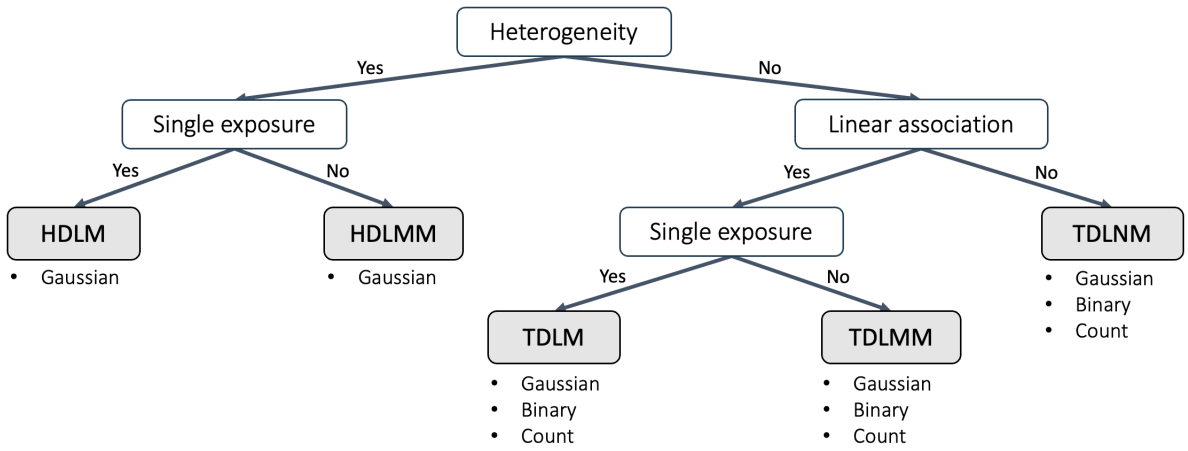


Figure 4.4: A decision tree for choosing tree structured DLMs. The bullet points below the models list the data types of response variables that each model can incorporate.

The **dlmtree** package offers tree structured DLMs shown in Figure 4.4. A main function `dlmtree` is designed to fit various tree structured DLMs, offering customizable analysis through its arguments. The function `dlmtree` with its main arguments is as follows.

```
dlmtree(formula, data, exposure.data, family, dlm.type, mixture, het, ...)
```

The function requires a formula specifying an outcome and covariates for the fixed effect, a data source, the data type of response variable, and three arguments for model specification. The data for the `formula`, including covariates and the outcome, are provided as an $(n \times p)$ data frame in the `data` argument. The exposures are not specified in the `formula` and are not needed in the `data` argument. Rather, the exposure data is specified separately as an $(n \times T)$ matrix of exposure measurements or a list of $(n \times T)$ matrices of exposure measurements in the `exposure.data` argument.

Table 4.2: Arguments for `dlmtree` function

Argument	Description
`formula`	Object of class `formula` for the fixed effect
`data`	A data frame containing covariates and an outcome used in `formula`
`exposure.data`	A numerical matrix of exposure data with the same length as data. For a mixture setting, a named list containing equally sized numerical matrices of exposure data having the same length as the data
`family`	'`gaussian`' for a continuous response, '`logit`' for binomial, '`zinb`' for count data
`dlm.type`	DLM type specification: '`linear`', '`nonlinear`', '`monotone`'
`mixture`	logical; A flag for mixture exposures if `TRUE`
`het`	logical; A flag for heterogeneity if `TRUE`

Table 4.2 provides descriptions of the arguments. All tree structured DLMs allow for continuous response variables, while TDLM, TDLNM, and TDLMM additionally allow for binary and count-valued response variables. Additional parameters, omitted here, include MCMC sampling parameters, the number of trees in the ensemble, effect shrinkage, and sparsity parameters for exposure and modifier selection. Model-specific hyperparameters, prefixed by the model name, can also be fine-tuned.

A fitted model is assigned a class, determined by the model-specifying arguments. The classes are: `tdlm`, `tdlmm`, `tdlnm`, `hdlm`, and `hdlmm`. Each class uses an S3 object oriented system with `summary` method. The `summary` method returns the model information, estimates and credible intervals for the fixed effects and lag effects, and time points of a significant effect, where the 95% credible intervals of the lag effects do not contain zero (in some applications this is referred to as *critical window*) of the exposure or mixture exposures. The `plot` method on the summary object further returns the visualization of the estimated main exposure effects (and interaction effects if applicable) with 95% credible intervals. Additionally for classes of models with heterogeneity: `hdlm` and `hdlmm`, the `shiny` method is built for various types of statistical inference. The **shiny** app interface provides tools for identifying important modifiers and their splitting points that contribute to heterogeneity. It also includes features for evaluating personalized exposure effects with a set of user-specified modifiers, and exposure effects specific to a subgroup defined by a set of modifiers of interest.

4.4 Example usage

We illustrate the example usage of tree structured DLMs through a set of vignettes based on simulated data. We demonstrate data preparation and the model fitting process for TDLM, TDLMM, HDLM, and HDLMM. An example of fitting TDLNM is provided in Appendix C.2.

4.4.1 Simulated dataset

We use the simulated dataset '`sbd_dlmtree`', which is publicly available in the **dlmtree** package GitHub repository. The dataset contains 10,000 simulated mother-child dyads with descriptions of maternal and birth information. The maternal covariates include maternal age, height, prior weight, prior body mass index (BMI), race, Hispanic designation, education attainment, smoking habits, marital status, and yearly income. Additionally, the birth information includes birth weight for gestational age z-scores (BWGAZ), gestational age, sex of a child, and estimated date of conception. The dataset contains five environmental chemicals measured at 37 weeks preceding the birth for

each dyad: fine particulate matter ($PM_{2.5}$), temperature, sulfur dioxide (SO_2), carbon monoxide (CO), and nitrogen dioxide (NO_2). Each exposure measurement is scaled by its interquartile range value. The dataset is constructed to have realistic distributions and correlations of the covariates and exposures. It contains a complex exposure-response relationship that includes interactions and heterogeneous effects.

In the following examples, we examine the distributed lag effects of maternal exposure to the environmental mixtures during the 37 weeks of gestation on BWGAZ. The results provided in the example usage are solely for demonstrative purposes of the model fitting process using the model parameters and simulated data and do not represent any actual findings.

4.4.2 Data preparation for model fitting

We first load the required packages: **dlmtree** and **dplyr**. We used **dplyr** for a better presentation of the data processing. We set the seed for reproducibility and load the external dataset 'sbd_dlmtree' from **dlmtree** package repository using an embedded function `get_sbd_dlmtree`.

```
# Libraries and seed
library(dlmtree)
library(dplyr)
set.seed(1)

# Download data as 'sbd'
sbd <- get_sbd_dlmtree()
```

The data frame `sbd` has a lagged format where each row has columns of lagged measurements of each exposure (e.g. wide format data). Often, datasets have a time-series format, which contains a single column of dates and multiple columns of corresponding measurements of response variables and exposures on that specific date only. This is particularly common in time-series studies, whereas the wide format data is more common in cohort studies. The model fitting function `dlmtree` is designed to use a data frame in a wide format hence a data frame with a time-series format must be pivoted to a wide format.

To prepare the birth data 'sbd' for model fitting, we create a data frame including the covariates and the response variable, BWGAZ. We note that the categorical columns in the dataset are of class

`factor`. We consider five components for exposure data: $PM_{2.5}$, temperature, SO_2, CO, and NO_2, and store them as a list of exposure matrices.

```
# Response and covariates
sbd_cov <- sbd %>% select(bwgaz, ChildSex, MomAge, GestAge, MomPriorBMI, Race,
                          Hispanic, MomEdu, SmkAny, Marital, Income,
                          EstDateConcept, EstMonthConcept, EstYearConcept)

# Exposure data
sbd_exp <- list(PM25 = sbd %>% select(starts_with("pm25_")),
                TEMP = sbd %>% select(starts_with("temp_")),
                SO2 = sbd %>% select(starts_with("so2_")),
                CO = sbd %>% select(starts_with("co_")),
                NO2 = sbd %>% select(starts_with("no2_")))

sbd_exp <- sbd_exp %>% lapply(as.matrix)
```

Each matrix of the exposure data list can be centered and scaled with caution of different interpretations of the resulting estimates. Specifically when using the TDLMM with lagged interaction, it is crucial to avoid centering the exposure data as it can lead to an inaccurate estimate of the marginal exposure effect when considering co-exposures.

4.4.3 TDLM: Linear relationship between an outcome and a single exposure

We first assume that we are interested in the linear association between BWGAZ and weekly exposure to $PM_{2.5}$ during the first 37 gestational weeks. We include the following covariates to control for fixed effects: child sex, maternal age, BMI, race, Hispanic designation, smoking habits, and month of conception. We fit TDLM with the following code.

```
tdlm.fit <- dlmtree(formula = bwgaz ~ ChildSex + MomAge + MomPriorBMI +
                                      Race + Hispanic + SmkAny + EstMonthConcept,
                    data = sbd_cov,
                    exposure.data = sbd_exp[["PM25"]],    # A single numeric matrix
                    family = "gaussian", dlm.type = "linear",
                    n.burn = 2500, n.iter = 10000, n.thin = 5)
```

The resulting fitted object of class `tdlm` has attributes of the fitted model information and posterior samples of parameters of interest. The `summary` method applied to the object `tdlm.fit` returns a clear overview of the model fit. The summary of the model is obtained with the following code.

```
tdlm.sum <- summary(tdlm.fit)
tdlm.sum

---
TDLM summary

Model run info:
- bwgaz ~ ChildSex + MomAge + MomPriorBMI + Race + Hispanic + SmkAny + EstMonthConcept
- sample size: 10,000
- family: gaussian
- 20 trees
- 2500 burn-in iterations
- 10000 post-burn iterations
- 5 thinning factor
- 0.95 confidence level

Fixed effect coefficients:
                       Mean  Lower  Upper
*(Intercept)          2.289  2.032  2.542
*ChildSexM           -2.105 -2.126 -2.085
MomAge                0.000 -0.001  0.002
*MomPriorBMI         -0.021 -0.022 -0.019
RaceAsianPI           0.069 -0.057  0.192
RaceBlack             0.078 -0.050  0.205
Racewhite             0.059 -0.060  0.181
*HispanicNonHispanic  0.255  0.233  0.278
*SmkAnyY             -0.403 -0.451 -0.356
EstMonthConcept2     -0.049 -0.109  0.010
*EstMonthConcept3    -0.145 -0.211 -0.077
*EstMonthConcept4    -0.230 -0.295 -0.160
*EstMonthConcept5    -0.207 -0.265 -0.147
*EstMonthConcept6    -0.205 -0.260 -0.153
EstMonthConcept7     -0.032 -0.083  0.023
*EstMonthConcept8     0.145  0.081  0.210
*EstMonthConcept9     0.393  0.326  0.460
*EstMonthConcept10    0.372  0.311  0.437
*EstMonthConcept11    0.330  0.271  0.387
*EstMonthConcept12    0.129  0.078  0.181
---
* = CI does not contain zero

DLM effect:
range = [-0.019, 0.008]
signal-to-noise = 0.021
critical windows: 11-20,36-37
```

```
              Mean  Lower  Upper
Period 1     0.003 -0.005  0.015
Period 2     0.000 -0.006  0.010
Period 3    -0.002 -0.010  0.004
Period 4    -0.002 -0.010  0.003
Period 5    -0.001 -0.007  0.004
Period 6    -0.001 -0.006  0.005
Period 7    -0.001 -0.006  0.006
Period 8    -0.001 -0.008  0.004
Period 9    -0.002 -0.011  0.003
Period 10   -0.002 -0.012  0.006
*Period 11 -0.016 -0.024 -0.007
*Period 12 -0.017 -0.024 -0.010
*Period 13 -0.017 -0.024 -0.012
*Period 14 -0.017 -0.022 -0.010
*Period 15 -0.017 -0.022 -0.011
*Period 16 -0.017 -0.022 -0.011
*Period 17 -0.017 -0.024 -0.011
*Period 18 -0.019 -0.030 -0.013
*Period 19 -0.018 -0.027 -0.011
*Period 20 -0.015 -0.024 -0.003
Period 21   -0.007 -0.019  0.002
Period 22   -0.002 -0.010  0.006
Period 23   -0.003 -0.012  0.003
Period 24   -0.002 -0.008  0.004
Period 25    0.000 -0.005  0.006
Period 26    0.000 -0.005  0.006
Period 27   -0.001 -0.005  0.005
Period 28   -0.001 -0.006  0.004
Period 29   -0.001 -0.007  0.004
Period 30   -0.002 -0.008  0.003
Period 31   -0.002 -0.008  0.004
Period 32   -0.001 -0.008  0.005
Period 33    0.002 -0.004  0.010
Period 34    0.004 -0.003  0.012
Period 35    0.006 -0.001  0.014
*Period 36  0.008  0.000  0.017
*Period 37  0.008  0.000  0.019
---
* = CI does not contain zero

residual standard errors: 0.004
---
```

The summary output first presents a section '`Model run info`' with the model fitting information including the formula, sample size, data type of the response variable, number of trees in the ensemble, MCMC parameters, and a confidence level. The next section, '`Fixed effect coefficients`', shows the estimates with credible intervals for the regression coefficients of the covariates. Lastly, the summary output returns '`DLM effect`' section including the range of DLM effects, signal-to-noise ratio, and estimated lagged effects with credible intervals. Each lag is marked with an asterisk if it is identified as a critical window based on a pointwise 0.95 probability credible interval. In context, TDLM estimated a negative association between BWGAZ and $PM_{2.5}$ with gestational weeks 11–20 as critical windows.

The cumulative effect of the exposure is defined as the effect of a one unit increment of the exposure across all time points. The following code returns the estimated cumulative effect with its 95% credible interval from an attribute `cumulative.effect` of the summary object `tdlm.sum`.

```
tdlm.sum$cumulative.effect
      mean       2.5%      97.5%
-0.1738753 -0.2124918 -0.1367665
```

In this context, TDLM estimates that a unit increment of exposure to $PM_{2.5}$ across all gestational weeks has a cumulative exposure effect of -0.17 (95% CrI: [-0.21, -0.14]) on BWGAZ.

For a more intuitive view of the overall trend of the distributed lag effects, the `plot` method may be applied to the summary object `tdlm.sum` to visualize the effect estimates as the following.

```
plot(tdlm.sum, main = "Estimated effect of PM2.5", xlab = "Time", ylab = "Effect")
```

Figure 4.5 shows the plot of the estimated exposure effect of $PM_{2.5}$ where the x-axis is the lags (or weeks) and the y-axis is the estimated exposure effect. The gray area showing the 95% credible intervals of exposure effects during weeks 11–20 does not cover the red line of a null effect, which indicates a critical window. The `plot` method also includes additional arguments of `main`, `xlab`, and `ylab` for customizing the main title, x-axis, and y-axis label. For a customized plot, the estimated values can be obtained from the attributes of the summary object `tdlm.sum`: `matfit`, `cilower`, and `ciupper`.

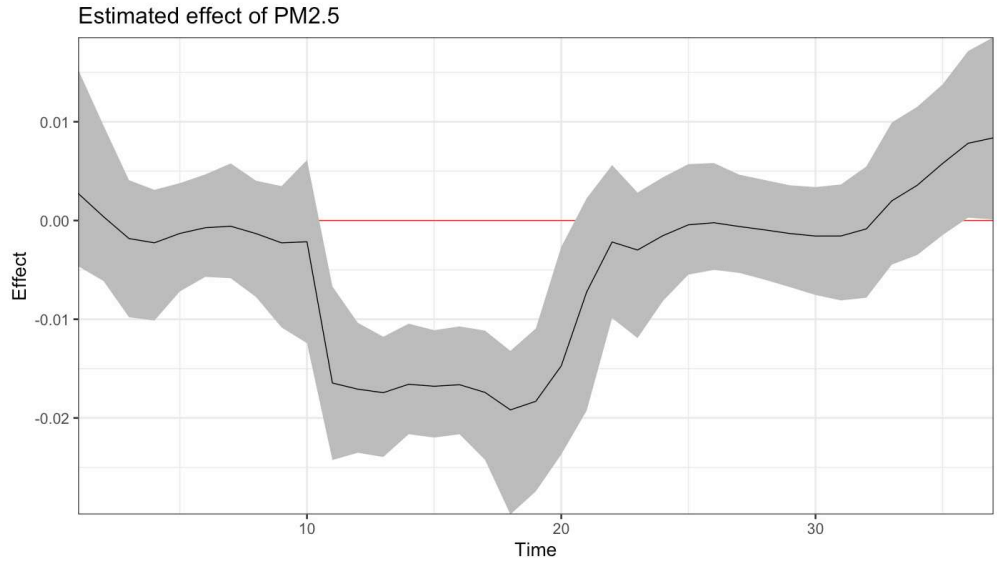


Figure 4.5: Estimated distributed lag effects of $PM_{2.5}$ on BWGAZ during 37 gestational weeks, using TDLM.

4.4.4 TDLMM: Linear relationship between an outcome and a mixture

Suppose we are interested in the linear association between BWGAZ and mixture exposures of five exposures: $PM_{2.5}$, temperature, SO_2, CO, and NO_2. We are mainly interested in the marginal effects of each exposure on BWGAZ, the pairwise lagged interaction between exposures, and which exposures are most correlated with BWGAZ. The key differences between the code for the model for multiple exposures below and the code for the previous single exposure model are that the `exposure.data` is provided with a list of matrices of five different exposures and we specify `mixture = TRUE`. The code is as follows.

```
tdlmm.fit <- dlmtree(formula = bwgaz ~ ChildSex + MomAge + MomPriorBMI +
                                       Race + Hispanic + SmkAny + EstMonthConcept,
                     data = sbd_cov,
                     exposure.data = sbd_exp,
                     family = "gaussian", dlm.type = "linear", mixture = TRUE,
                     mixture.interactions = "noself",
                     n.burn = 2500, n.iter = 10000, n.thin = 5)
```

The model assumption regarding lagged interaction can be additionally specified for the TDLMM fitting with an argument `mixture.interactions`. The options include no interaction (‘none’), no-

self interactions ('noself'), and all interactions ('all'). We use TDLMM with no-self lagged interaction, which is the default argument.

The summary method can be applied to the fitted object tdlmm.fit. As TDLMM allows for pairwise lagged interaction between mixture exposures, the estimated exposure effect for each exposure varies with the levels of the co-exposures. The summary method on class tdlmm offers additional control argument marginalize to address this. The argument marginalize requires a fixed level used for co-exposure marginalization. The default is the empirical means of co-exposures, which provides the distributed lag function for a single exposure estimated when all other exposures are fixed at their means. This is equivalent to integrating out all co-exposures. The following code returns the summary of the tdlmm.fit with marginalization using the empirical means of co-exposures.

```
# Marginalization with co-exposure fixed at the empirical means
tdlmm.sum <- summary(tdlmm.fit, marginalize = "mean")
```

Other marginalization options are available for conducting different types of inferences. The second option is to specify a number between 0 and 100, representing a percentile of the co-exposures that will be used for marginalization. The following code returns the summary of the tdlmm.fit with marginalization fixing all co-exposures at their 25th percentile values.

```
# Marginalization with co-exposure fixed at 25th percentile
tdlmm.sum.percentile <- summary(tdlmm.fit, marginalize = 25)
```

The last option is to specify the exact levels of co-exposures. This option requires the argument marginalize to be a numeric vector of the same length as the number of exposures used for the model fitting. The specified values in the vector must also follow the order of the exposures in the fitted model. This method of marginalization offers flexibility as these exposure levels can be specified based on pre-informed levels using existing data or to address hypothetical questions. For example, the following code can be used to obtain the marginal exposure effect of $PM_{2.5}$ when temperature, SO_2, CO, and NO_2 are all fixed to 1. The marginalized effects of other exposures are calculated in a similar manner.

```
# Marginalization with co-exposure fixed at exact levels for each exposure
tdlmm.sum.level <- summary(tdlmm.fit, marginalize = c(1, 1, 1, 1, 1))
```

Below is the summary result of `tdlmm.sum` with the default argument using empirical means.

```
tdlmm.sum
---
TDLMM summary

Model run info:
- bwgaz ~ ChildSex + MomAge + MomPriorBMI + Race + Hispanic + SmkAny + EstMonthConcept
- sample size: 10,000
- family: gaussian
- 20 trees (alpha = 0.95, beta = 2)
- 2500 burn-in iterations
- 10000 post-burn iterations
- 5 thinning factor
- 5 exposures measured at 37 time points
- 10 two-way interactions (no-self interactions)
- 1 kappa sparsity prior
- 0.95 confidence level

Fixed effects:
                       Mean  Lower  Upper
*(Intercept)          0.172  0.043  0.307
*ChildSexM           -2.063 -2.085 -2.041
 MomAge               0.001 -0.001  0.002
*MomPriorBMI         -0.020 -0.022 -0.019
 RaceAsianPI          0.027 -0.058  0.117
 RaceBlack            0.033 -0.063  0.124
 Racewhite            0.016 -0.067  0.100
*HispanicNonHispanic  0.248  0.224  0.272
*SmkAnyY             -0.393 -0.441 -0.346
 EstMonthConcept2     0.073 -0.003  0.145
*EstMonthConcept3     0.107  0.009  0.211
*EstMonthConcept4     0.158  0.038  0.282
*EstMonthConcept5     0.255  0.126  0.388
*EstMonthConcept6     0.200  0.064  0.333
*EstMonthConcept7     0.223  0.084  0.354
*EstMonthConcept8     0.199  0.068  0.331
*EstMonthConcept9     0.291  0.164  0.418
*EstMonthConcept10    0.182  0.070  0.296
*EstMonthConcept11    0.135  0.040  0.236
 EstMonthConcept12    0.006 -0.062  0.077
---
```

```
* = CI does not contain zero

--
Exposure effects: critical windows
* = Exposure selected by Bayes Factor
(x.xx) = Relative effect size

 *PM25 (0.7): 11-20
 *TEMP (0.7): 5-19
 *SO2 (0.21):
 *CO (0.63):
 *NO2 (0.26): 23
--
Interaction effects: critical windows

 PM25/TEMP (0.8):
 12/6-19
 13/6-19
 14/6-20
 15/6-20
 16/6-20
 17/6-21
 18/5-22
 19/5-22
 20/6-21
---
residual standard errors: 0.005
```

The summary output of the TDLMM fit contains similar information to that of the TDLM. The '`Model run info`' section in the output includes additional information specific to mixture exposures: the number of exposures included in the model, the number of pairwise interactions, and a sparsity parameter for exposure selection. The summary output does not include the lagged effects for each exposure to prevent overwhelming the output with excessive information. The summary presents the critical window of the marginal effects of each exposure with the relative effect size, indicating the effect size of exposure relative to that of other exposures. The summary also returns the critical window of lagged interaction effects with relative effect size. In our context, all five exposures are considered to be significantly associated with BWGAZ, based on a Bayes factor threshold of 0.5. The TDLMM estimated gestational weeks 11–20, 5–19, and 23 as critical

windows of $PM_{2.5}$, temperature, and NO_2, respectively. The model fit also identified significant $PM_{2.5}$–temperature lagged interaction effects.

More useful statistical inferences are possible with `tdlmm.fit` and `tdlmm.sum`. First, a function `adj_coexposure` can be used to obtain the marginalized exposure effect while accounting for the expected change in co-exposures. In comparison to using the argument `marginalize` in `summary` method, the function `adj_coexposure` uses spline-based method to predict the expected changes in co-exposures corresponding with a pre-defined change in an exposure of interest and calculates the marginalized effects of the primary exposure with co-exposure at the predicted levels. The following code shows an example usage of the function with its output omitted.

```
# Lower and upper exposure levels specified as 25th and 75th percentiles
tdlmm.coexp <- adj_coexposure(sbd_exp, tdlmm.fit, contrast_perc = c(0.25, 0.75))
```

The function requires exposure data, the model fit of class `tdlmm`, and an argument `contrast_perc` which can be specified with a vector of two percentiles used for co-exposure prediction. Another argument `contrast_exp` is available for specifying exact exposure levels for each exposure.

Second, the marginal cumulative effect of each exposure can be obtained with the summary object `tdlmm.sum`. The object has a list attribute `DLM`, which contains estimates of marginal exposure effect and cumulative effect with their credible intervals. These estimated effects correspond to the argument `marginalize` specified within `summary` method. The code below returns the estimates of the cumulative effect of $PM_{2.5}$.

```
tdlmm.sum$DLM$PM25$cumulative
$mean
[1] -0.3790024

$ci.lower
      2.5%
-0.5318544

$ci.upper
     97.5%
-0.2305391
```

$PM_{2.5}$ can be replaced with other exposures if desired, e.g., `tdlmm.sum$DLM$TEMP$cumulative` for temperature. In this context, the TDLMM estimates that the marginal cumulative effect of a unit increase of $PM_{2.5}$ across all gestational weeks is -0.38 (95% CrI: [-0.53, -0.23]).

The `plot` method on the summary object `tdlmm.sum` requires a single exposure or a pair of exposures. The `plot` method returns the marginal effect of an exposure when specified with a single exposure. For instance, for the first three exposures, $PM_{2.5}$, temperature, and SO_2, we can use the following code.

```
library(gridExtra)

p1 <- plot(tdlmm.sum, exposure1 = "PM25", main = "PM2.5")
p2 <- plot(tdlmm.sum, exposure1 = "TEMP", main = "Temperature")
p3 <- plot(tdlmm.sum, exposure1 = "SO2", main = "SO2")

grid.arrange(p1, p2, p3, nrow = 1)
```

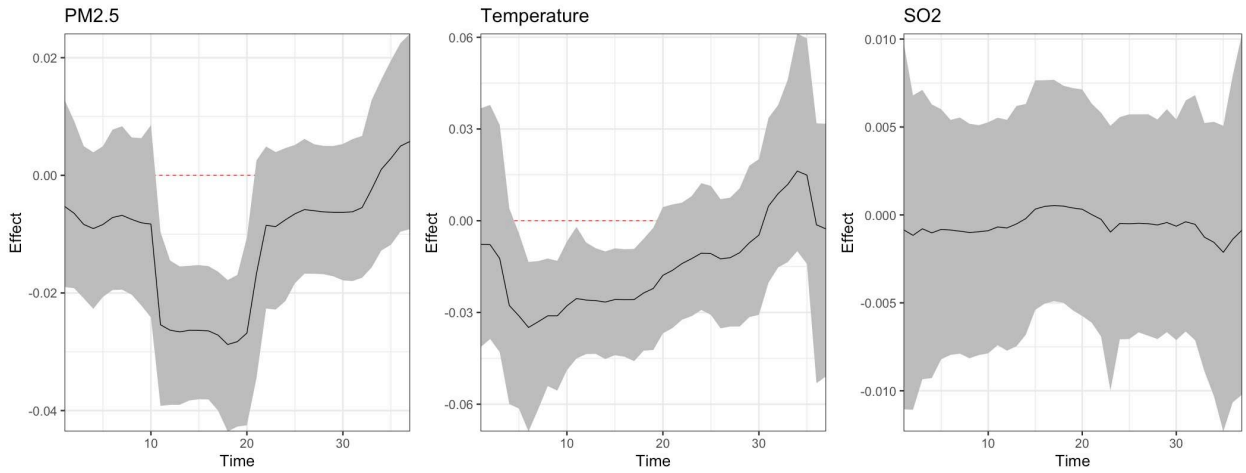


Figure 4.6: Estimated marginal distributed lag effects of $PM_{2.5}$, temperature, and SO_2 on BWGAZ during 37 gestational weeks, using TDLMM.

The `plot` method with arguments of two exposures visualizes an interaction surface of two specified exposures. The following code plots an estimated pairwise interaction surface of two specified exposures:

```
plot(tdlmm.sum, exposure1 = "PM25", exposure2 = "TEMP")
```

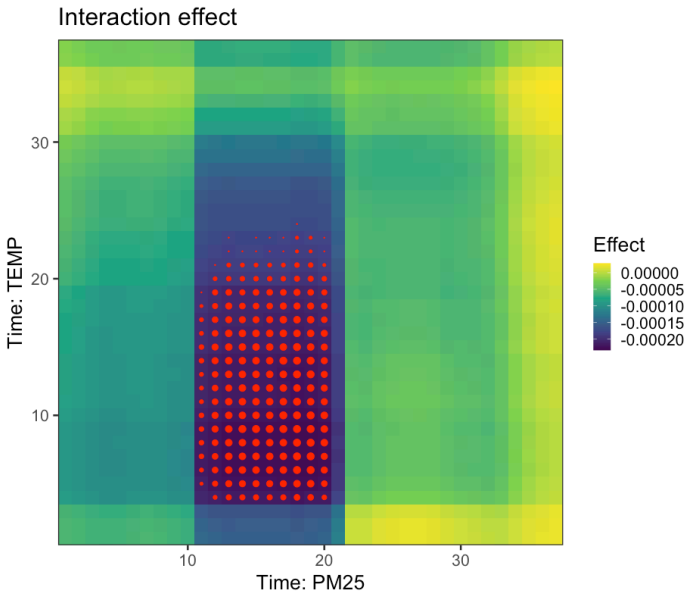


Figure 4.7: Estimated lagged interaction effects between $PM_{2.5}$ and temperature, using TDLMM.

Figure 4.7 shows the interaction surface between $PM_{2.5}$ and temperature, estimated with TDLMM. The gradient colors on the grid indicate the estimated interaction effects where the x-axis is the exposure time span of $PM_{2.5}$ and the y-axis is that of temperature. The red dots indicate that the credible interval of the effect does not contain zero with larger dots indicating a higher probability of credible intervals. Figure 4.7 indicates significant negative interaction effects at weeks 12–20 of $PM_{2.5}$ and 6–19 weeks of temperature at 95% confidence level, implying that increased exposure to $PM_{2.5}$ may decrease the exposure effect of temperature, and vice versa.

4.4.5 HDLM & HDLMM: Introducing heterogeneity to exposure effects

We illustrate heterogeneous models for a single exposure (HDLM) and for a mixture exposure (HDLMM) for estimating heterogeneous exposure effects. We focus our demonstration on a **shiny** interface built for HDLM and HDLMM for examining the most significant modifying factors, the personalized exposure effects, and the subgroup-specific exposure effects.

Suppose we are interested in the linear association between BWGAZ and a single exposure $PM_{2.5}$. We additionally assume that the exposure effect of $PM_{2.5}$ may be modified by child sex, maternal age, BMI, race, Hispanic designation, and smoking habits across the population. Specif-

ically for fitting heterogeneous models, an additional argument `hdlm.modifiers` can be set to a vector of modifier names. These modifiers must be included in the data frame provided to the `data` argument (`sbd_cov` in this example). By default, the argument is set to include all covariates included in the `formula` argument as modifiers. Also, the possible number of splitting points of modifiers for heterogeneity can be specified with an integer argument `hdlm.modifier.splits` to manage the computational cost. We specify `het = TRUE` to fit the HDLM with the following code.

```
hdlm.fit <- dlmtree(formula = bwgaz ~ ChildSex + MomAge + MomPriorBMI +
                                        Race + Hispanic + SmkAny + EstMonthConcept,
                    data = sbd_cov,
                    exposure.data = sbd_exp[["PM25"]],
                    family = "gaussian", dlm.type = "linear", het = TRUE,
                    hdlm.modifiers = c("ChildSex", "MomAge", "MomPriorBMI",
                                       "Race", "Hispanic", "SmkAny"),
                    hdlm.modifier.splits = 10,
                    n.burn = 2500, n.iter = 10000, n.thin = 5)
```

The `summary` method applied to the object `hdlm.fit` returns the overview of the model fit. The following code returns the summary.

```
hdlm.sum <- summary(hdlm.fit)
hdlm.sum

---
HDLM summary

Model run info:
- bwgaz ~ ChildSex + MomAge + MomPriorBMI + Race + Hispanic + SmkAny + EstMonthConcept
- sample size: 10,000
- family: gaussian
- 20 trees
- 2500 burn-in iterations
- 10000 post-burn iterations
- 5 thinning factor
- 0.5 modifier sparsity prior
- 0.95 confidence level

Fixed effects:
                      Mean  Lower  Upper
*(Intercept)         1.463  1.099  1.818
 ChildSexM           0.120 -0.294  0.557
```

```
 MomAge               0.001 -0.004  0.007
*MomPriorBMI         -0.020 -0.023 -0.017
 RaceAsianPI          0.064 -0.062  0.195
 RaceBlack            0.074 -0.052  0.205
 Racewhite            0.055 -0.066  0.180
 HispanicNonHispanic -0.015 -0.050  0.019
 SmkAnyY             -0.316 -0.722  0.510
 EstMonthConcept2    -0.047 -0.100  0.006
*EstMonthConcept3    -0.126 -0.188 -0.062
*EstMonthConcept4    -0.205 -0.272 -0.143
*EstMonthConcept5    -0.200 -0.249 -0.148
*EstMonthConcept6    -0.201 -0.249 -0.151
 EstMonthConcept7    -0.041 -0.093  0.017
*EstMonthConcept8     0.145  0.089  0.206
*EstMonthConcept9     0.388  0.327  0.454
*EstMonthConcept10    0.378  0.313  0.440
*EstMonthConcept11    0.324  0.269  0.376
*EstMonthConcept12    0.126  0.076  0.171
---
* = CI does not contain zero

Modifiers:
               PIP
ChildSex    1.0000
MomAge      0.8515
MomPriorBMI 0.7880
Race        0.1940
Hispanic    1.0000
SmkAny      0.4065
---
PIP = Posterior inclusion probability

residual standard errors: 0.004
---
To obtain exposure effect estimates, use the 'shiny(fit)' function.
```

As before, the summary output includes '`Model run info`' and '`Fixed effects`' sections with the estimates and the 95% credible intervals of the regression coefficients of the fixed effect. The summary additionally shows the sparsity hyperparameter set for modifier selection and the posterior inclusion probability (PIP) of the modifiers included in the model. The fitted HDLM identified

child sex, maternal age, BMI, and Hispanic designation to be the modifiers that contributed the most heterogeneity to the exposure effect of $PM_{2.5}$.

A similar model fitting process can be done when examining the heterogeneous exposure effect of a mixture of five exposures on BWGAZ. The following code additionally specifies `mixture = TRUE` and fits HDLMM with the same potential modifiers:

```
hdlmm.fit <- dlmtree(formula = bwgaz ~ ChildSex + MomAge + MomPriorBMI +
                                       Race + Hispanic + SmkAny + EstMonthConcept,
                     data = sbd_cov,
                     exposure.data = sbd_exp,
                     family = "gaussian",
                     dlm.type = "linear", mixture = TRUE, het = TRUE,
                     hdlm.modifiers = c("ChildSex", "MomAge", "MomPriorBMI",
                                        "Race", "Hispanic", "SmkAny"),
                     hdlm.modifier.splits = 10,
                     n.burn = 2500, n.iter = 10000, n.thin = 5)
```

As previously, the `summary` method on `hdlmm` model object similarly returns the summary of the fitted HDLMM. The summary output, omitted here, is similar to that of HDLM, with additional information such as the number of exposures, number of pairwise interactions, and sparsity parameters for modifier selection and exposure selection. Unlike the `summary` method applied to the model of class `tdlmm` in Section 4.4.4, marginalization methods are unavailable for the fitted model `hdlmm.fit` as the marginalization of co-exposure with heterogeneity is not well defined.

Exposure effects are estimated at the individual level and can be summarized at either the individual or subgroup level. This flexibility makes summarizing and visualizing the estimated effects challenging. A built-in **shiny** app with an object of class `hdlm` and `hdlmm` provides a comprehensive analysis of the exposure effects. HDLM and HDLMM share the same **shiny** interface but the `shiny` method applied to class `hdlmm` additionally includes an option in the panel to select an exposure of interest from mixture exposures. We present the **shiny** app interface using the fitted HDLM for a single exposure, using the same argument specification for fitting the model `hdlmm.fit`. The **shiny** app is launched with the following code.

```
shiny(hdlm.fit)
```

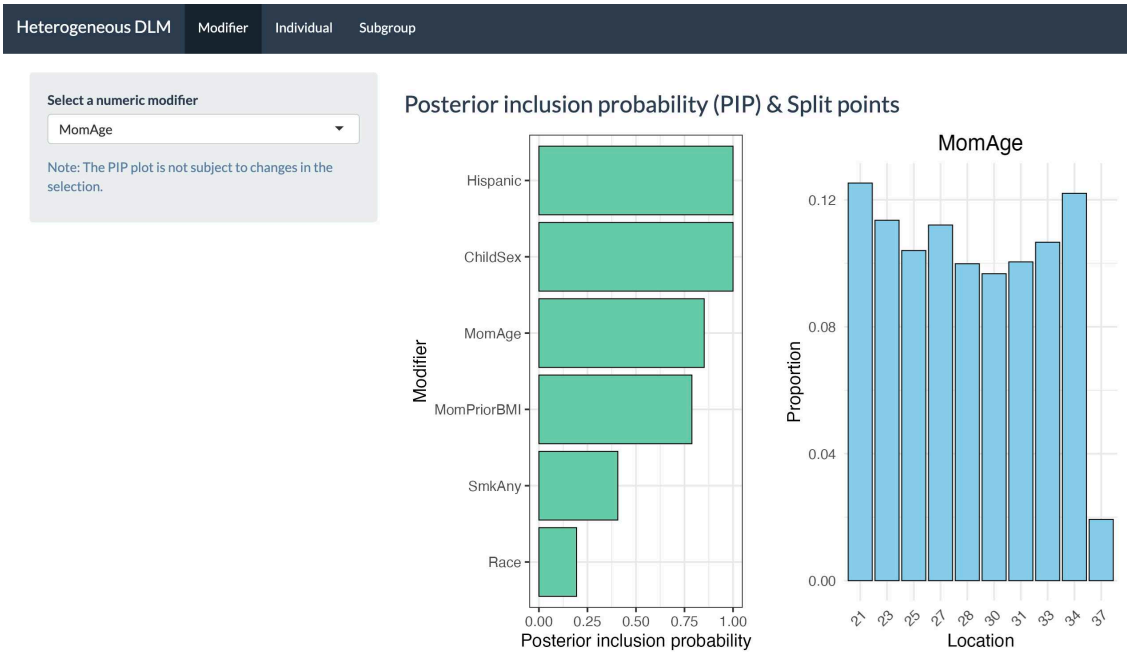


Figure 4.8: R **shiny** user interface for the fitted HDLM

Figure 4.8 shows the main screen, also the first tab, of the **shiny** app for the fitted HDLM. The **shiny** interface includes three tabs. The first tab labeled 'Modifier' presents two panels including a bar plot of modifier PIPs and the proportions of split points of a user-selected continuous modifier used to split the internal nodes of modifier trees.

In the 'Individual' tab, the user can adjust the levels of modifiers to obtain the individualized estimate of the distributed lag effects. In our context, the **shiny** app provides the personalized exposure effect and critical windows when the sex of a child, age, BMI, race, Hispanic designation, and smoking habits of a mother are specified. Figure 4.9 shows the estimated personalized exposure effect of $PM_{2.5}$ during gestational weeks for a 29-year-old white non-Hispanic mother with a BMI of 24, whose child is male and who does not smoke.

The last tab labeled 'Subgroup' offers subgroup-specific analyses. In the top panel, shown in Figure 4.10, the user can select one or two modifiers to group the samples into multiple subgroups and obtain personalized exposure effects for individuals in each subgroup. Each line of the resulting plot represents an exposure effect of an individual accounting for all modifiers of that individual. This is useful for simultaneously assessing how much exposure effects vary among

Figure 4.9: Personalized exposure effect in R **shiny** app

individuals and across subgroups. The bottom panel allows users to select one or two modifiers for subgroup-specific distributed lag effects. Subgroup-specific distributed lag effects are calculated by marginalizing out the modifiers not specified in the panel. Two modifiers at most can be specified for analyzing how much each modifier affects heterogeneity in the exposure effects. For example, Figure 4.11 shows the estimated exposure-time-response function for eight subgroups, grouped by two categorical modifiers: race and child sex. The difference in subgroup-specific exposure effects indicates that race does not contribute much heterogeneity in the exposure effect of $PM_{2.5}$ while child sex introduces a considerable amount, which aligns with the modifier PIPs in the summary output. Using the simulated dataset, the subgroup-specific effects suggest that mothers with a male child may be more vulnerable to $PM_{2.5}$.

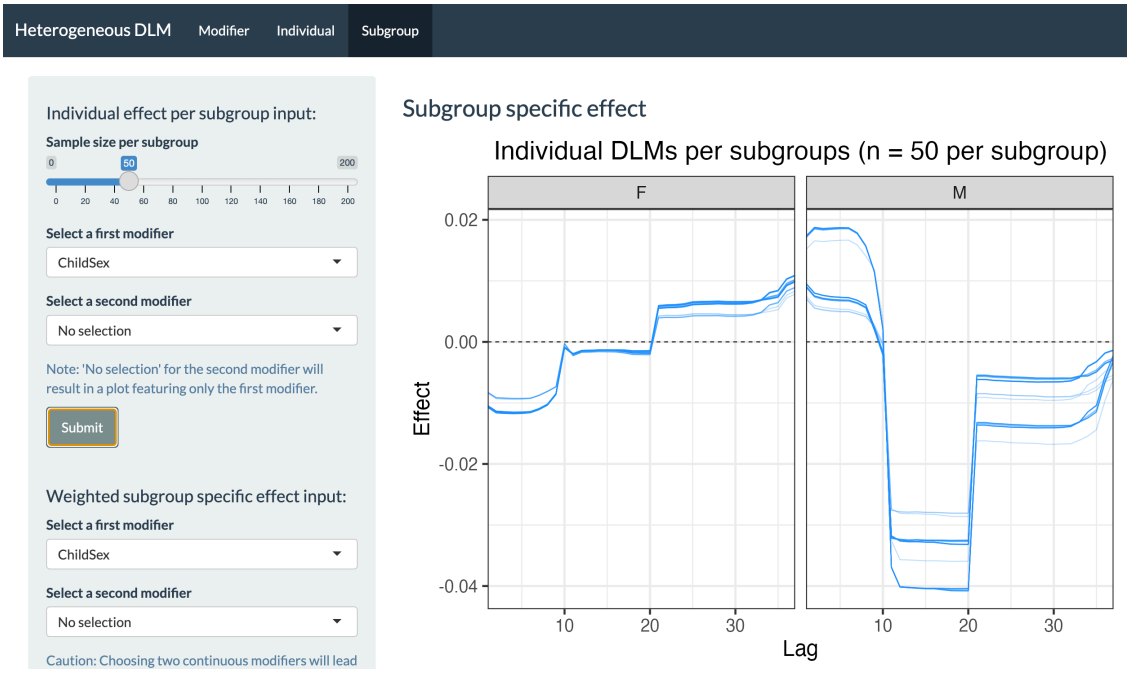


Figure 4.10: Personalized exposure effects with subgroups in R **shiny** app

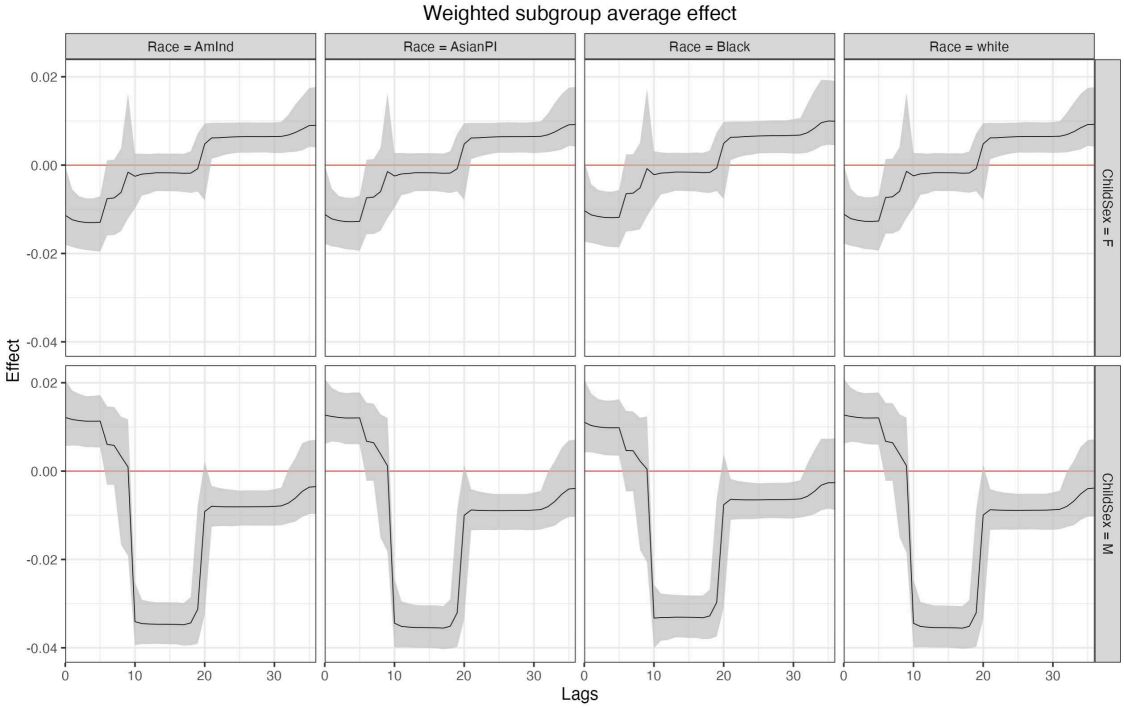


Figure 4.11: Subgroup-specific effects, grouped by race and child sex, in R **shiny** app

4.5 Summary

We introduced the R package **dlmtree**, a user friendly software for addressing a wide range of research questions regarding the relationship between a longitudinally assessed exposure or mixture and a scalar outcome. Our software provides functionality to estimate distributed lag linear or nonlinear models, quantify main and interaction effects, and account for heterogeneity in the exposure-time-response function. Furthermore, the methods in our package have been carefully optimized for computation efficiency under a custom C++ language framework with a convenient R wrapper function, `dlmtree`, designed for accessibility by all researchers. In this chapter, we provided an overview of the regression tree approaches used for estimating DLMs, and through a collection of vignettes, we highlighted a variety of tools available for processing data, fitting models, conducting inference, and visualizing the results. Our goal in making this package available is to bring robust data science tools to expand the range of questions that can be asked and answered with longitudinally assessed data.

Chapter 5

Conclusion

In this dissertation, we proposed novel statistical frameworks to estimate the exposure effects of environmental mixtures across different scenarios. In Chapter 2, we proposed employing PG augmentation within the TDLMM framework to allow for the count data with zero-inflation and overdispersion in a mixture exposure setting. Through simulation studies, we showed that the proposed method can accurately estimate exposure effects and identify windows of susceptibility with high precision, outperforming spline-based methods for a single exposure. Additionally, for mixture exposures, the method was able to perform exposure selection and incorporate pairwise lagged interaction between exposures. We applied our method to Colorado birth registry data with LBIC study design to estimate the association between $PM_{2.5}$ and maximal daily temperature and pregnancy loss. Using our method, we estimated that $PM_{2.5}$ is negatively associated with pregnancy loss with a window of susceptibility spanning across the entire gestational period. The exposure effect was more pronounced during weeks 10 to 24. Using Bayesian g-computation, the estimated exposure effect translated into 2.35 EPLD per month across all counties in Colorado for approximately 2 $\mu g/m^3$ increment of $PM_{2.5}$. We did not identify any windows of susceptibility for maximal temperature.

In Chapter 3, we proposed a novel framework to estimate heterogeneous exposure-time-response function in a mixture exposure setting. Aligning with precision environmental health objectives, we proposed HDLMM using a tree triplet structure that consists of a modifier tree for addressing heterogeneity and a DLM tree pair to structure the time-varying effects and pairwise lagged interaction. Via simulation studies, we showed that our model was able to identify modifiers that contribute to heterogeneity through modifier selection and exposures in environmental mixtures through exposure selection. Our proposed method was also able to accurately estimate heterogeneous exposure effects and identify windows of susceptibility at both individual and subgroup levels. For conducting inference for various hypothetical scenarios,

we introduced estimands suitable for the HDLMM framework: CATE and GATE. We applied the HDLMM to Colorado birth administrative data. We found several maternal factors that contributed heterogeneity to the exposure effect of $PM_{2.5}$. We first observed that $PM_{2.5}$ had a larger negative effect on mothers younger than 25 and those with a BMI between 23.01 and 34.13. The exposure effect was also modified by race and Hispanic designation as $PM_{2.5}$ had a stronger negative effect on non-Hispanic Asian, Pacific Islanders, and White mothers with longer windows of susceptibility. The exposure effect of temperature was minuscule, but the mother's age and race contributed to varying trends of the effects. In comparison to the homogeneous TDLMM, the findings from our heterogeneous model illustrated that estimating exposure effects without heterogeneity assumption could lead to biased estimates and incorrect identification of windows of susceptibility.

In Chapter 4, we introduced a comprehensive R package **dlmtree** for implementing various tree structured DLMs. We first provided a conceptual review of the tree structured DLM and its extensions. The extensions included nonlinear associations, generalized linear models, mixture exposures, and heterogeneity in exposure effects. With simulated data, we demonstrated data preparation and model fitting process for each model with the accessible main function. We highlighted functions for model summary, visualization, and inference. Finally for heterogeneous models, we presented a built-in shiny interface for interactive analysis such as modifier selection, personalized exposure effects, and subgroup-specific effects.

5.1 Future work

We suggest future research directions and enumerate possible extensions of the proposed models. First, with our proposed model, the TDLMM can incorporate continuous, binary, and count outcomes. An immediate extension would be to consider multinomial or ordinal response variables, which frequently arise in cases such as the severity of respiratory illnesses among children or risk factors for diseases. For the distributed lag models with heterogeneity, current methods and software are limited to continuous response variables, despite their broad applicability. Extending

the heterogeneous framework to accommodate binary and count outcomes would be a reasonable approach. Second, the existing framework of the DLM does not address the possibility of differences in the total time span of exposure across populations, i.e., the gestational weeks of pregnant mothers are likely to vary. Because of this limitation, studies, including ours, typically include only full-term births or similar selection criteria required to meet model assumptions. However, including only full-term births is a form of conditioning on a mediator (gestational age) and can result in bias (Neophytou et al., 2021). Therefore, there is a need for models within the DLM context that can account for variations in exposure time spans, as no established framework currently exists for addressing this. Lastly, further enhancement for the R package **dlmtree** includes researching and implementing a faster BART algorithm, integrating additional tree structured DLM methods, and developing more statistical inferential tools.

5.2 Impact

We anticipate that the proposed methods and software in this dissertation will contribute to the broader applicability of the DLM framework in various fields. First, in Chapter 2, we proposed a statistical approach that accommodates count data of zero-inflation and overdispersion, extending the TDLMM applicable for count outcomes. Our method contributes to the generalized linear model extensions of the DLM framework for both single and mixture exposures. Our method, motivated by the context of environmental health science, will facilitate studies of other various fields with count outcomes and longitudinally assessed mixture exposures data. Second, the HDLMM framework we proposed in Chapter 3 is in line with precision environmental health objectives and is the first data-driven statistical method to simultaneously estimate who is most vulnerable, when they are most vulnerable, and which exposure is most harmful. Application of the proposed method to large cohort study data, such as Environmental influences on Child Health Outcome (ECHO; Park et al., 2024) with its rich information on personal factors will improve the estimation of personalized effects and the identification of vulnerable communities. This, in turn, will lead to better-informed clinical decisions, targeted interventions, and treatments for improving public

health. Lastly, the R package **dlmtree**, introduced in Chapter 4, will facilitate the implementation of DLMs for researchers, especially those with limited experience in this modeling framework. Featuring five distinct tree structured DLMs with user-friendly functions, we expect the software to be a versatile statistical toolbox for enabling further analysis in environmental health science and other fields.

Bibliography

Abatzoglou, J. T. (2013). Development of gridded surface meteorological data for ecological applications and modelling. *International journal of climatology*, 33(1):121–131. 11, 34

Anenberg, S. C., Haines, S., Wang, E., Nassikas, N., and Kinney, P. L. (2020). Synergistic health effects of air pollution, temperature, and pollen exposure: a systematic review of epidemiological evidence. *Environmental Health*, 19(1):130. 31

Antonelli, J., Wilson, A., and Coull, B. A. (2024). Multiple exposure distributed lag models with variable selection. *Biostatistics*, 25(1):1–19. 4, 9, 32, 35

Baccarelli, A., Dolinoy, D. C., and Walker, C. L. (2023). A precision environmental health approach to prevention of human disease. *Nature Communications*, 14(1):2449. 31

Bekkar, B., Pacheco, S., Basu, R., and DeNicola, N. (2020). Association of Air Pollution and Heat Exposure With Preterm Birth, Low Birth Weight, and Stillbirth in the US: A Systematic Review. *JAMA Network Open*, 3(6):e208243–e208243. 1

Bell, M. L., Gasparrini, A., and Benjamin, G. C. (2024). Climate change, extreme heat, and health. *New England Journal of Medicine*, 390(19):1793–1801. 1

Bello, G. A., Arora, M., Austin, C., Horton, M. K., Wright, R. O., and Gennings, C. (2017). Extending the Distributed Lag Model framework to handle chemical mixtures. *Environmental Research*, 156:253–264. 4, 9, 32

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010). A Spatio-Temporal Downscaler for Output From Numerical Models. *Journal of agricultural, biological, and environmental statistics*, 15(2):176–197. 34

Billionnet, C., Sherrill, D., Annesi-Maesano, I., et al. (2012). Estimating the health effects of exposure to multi-pollutant mixture. *Annals of Epidemiology*, 22(2):126–141. 31

Brunst, K. J., Tignor, N., Just, A., Liu, Z., Lin, X., Hacker, M. R., Bosquet Enlow, M., Wright, R. O., Wang, P., Baccarelli, A. A., and Wright, R. J. (2018). Cumulative lifetime maternal stress and epigenome-wide placental DNA methylation in the PRISM cohort. *Epigenetics*, 13(6):665–681. 31

Burke, M., Childs, M. L., de la Cuesta, B., Qiu, M., Li, J., Gould, C. F., Heft-Neal, S., and Wara, M. (2023). The contribution of wildfire to pm2.5 trends in the usa. *Nature*, 622(7984):761–766. 1

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480. 17, 38

Casey, J. A., James, P., Rudolph, K. E., Wu, C.-D., and Schwartz, B. S. (2016). Greenness and Birth Outcomes in a Range of Pennsylvania Communities. *International Journal of Environmental Research and Public Health*, 13(3):311. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute. 31

Chen, Y.-H., Mukherjee, B., and Berrocal, V. J. (2019). Distributed Lag Interaction Models with Two Pollutants. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 68(1):79–97. 3, 10

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1). arXiv:0806.3286 [stat]. 5, 16, 38, 39, 54, 103

Chiu, Y.-H. M., Carroll, K. N., Coull, B. A., Kannan, S., Wilson, A., and Wright, R. J. (2022). Prenatal Fine Particulate Matter, Maternal Micronutrient Antioxidant Intake, and Early Childhood Repeated Wheeze: Effect Modification by Race/Ethnicity and Sex. *Antioxidants*, 11(2):366. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. 6, 31

Chiu, Y.-H. M., Hsu, H.-H. L., Coull, B. A., Bellinger, D. C., Kloog, I., Schwartz, J., Wright, R. O., and Wright, R. J. (2016). Prenatal particulate air pollution and neurodevelopment in urban chil-

dren: Examining sensitive windows and sex-specific associations. *Environment International*, 87:56–65. 1

Chiu, Y.-H. M., Wilson, A., Hsu, H.-H. L., Jamal, H., Mathews, N., Kloog, I., Schwartz, J., Bellinger, D. C., Xhani, N., Wright, R. O., Coull, B. A., and Wright, R. J. (2023). Prenatal ambient air pollutant mixture exposure and neurodevelopment in urban children in the Northeastern United States. *Environmental Research*, 233:116394. 53

Crouse, D. L., Peters, P. A., Hystad, P., Brook, J. R., van Donkelaar, A., Martin, R. V., Villeneuve, P. J., Jerrett, M., Goldberg, M. S., Pope III, C. A., et al. (2015). Ambient pm2.5, o3, and no2 exposures and associations with mortality over 16 years of follow-up in the canadian census health and environment cohort (canchec). *Environmental health perspectives*, 123(11):1180–1186. 1

Dadvand, P., Parker, J., Bell, M. L., Bonzini, M., Brauer, M., Darrow, L. A., Gehring, U., Glinianaia, S. V., Gouveia, N., Ha, E.-h., et al. (2013). Maternal exposure to particulate air pollution and term birth weight: a multi-country evaluation of effect and heterogeneity. *Environmental health perspectives*, 121(3):267–373. 2

Darrow, L. A., Huang, M., Warren, J. L., Strickland, M. J., Holmes, H. A., Newman, A. J., and Chang, H. H. (2024). Preterm and Early-Term Delivery After Heat Waves in 50 US Metropolitan Areas. *JAMA Network Open*, 7(5):e2412055–e2412055. 1

Dearborn, L. C., Hazlehurst, M. F., Loftus, C. T., Szpiro, A. A., Carroll, K. N., Moore, P. E., Adgent, M. A., Barrett, E. S., Nguyen, R. H., Sathyanarayana, S., et al. (2023). Role of air pollution in the development of asthma among children with a history of bronchiolitis in infancy. *Epidemiology*, 34(4):554–564. 1

Demateis, D., Keller, K. P., Rojas-Rueda, D., Kioumourtzoglou, M.-A., and Wilson, A. (2024). Penalized Distributed Lag Interaction Model: Air Pollution, Birth Weight and Neighborhood Vulnerability. arXiv:2401.02939 [stat]. 6, 32, 54

Di, Q., Dai, L., Wang, Y., Zanobetti, A., Choirat, C., Schwartz, J. D., and Dominici, F. (2017). Association of Short-term Exposure to Air Pollution With Mortality in Older Adults. *JAMA*, 318(24):2446–2456. 1

Gao, Y. and Kowal, D. R. (2024). Bayesian adaptive and interpretable functional regression for exposure profiles. *The Annals of Applied Statistics*, 18(1):642–663. 3

Gasparrini, A. and Armstrong, B. (2013). Distributed lag non-linear models in r: the package dlnm. *London School of Hygiene and Tropical Medicine, UK. dlnm version*, 1(7):05–16. 54

Gasparrini, A., Armstrong, B., and Kenward, M. G. (2010). Distributed lag non-linear models. *Statistics in Medicine*, 29(21):2224–2234. 3, 9, 32, 53

Hastie, T. and Tibshirani, R. (2000a). Bayesian Backfitting. *Statistical Science*, 15(3):196–213. Publisher: Institute of Mathematical Statistics. 110

Hastie, T. and Tibshirani, R. (2000b). Bayesian backfitting (with comments and a rejoinder by the authors. *Statistical Science*, 15(3):196–223. Publisher: Institute of Mathematical Statistics. 18, 103

Hsu, H.-H. L., Chiu, Y.-H. M., Coull, B. A., Kloog, I., Schwartz, J., Lee, A., Wright, R. O., and Wright, R. J. (2015). Prenatal Particulate Air Pollution and Asthma Onset in Urban Children. Identifying Sensitive Windows and Sex Differences. *American Journal of Respiratory and Critical Care Medicine*, 192(9):1052–1059. 1

Hsu, H.-H. L., Wilson, A., Schwartz, J., Kloog, I., Wright, R. O., Coull, B. A., and Wright, R. J. (2023). Prenatal Ambient Air Pollutant Mixture Exposure and Early School-age Lung Function. *Environmental Epidemiology*, 7(2):e249. 32, 53

Karlsson, M. and Ziebarth, N. R. (2018). Population health effects and health-related costs of extreme temperatures: Comprehensive evidence from germany. *Journal of Environmental Economics and Management*, 91:93–117. 1

Kioumourtzoglou, M.-A., Raz, R., Wilson, A., Fluss, R., Nirel, R., Broday, D. M., Yuval, Hacker, M. R., McElrath, T. F., Grotto, I., Koutrakis, P., and Weisskopf, M. G. (2019). Traffic-Related Air Pollution and Pregnancy Loss. *Epidemiology (Cambridge, Mass.)*, 30(1):4–10. 8, 11, 30

Koyck, L. M. (1954). *Distributed lags and investment analysis*, volume 4. North-Holland Publishing Company Amsterdam. 53

Lee, A., Leon Hsu, H.-H., Mathilda Chiu, Y.-H., Bose, S., Rosa, M. J., Kloog, I., Wilson, A., Schwartz, J., Cohen, S., Coull, B. A., Wright, R. O., and Wright, R. J. (2018). Prenatal fine particulate exposure and early childhood asthma: Effect of maternal stress and fetal sex. *Journal of Allergy and Clinical Immunology*, 141(5):1880–1886. 6, 31

Leung, M., Kioumourtzoglou, M.-A., Raz, R., and Weisskopf, M. G. (2021). Bias due to Selection on Live Births in Studies of Environmental Exposures during Pregnancy: A Simulation Study. *Environmental Health Perspectives*, 129(4):47001. 11

Leung, M., Rowland, S. T., Coull, B. A., Modest, A. M., Hacker, M. R., Schwartz, J., Kioumourtzoglou, M.-A., Weisskopf, M. G., and Wilson, A. (2023). Bias Amplification and Variance Inflation in Distributed Lag Models Using Low-Spatial-Resolution Data. *American Journal of Epidemiology*, 192(4):644–657. 10, 24, 30

Linero, A. R. (2018). Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection. *Journal of the American Statistical Association*, 113(522):626–636. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2016.1264957. 17, 39

Liu, C., Chen, R., Sera, F., Vicedo-Cabrera, A. M., Guo, Y., Tong, S., Coelho, M. S., Saldiva, P. H., Lavigne, E., Matus, P., et al. (2019). Ambient particulate air pollution and daily mortality in 652 cities. *New England Journal of Medicine*, 381(8):705–715. 1

Liu, S. H., Bobb, J. F., Lee, K. H., Gennings, C., Claus Henn, B., Bellinger, D., Austin, C., Schnaas, L., Tellez-Rojo, M. M., Hu, H., Wright, R. O., Arora, M., and Coull, B. A. (2018).

Lagged kernel machine regression for identifying time windows of susceptibility to exposures of complex mixtures. *Biostatistics (Oxford, England)*, 19(3):325–341. 32

Makalic, E. and Schmidt, D. F. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182. arXiv:1508.03884 [stat]. 17

Medina-Ramón, M., Zanobetti, A., Cavanagh, D. P., and Schwartz, J. (2006). Extreme temperatures and mortality: assessing effect modification by personal characteristics and specific cause of death in a multi-city case-only analysis. *Environmental health perspectives*, 114(9):1331–1336. 1

Mork, D., Kioumourtzoglou, M.-A., Weisskopf, M., Coull, B. A., and Wilson, A. (2023). Heterogeneous Distributed Lag Models to Estimate Personalized Effects of Maternal Exposures to Air Pollution. *Journal of the American Statistical Association*, pages 1–13. arXiv:2109.13763 [stat]. 33, 54, 60, 111, 144

Mork, D. and Wilson, A. (2022). Treed distributed lag nonlinear models. *Biostatistics (Oxford, England)*, 23(3):754–771. 39, 54, 59, 60, 133

Mork, D. and Wilson, A. (2023). Estimating perinatal critical windows of susceptibility to environmental mixtures via structured Bayesian regression tree pairs. *Biometrics*, 79(1):449–461. 3, 4, 5, 9, 10, 14, 15, 18, 23, 29, 32, 35, 37, 38, 39, 40, 54, 56, 57, 60, 103, 129, 138

Mork, D. and Wilson, A. (2024). Incorporating prior information into distributed lag nonlinear models with zero-inflated monotone regression trees. *Bayesian Analysis*, 1(1):1–29. 59

Mortimer, K., Neugebauer, R., Lurmann, F., Alcorn, S., Balmes, J., and Tager, I. (2008). Air pollution and pulmonary function in asthmatic children: effects of prenatal and lifetime exposures. *Epidemiology*, 19(4):550–557. 2

Muggeo, V. M. R. (2007). Bivariate distributed lag models for the analysis of temperature-by-pollutant interaction effect on mortality. *Environmetrics*, 18(3):231–243. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/env.829. 3

Neelon, B. (2019). Bayesian Zero-Inflated Negative Binomial Regression Based on Pólya-Gamma Mixtures. *Bayesian analysis*, 14(3):829–855. 12, 18, 29, 60

Neophytou, A. M., Kioumourtzoglou, M.-A., Goin, D. E., Darwin, K. C., and Casey, J. A. (2021). Educational note: addressing special cases of bias that frequently occur in perinatal epidemiology. *International journal of epidemiology*, 50(1):337–345. 85

Odden, M. C., Rawlings, A. M., Khodadadi, A., Fern, X., Shlipak, M. G., Bibbins-Domingo, K., Covinsky, K., Kanaya, A. M., Lee, A., Haan, M. N., Newman, A. B., Psaty, B. M., and Peralta, C. A. (2020). Heterogeneous Exposure Associations in Observational Cohort Studies: The Example of Blood Pressure in Older Adults. *American Journal of Epidemiology*, 189(1):55–67. 33

Palda, K. S. (1965). The measurement of cumulative advertising effects. *The Journal of Business*, 38(2):162–179. 53

Park, C. H., Blaisdell, C. J., Arteaga, S. S., Mash, C., Laessig, S., Hanspal, M., Luetkemeier, E., Thompson, L. C., and Gillman, M. W. (2024). How the environmental influences on child health outcome (echo) cohort can spur discoveries in environmental epidemiology. *American Journal of Epidemiology*, page kwae073. 85

Park, S. K., Tao, Y., Meeker, J. D., Harlow, S. D., and Mukherjee, B. (2014). Environmental Risk Score as a New Tool to Examine Multi-Pollutants in Epidemiologic Research: An Example from the NHANES Study Using Serum Lipid Levels. *PLoS ONE*, 9(6):e98632. 31

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Polya-Gamma latent variables. arXiv:1205.0310 [stat]. 60

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 19

Rosa, M. J., Hsu, H.-H. L., Just, A. C., Brennan, K. J., Bloomquist, T., Kloog, I., Pantic, I., Mercado García, A., Wilson, A., Coull, B. A., Wright, R. O., Téllez Rojo, M. M., Baccarelli,

A. A., and Wright, R. J. (2019a). Association between prenatal particulate air pollution exposure and telomere length in cord blood: Effect modification by fetal sex. *Environmental Research*, 172:495–501. 60

Rosa, M. J., Nentin, F., Bosquet Enlow, M., Hacker, M. R., Pollas, N., Coull, B., and Wright, R. J. (2019b). Sex-specific associations between prenatal negative life events and birth outcomes. *Stress*, 22(6):647–653. 1

Salam, M. T., Millstein, J., Li, Y.-F., Lurmann, F. W., Margolis, H. G., and Gilliland, F. D. (2005). Birth outcomes and prenatal exposure to ozone, carbon monoxide, and particulate matter: Results from the childrens health study. *Environmental Health Perspectives*, 113(11):1638–1644. 1

Schliep, E. M., Schafer, T. L., and Hawkey, M. (2021). Distributed lag models to identify the cumulative effects of training and recovery in athletes using multivariate ordinal wellness data. *Journal of Quantitative Analysis in Sports*, 17(3):241–254. 53

Schwartz, J. (2000). The distributed lag between air pollution and daily deaths. *Epidemiology (Cambridge, Mass.)*, 11(3):320–326. 9, 32, 53

Schwerdtfeger, K. L. and Shreffler, K. M. (2009). Trauma of Pregnancy Loss and Infertility for Mothers and Involuntarily Childless Women in the Contemporary United States. *Journal of loss & trauma*, 14(3):211–227. 8

Strand, L. B., Barnett, A. G., and Tong, S. (2011). Methodological challenges when estimating the effects of season and seasonal exposures on birth outcomes. *BMC Medical Research Methodology*, 11:49. 11

Vose, R., Easterling, D. R., Kunkel, K., LeGrande, A., and Wehner, M. (2017). Temperature changes in the united states. *Climate science special report: Fourth national climate assessment*, 1(GSFC-E-DAA-TN49028). 1

Walter, K. (2023). Early Pregnancy Loss. *JAMA*, 329(16):1426. 8

Wang, Y., Ghassabian, A., Gu, B., Afanasyeva, Y., Li, Y., Trasande, L., and Liu, M. (2023). Semi-parametric distributed lag quantile regression for modeling time-dependent exposure mixtures. *Biometrics*, 79(3):2619–2632. 4, 9, 32

Warren, J., Fuentes, M., Herring, A., and Langlois, P. (2012). Spatial-temporal modeling of the association between air pollution exposure and preterm birth: identifying critical windows of exposure. *Biometrics*, 68(4):1157–1167. 54

Warren, J. L., Chang, H. H., Warren, L. K., Strickland, M. J., Darrow, L. A., and Mulholland, J. A. (2022). Critical window variable selection for mixtures: Estimating the impact of multiple air pollutants on stillbirth. *The Annals of Applied Statistics*, 16(3):1633–1652. Publisher: Institute of Mathematical Statistics. 4, 9, 32

Warren, J. L., Luben, T. J., and Chang, H. H. (2020). A spatially varying distributed lag model with application to an air pollution and term low birth weight study. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 69(3):681–696. 3

Wilcox, A. J., Weinberg, C. R., O'Connor, J. F., Baird, D. D., Schlatterer, J. P., Canfield, R. E., Armstrong, E. G., and Nisula, B. C. (1988). Incidence of Early Loss of Pregnancy. *New England Journal of Medicine*, 319(4):189–194. 8

Wilson, A., Chiu, Y.-H. M., Hsu, H.-H. L., Wright, R. O., Wright, R. J., and Coull, B. A. (2017a). Bayesian distributed lag interaction models to identify perinatal windows of vulnerability in children's health. *Biostatistics (Oxford, England)*, 18(3):537–552. 3, 6, 32, 54

Wilson, A., Chiu, Y.-H. M., Hsu, H.-H. L., Wright, R. O., Wright, R. J., and Coull, B. A. (2017b). Potential for Bias When Estimating Critical Windows for Air Pollution in Children's Health. *American Journal of Epidemiology*, 186(11):1281–1289. 2, 9, 32

Wilson, A., Hsu, H.-H. L., Chiu, Y.-H. M., Wright, R. O., Wright, R. J., and Coull, B. A. (2022). Kernel machine and distributed lag models for assessing windows of susceptibility to environmental mixtures in children's health studies. *The Annals of Applied Statistics*, 16(2):1090–1110. 10, 32

Wright, R. O. (2017). Environment, susceptibility windows, development, and child health. *Current Opinion in Pediatrics*, 29(2):211–217. 1, 9, 32

Zanobetti, A., Wand, M. P., Schwartz, J., and Ryan, L. M. (2000). Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics (Oxford, England)*, 1(3):279–292. 3, 9, 54

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, 27(8). 19

Zeka, A., Melly, S. J., and Schwartz, J. (2008). The effects of socioeconomic status and indices of physical environment on reduced birth weight and preterm births in Eastern Massachusetts. *Environmental Health*, 7(1):60. 1

Appendix A

Treed Distributed Lag Mixture Model With Zero-Inflated Count Data to Investigate the Association Between Air Pollution and Pregnancy Loss

A.1 Additional figures for the data description

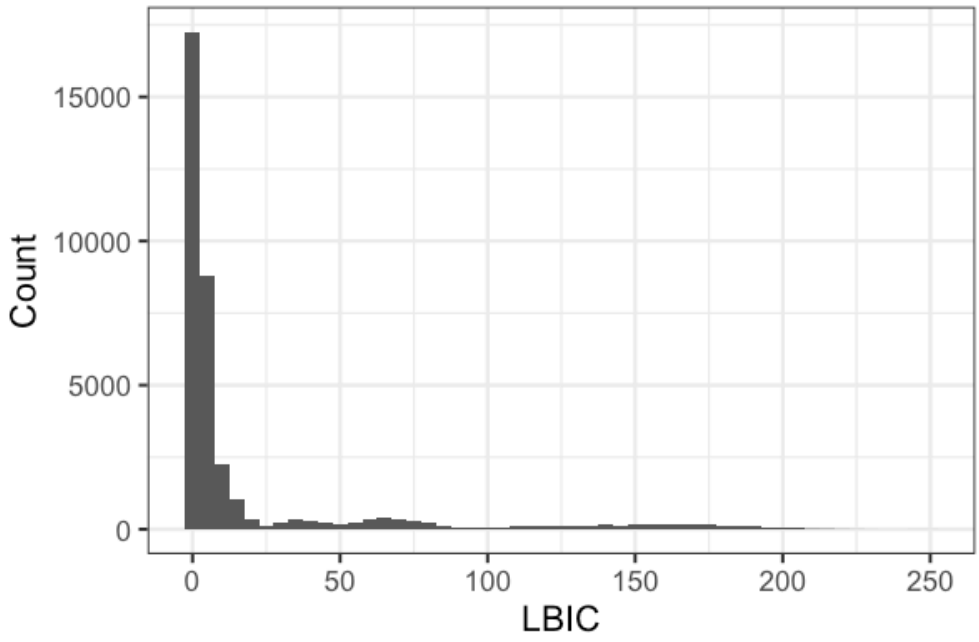


Figure A.1: LBIC histogram

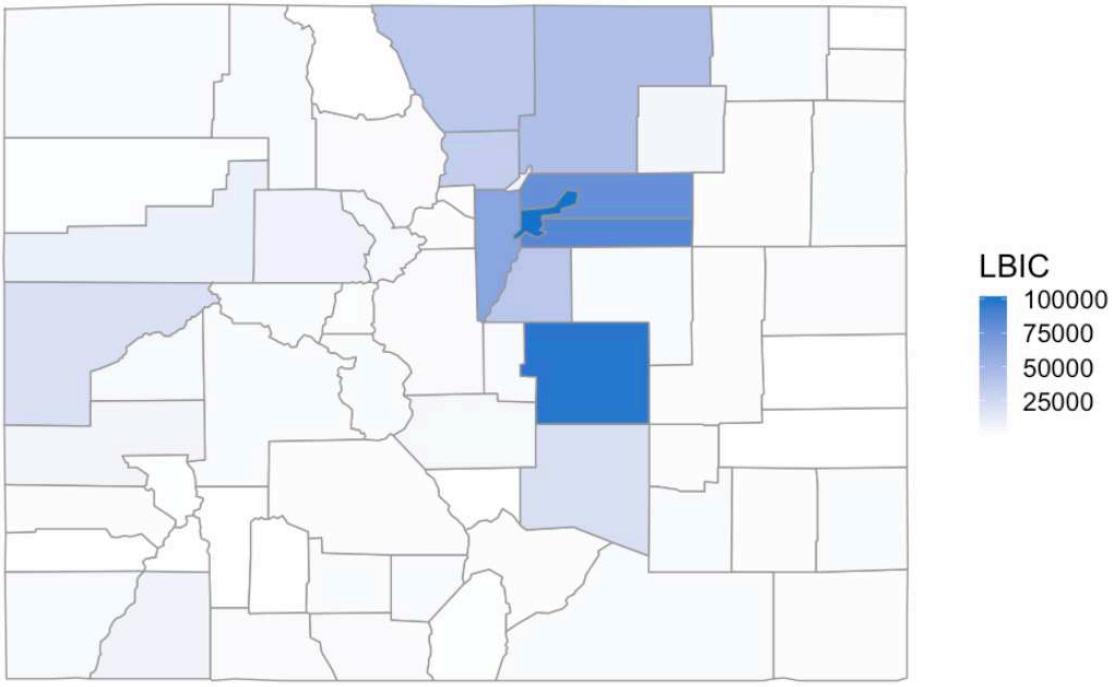


Figure A.2: LBIC per county

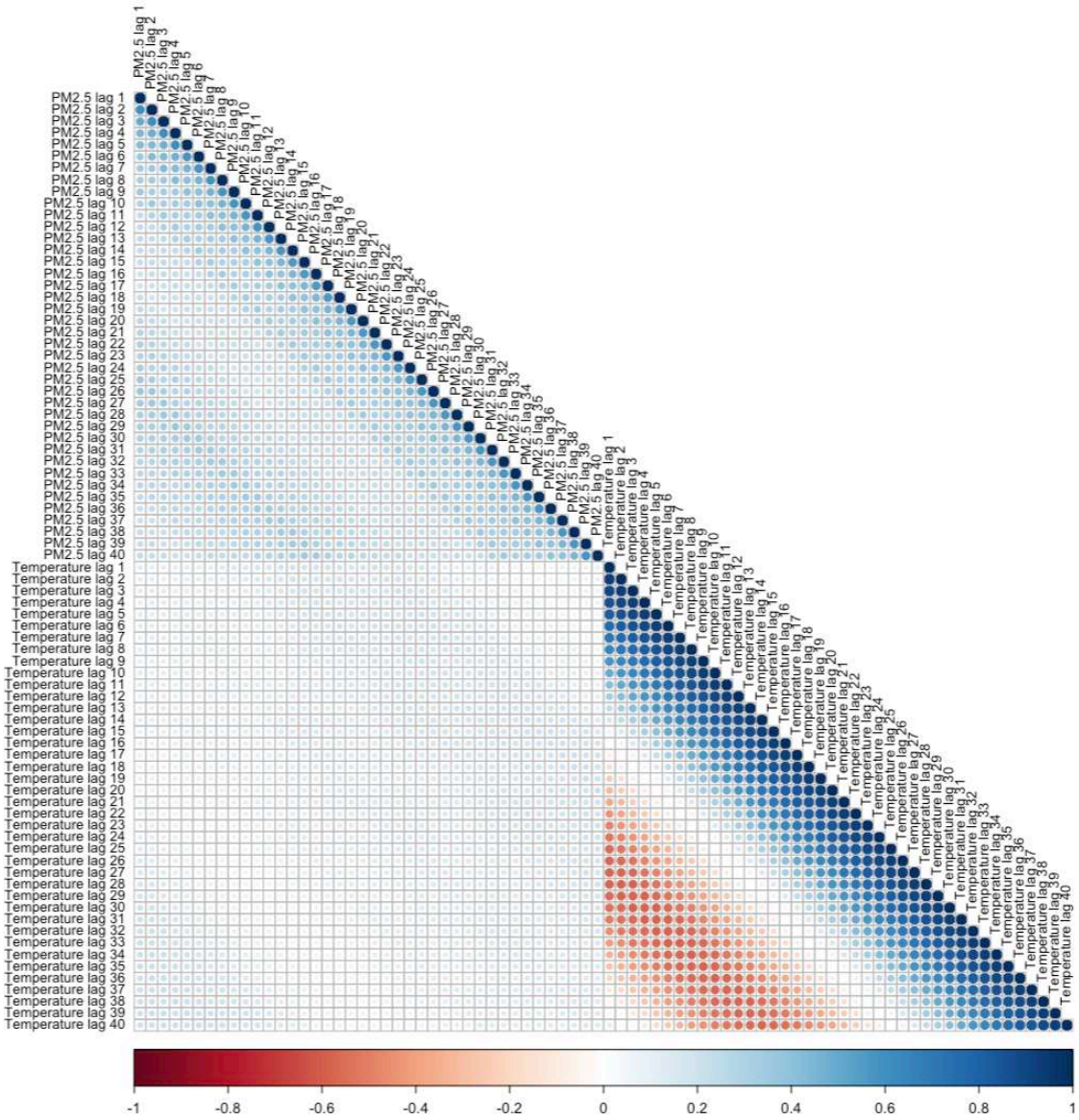


Figure A.3: Autocorrelation and correlation between exposures

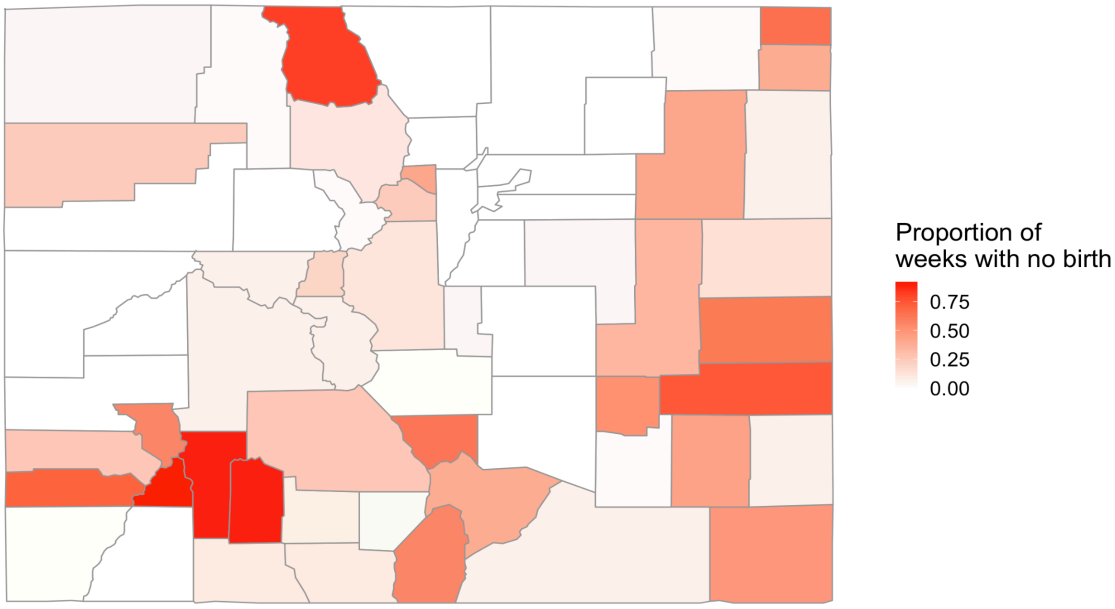


Figure A.4: Proportion of weeks with no birth per county

A.2 Outline of MCMC algorithm

Step 0: Initial sampling

We initialize parameters with specified prior distributions. First, w_i is initialized as 0.5 if $y_i > 0$ and 1 otherwise. Two diagonal matrices of Pólya-Gamma (PG) variables, $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ are initially set to identity matrices. Lastly, the dispersion parameter r is initialized at 5 considering its support from 1 to 10. The design matrices of the ZI and NB model, $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, are standardized, and exposure data is scaled to have a standard deviation of 1. We sample the full conditional posterior distributions of parameters of interest with Gibbs sampler repeating steps 1–5.

Step 1: Update w

1. Calculate $\pi_i = \frac{\exp(z_{1i}'\boldsymbol{\gamma}_1)}{1+\exp(z_{1i}'\boldsymbol{\gamma}_1)}, \quad 1 \leq i \leq n.$
2. Calculate $\nu_i = 1 - \psi_i = 1 - \frac{\exp(z_{2i}'\boldsymbol{\gamma}_2 + f_i)}{1+\exp(z_{2i}'\boldsymbol{\gamma}_2 + f_i)}.$
3. Calculate $\theta_i^{ZI} = \frac{\pi_i}{\nu_i^r(1-\pi_i)+\pi_i}.$

   Below illustrates the derivation of the Bernoulli distribution probability θ_i^{ZI} to update w_i and estimate π_i:

$$\begin{aligned}
\theta_i^{ZI} &= \mathbb{P}(\mathrm{w}_i = 1 | y_i = 0) \\
&= \frac{\mathbb{P}(\mathrm{w}_i = 1, y_i = 0)}{\mathbb{P}(y_i = 0)} \\
&= \frac{\mathbb{P}(y_i = 0 | \mathrm{w}_i = 1)\mathbb{P}(\mathrm{w}_i = 1)}{\mathbb{P}(y_i = 0 | \mathrm{w}_i = 1)\mathbb{P}(\mathrm{w}_i = 1) + \mathbb{P}(y_i = 0 | \mathrm{w}_i = 0)\mathbb{P}(\mathrm{w}_i = 0)}.
\end{aligned}$$

   As $\mathrm{w}_i = 1$ implies that an observation i belongs to a zero-mass, $\mathbb{P}(y_i = 0 | \mathrm{w}_i = 1) = 1$. $\mathbb{P}(y_i = 0 | \mathrm{w}_i = 0)$ is the probability of observing a zero from a negative binomial distribution which yields

$$\mathbb{P}(y_i = 0 | \mathrm{w}_i = 0) = \frac{\Gamma(0+r)}{\Gamma(r)0!}(1-\psi_i)^r \psi_i^0 = (1-\psi_i)^r = \nu_i^r.$$

Therefore,

$$\begin{aligned}\theta_i^{ZI} &= \frac{\mathbb{P}(y_i = 0|\mathrm{w}_i = 1)\mathbb{P}(\mathrm{w}_i = 1)}{\mathbb{P}(y_i = 0|\mathrm{w}_i = 1)\mathbb{P}(\mathrm{w}_i = 1) + \mathbb{P}(y_i = 0|\mathrm{w}_i = 0)\mathbb{P}(\mathrm{w}_i = 0)} \\ &= \frac{1 \cdot \pi_i}{1 \cdot \pi_i + (1 - \psi_i)^r(1 - \pi_i)} \\ &= \frac{\pi_i}{\nu_i^r(1 - \pi_i) + \pi_i}.\end{aligned}$$

4. Update the latent variable, $\mathbf{w}$ as follows:

$$\mathrm{w}_i = \begin{cases} 0 & \text{if } y_i > 0 \\ \sim \text{Bernoulli}(\theta_i^{ZI}) & \text{if } y_i = 0. \end{cases}$$

Step 2: Update $\boldsymbol{\gamma}_1$

With the updated $\mathbf{w}$, we sample $\boldsymbol{\gamma}_1$. The process follows a Bayesian logistic regression with PG augmentation where $\mathbf{w}$ is treated as a binary response variable, and $\boldsymbol{Z}_1$ is a design matrix.

1. Update $\boldsymbol{\gamma}_1$

   (a) Calculate $\boldsymbol{\Sigma}_{\gamma_1} = \boldsymbol{Z}_1'\boldsymbol{\Omega}_1\boldsymbol{Z}_1 + \boldsymbol{I}/c$. ($c = 100$ following the $\boldsymbol{\gamma}_1$ prior)

   (b) Calculate $\boldsymbol{l}_1 = \left(\frac{\mathrm{w}_1 - 0.5}{\omega_{11}}, \ldots, \frac{\mathrm{w}_n - 0.5}{\omega_{1n}}\right)$.

   (c) Sample $\boldsymbol{\gamma}_1 \sim \text{MVN}(\boldsymbol{\Sigma}_{\gamma_1}\boldsymbol{Z}_1'\boldsymbol{\Omega}_1\boldsymbol{l}_1, \boldsymbol{\Sigma}_{\gamma_1})$.

2. Update the PG variable, $\boldsymbol{\omega}_1$

   (a) Update $\omega_{1i} \sim \text{PG}(1, z_{1i}'\boldsymbol{\gamma}_1)$.

   (b) Set $\boldsymbol{\Omega}_1 = \text{diag}(\boldsymbol{\omega}_1)$.

Step 3: Update $\boldsymbol{\gamma}_2$ with distributed lag function, $\boldsymbol{f}$

In step 3, we only consider observations with $\mathrm{w}_i = 0$. For notational convenience, we use an asterisk to specify $\mathrm{w}_i = 0$. For example, $\boldsymbol{Z}_2^*$ denotes a matrix of $\boldsymbol{Z}_2$ where each column is multiplied

by $\mathbf{1} - \mathbf{w}$ where $\mathbf{1}$ is a vector of 1's. Similarly for a vector, $\boldsymbol{f}^* = \boldsymbol{f} \circ (\mathbf{1} - \mathbf{w})$ where $\circ$ denotes Hadamard product. The asterisk notation does not change the dimension but simply zeros out the observations with $\mathrm{w}_i = 1$. Similarly to step 2, we sample $\boldsymbol{\gamma}_2$ with PG augmentation including the exposure effect, $\boldsymbol{f}$.

1. Update $\boldsymbol{\gamma}_2$

   (a) Calculate $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2} = \boldsymbol{Z}_2^{*'}\boldsymbol{\Omega}_2\boldsymbol{Z}_2^* + \boldsymbol{I}/c$. ($c = 100$ following the $\boldsymbol{\gamma}_2$ prior)

   (b) Calculate $\boldsymbol{l}_2 = \left(\frac{y_1 - r}{\omega_{21}}, \ldots, \frac{y_n - r}{\omega_{2n}}\right)$.

   (c) Sample $\boldsymbol{\gamma}_2 \sim \text{MVN}(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}\boldsymbol{Z}_2^{*'}\boldsymbol{\Omega}_2(\boldsymbol{l}_2^* - \boldsymbol{f}^*), \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2})$.

2. Update the PG variable, $\boldsymbol{\omega}_2$

   (a) Update $\omega_{2i} \sim \text{PG}(y_i + r, z_{2i}'\boldsymbol{\gamma}_2 + f_i^*)$.

   (b) Set $\boldsymbol{\Omega}_2 = \text{diag}(\omega_{21}, \ldots, \omega_{2n})$.

Step 4: Update $\boldsymbol{f}$

To update $\boldsymbol{f}$, we iterate through A trees in the ensemble and update trees with the transition steps uniformly selected from grow, prune, change, and switch-component. To obtain the terminal node-specific effects, we calculate the partial residual of each tree and iterate through the Bayesian backfitting algorithm as proposed by Chipman et al. (2010) and Hastie and Tibshirani (2000b). The full derivation specific to treed DLM structure can be found in Supporting Information of Mork and Wilson (2023). For each tree $\mathcal{T}_a$ for $a = 1, \ldots, A$, we repeat the following:

1. Randomly select a transition step from grow, prune, change, and switch-component with equal probability.

2. Update tree with the marginal likelihood of the partial residual $\boldsymbol{R}_a$ using the Metropolis-Hastings (MH) algorithm and Bayesian backfitting. The marginal likelihood of $\boldsymbol{R}_a$ after

integrating out all parameters specific to $\mathcal{T}_a$ denoted $\boldsymbol{\Theta}_a$, including Λ_a and $\mathcal{I}_a$, is

$$\mathbb{P}(\boldsymbol{R}_a|-) \propto |\tau_a^2\nu^2|_{B_a}^{-1/2}|\boldsymbol{\Sigma}_{\boldsymbol{\Theta}_a}|_{B_a}^{1/2}\exp\left\{-\frac{1}{2}\boldsymbol{R}_a'(\boldsymbol{\Sigma}_l^{-1}-\boldsymbol{\Sigma}_l^{-1}\boldsymbol{X}_a\boldsymbol{\Sigma}_{\boldsymbol{\Theta}_a}^{-1}\boldsymbol{X}_a'\boldsymbol{\Sigma}_l^{-1})\boldsymbol{R}_a\right\},$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}_a} = \left(\boldsymbol{X}_a'\boldsymbol{\Sigma}_l'\boldsymbol{X}_a + \frac{1}{\tau_a^2\nu^2}\boldsymbol{I}\right)^{-1}$, $\boldsymbol{\Sigma}_l = (\boldsymbol{\Omega}_2 - \boldsymbol{\Omega}_2\boldsymbol{Z}_2^{*'}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}\boldsymbol{Z}_2^*\boldsymbol{\Omega}_2)^{-1}$, $\boldsymbol{X}_a$ is exposure data specific to $\mathcal{T}_a$, and B_a is the total number of terminal nodes in tree $\mathcal{T}_a$. The shrinkage hyperparameter terms: τ_a^2 and ν^2 can change with different shrinkage choices.

3. Sample $\boldsymbol{\Theta}_a|- \sim \text{MVN}(\boldsymbol{\Sigma}_{\boldsymbol{\Theta}_a}\boldsymbol{X}_a'\boldsymbol{\Sigma}_l^{-1}\boldsymbol{R}_a, \boldsymbol{\Sigma}_{\boldsymbol{\Theta}_a})$.

Step 5: Update r

Finally, the dispersion parameter is sampled with a random walk and MH algorithm. Note that r has a support of $(0, 10)$. Let r_p and r_c denote the proposed and current values of r, respectively.

1. Propose r_p = $\begin{cases} r_{\text{c}} - 1 & \text{by probability } 1/2 \\ r_{\text{c}} + 1 & \text{by probability } 1/2. \end{cases}$
2. Accept r_p with MH ratio:

$$\min\left\{1, \frac{\mathbb{P}(\boldsymbol{y}|r_{\text{p}},\boldsymbol{\nu})\mathbb{P}(r_{\text{c}}|r_{\text{p}})}{\mathbb{P}(\boldsymbol{y}|r_{\text{c}},\boldsymbol{\nu})\mathbb{P}(r_{\text{p}}|r_{\text{c}})}\right\} = \min\left\{1, \prod_{\text{w}_i=1}\left[\frac{\mathbb{P}(y_i|r_{\text{p}},\nu_i)}{\mathbb{P}(y_i|r_{\text{c}},\nu_i)}\right]\right\}.$$

A.3 Degrees of freedom selection for spline-based method

For the competing method of scenario 1 of the simulation, we fit the splines model with different degrees of freedom and choose the one with the best average performance. We report below performance measures: RMSE, confidence interval (CI) coverage, CI width, and precision per degree of freedom. Each boxplot represents a distribution of the performance measures of 100 independent data sets. The numbers on the bottom are averages of the performance measures.

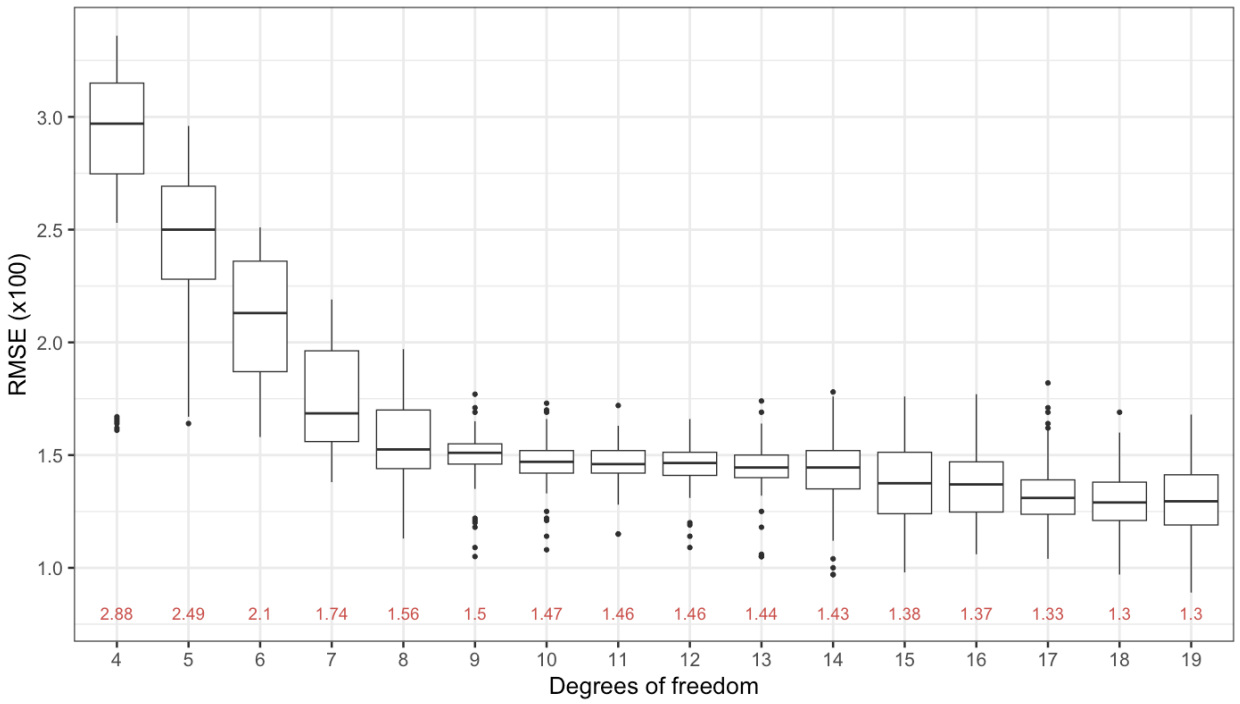


Figure A.5: RMSE

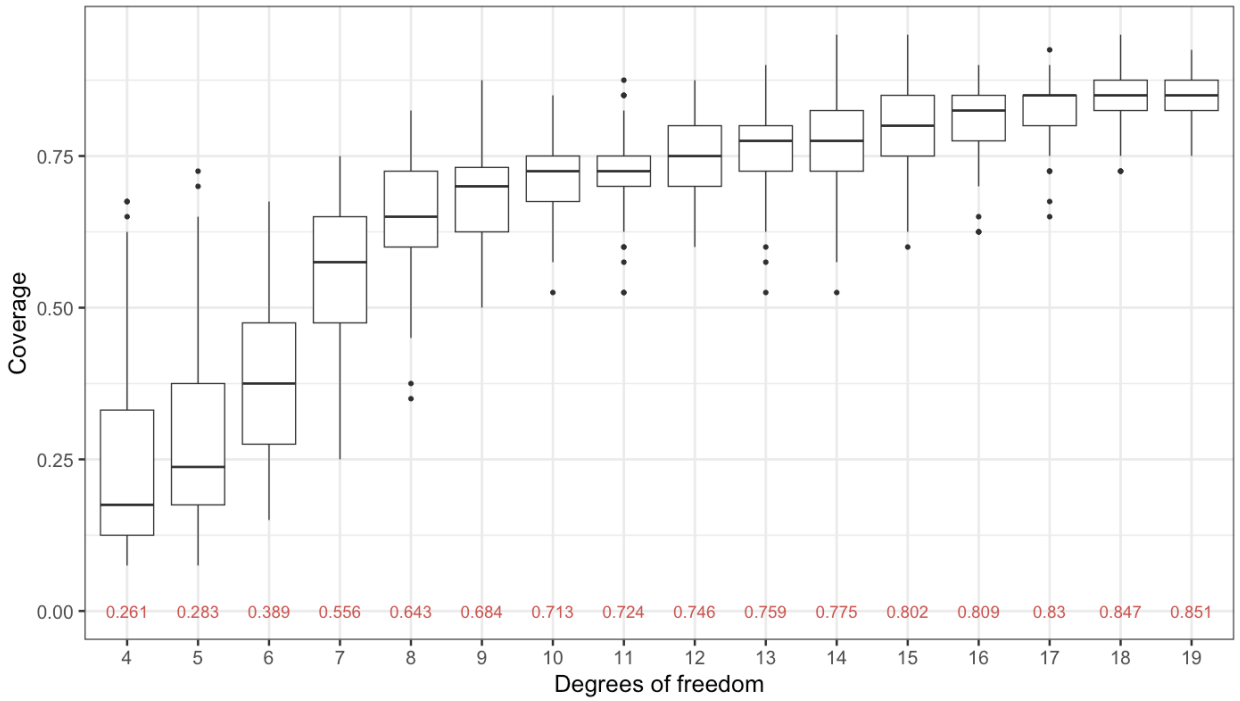


Figure A.6: Coverage

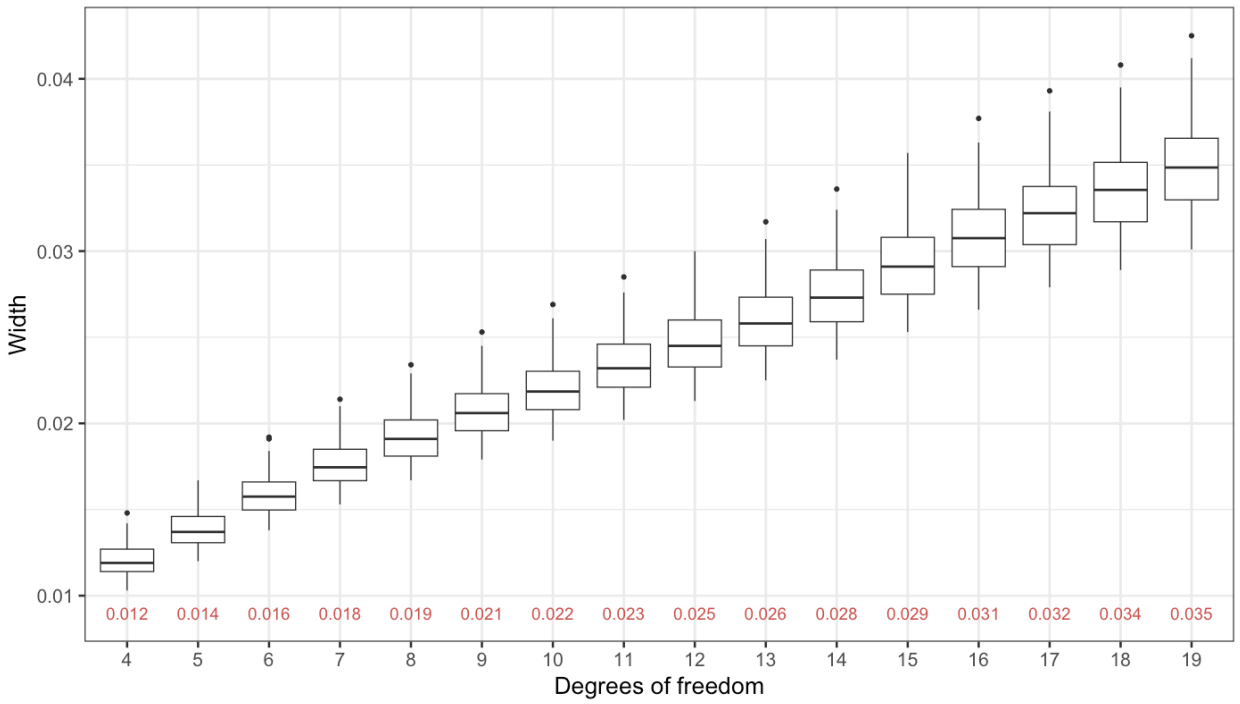


Figure A.7: Confidence interval width

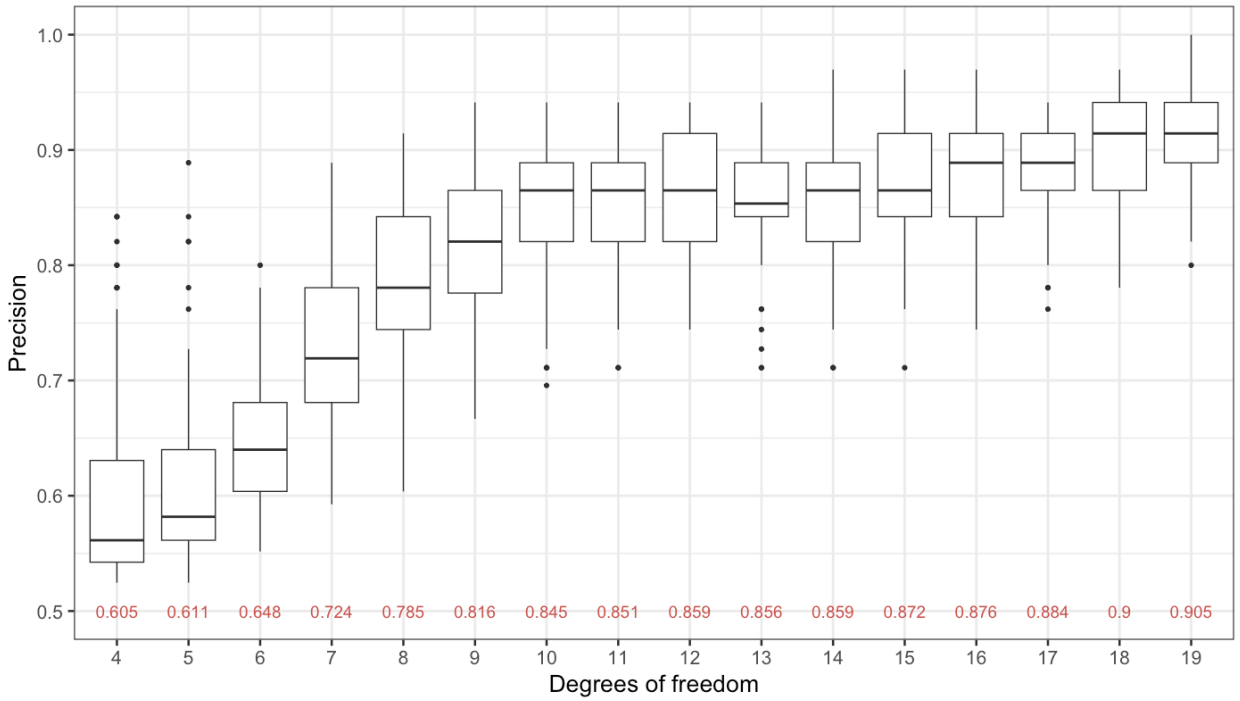


Figure A.8: Precision

Appendix B

Heterogeneous Distributed Lag Mixture Model for Precision Environmental Health With Longitudinally Assessed Mixture Exposures

B.1 Data description

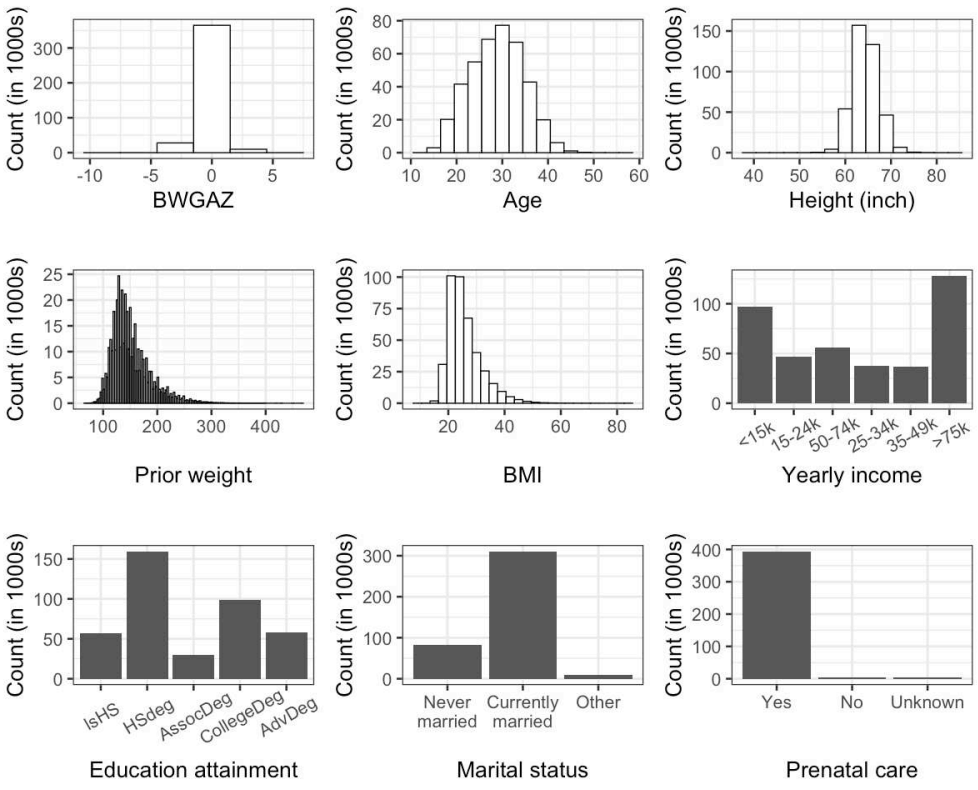


Figure B.1: Descriptive statistics of variables in the Colorado birth cohort data.

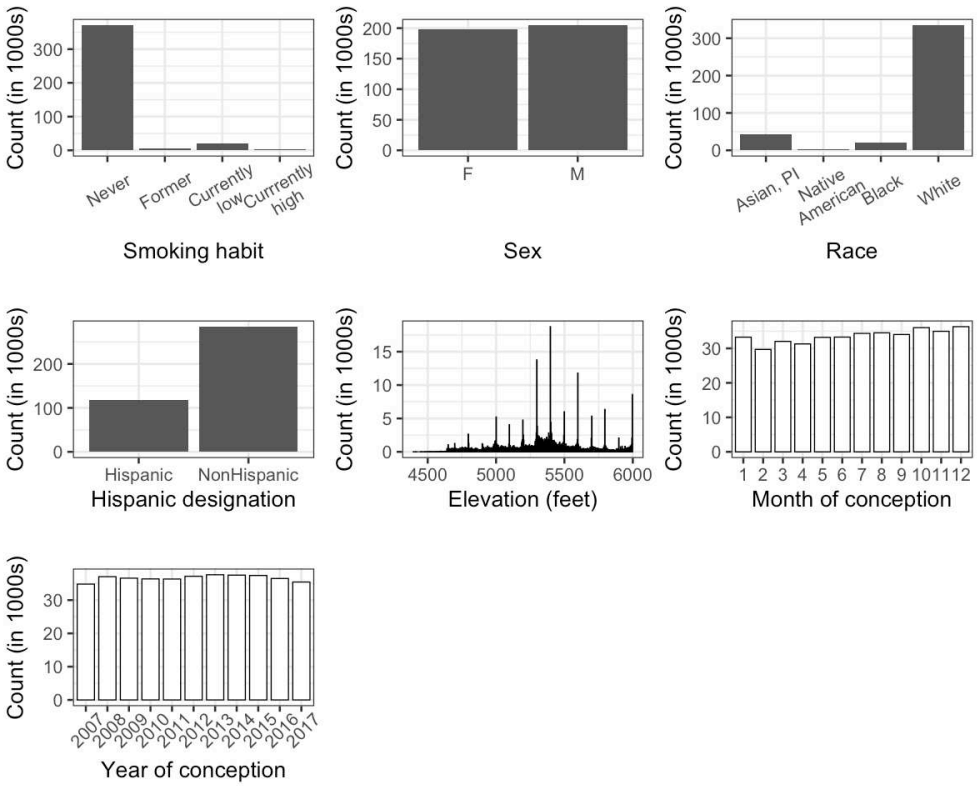


Figure B.2: Descriptive statistics of variables in the Colorado birth cohort data (Continued).

The data contains a set of maternal covariates including pre-pregnancy weight, height, income, education attainment, marital status, prenatal care, race, smoking habits, and Hispanic designations. The data also includes child sex, date of birth, and census tract of maternal residence at birth. Figure B.1 and B.2 show the distributions of the covariates included in the data.

We linked exposure data of $PM_{2.5}$ and temperature to BWGAZ using the maternal residence and estimated date of conception from the administrative birth data. Notably for maximal daily temperature data, we calculated the weekly average of the temperatures of all centroids that overlapped with each census tract.

B.2 Prior specification

B.2.1 Modifier tree structure

The modifier tree requires two elements to impose a prior on its tree structure: splitting probability and splitting rule assignments with modifier selection. The splitting probability is the probability that a terminal node of a tree is split into two child nodes and becomes an internal node. The splitting probability of node η with its depth d is defined as $\mathbb{P}_{split}(\eta) = \alpha(1+d)^{-\beta}$ where $\alpha \in (0,1)$ and $\beta \in [0,\infty)$ control for the tree shape and size, i.e., the number of terminal nodes of the modifier tree. We specify $\alpha = 0.95$ and $\beta = 2$.

For the splitting rule assignment, we define a set of splitting rules $\rho = \{m_l, K\}$ where m_l is a modifier for $l = 1, \ldots, L$ and K represents a splitting rule. If the modifier is continuous, K is an inequality splitting the modifier space into two non-overlapping intervals. For categorical modifiers, K is a subset of the modifier's categories. We establish the prior for ρ of a node η by defining the probability of selecting a modifier from the modifier candidate set and sampling a splitting value or group based on the chosen modifier, that is, $\mathbb{P}(\rho|\eta) = \mathbb{P}(m_l|\eta)\mathbb{P}(K|m_l,\eta)$. A modifier is chosen from the available modifiers, which consists of modifiers whose rules can split the node into two non-empty subgroups. We denote the L-vector of probabilities for the L candidate modifiers included in the split as $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_L)$. Because some candidate modifiers may not be available to split for a given terminal node, we normalize this vector of probabilities among those that can be split on. The probability of a modifier selection is $\mathbb{P}(m_l|\eta) = \psi_l / \sum_{l=1}^{L} \psi_l \mathbb{I}(m_l \in \mathcal{Z})$ where $\mathcal{Z}$ is the available modifiers at node η. Then if m_l is continuous, the splitting value is randomly sampled from the possible splitting values with a uniform probability of $\mathbb{P}(K|m_l,\eta) = 1/(n_{m_l,\eta} - 1)$ where $n_{m_l,\eta}$ is the total number of splitting values. If m_l is categorical, K is sampled with probability $\mathbb{P}(K|m_l,\eta) = 1/(2^{n_{m_l,\eta}} - 1)$ where $n_{m_l,\eta}$ is the number of categories.

B.2.2 DLM tree structure

In the case of the DLM tree structure, we need to specify priors for its splitting probability and component to be assigned to the tree. The splitting probability for each terminal node is defined in the same way as described for the modifier tree. A key distinction from the modifier tree is that the DLM tree divides the total exposure time span rather than partitioning the modifier space. The splitting rule is simplified because the only variable available to split on is exposure time. The rule is, therefore, only a decision on where to split time with the prior being discrete uniform over all potential splitting rules for that node.

B.3 Outline of MCMC algorithm

B.3.1 Initial processing for Bayesian backfitting

The goal of the algorithm is to estimate f in (1) of the main text. We write $f(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iQ}, \boldsymbol{m}_i)$ as $\boldsymbol{f}$ for convenience. We first integrate $\boldsymbol{\gamma}$ out, which gives us

$$\boldsymbol{y}|\boldsymbol{f}, \sigma^2 \sim \text{MVN}(\boldsymbol{f}, \sigma^2 \boldsymbol{V}_Z), \tag{B.1}$$

where $\boldsymbol{V}_Z = (\boldsymbol{I} - \boldsymbol{Z}\boldsymbol{V}_\gamma \boldsymbol{Z}')^{-1}$ and $\boldsymbol{V}_\gamma = (\boldsymbol{Z}'\boldsymbol{Z} + \frac{1}{c}\boldsymbol{I})^{-1}$. Using the derivation from (B.1), we construct a partial residual to employ Bayesian backfitting (Hastie and Tibshirani, 2000a). The partial residual for ath tree triplet is defined as

$$\boldsymbol{R}_a = \boldsymbol{y} - \sum_{\substack{a'=1 \\ a' \neq a}}^{A} g(t|\mathcal{T}_a, \mathcal{D}_a).$$

Since

$$\sum_{\substack{a'=1 \\ a' \neq a}}^{A} g(t|\mathcal{T}_a, \mathcal{D}_a) = \boldsymbol{f} - \boldsymbol{X}_a \boldsymbol{\delta}_a,$$

where $\boldsymbol{X}_a$ is $(n \times T)$ matrix of exposure measurement specific to $\mathcal{T}_a$ and $\boldsymbol{\delta}_a$ is a $(T \times 1)$ vector of lag effects, we know that $\boldsymbol{R}_a \sim \text{MVN}(\boldsymbol{X}_a \boldsymbol{\delta}_a, \sigma^2 \boldsymbol{V}_Z)$ by (B.1).

To separate the update process for the tree triplets into updating tree structures and tree triplet terminal node-specific effects, we marginalize out $\boldsymbol{\delta}_a$ and σ^2. With the shrinkage prior specification on the lag effects, i.e., $\boldsymbol{\delta}_a \sim \text{MVN}(0, \tau_a^2\nu^2\sigma^2\boldsymbol{I})$, we have

$$\begin{aligned}\mathbb{P}(\boldsymbol{R}_a|\mathcal{T}_a, -) &\propto \int_{\sigma^2}\int_{\boldsymbol{\delta}_a} \mathbb{P}(\boldsymbol{R}_a, \boldsymbol{\delta}_a, \sigma^2|\mathcal{T}_a)d\boldsymbol{\delta}_a d\sigma^2 \\ &\propto (\tau_a^2\nu^2)^{-P_a/2}|\boldsymbol{V}_{\delta_a}|^{1/2}\left(\frac{\boldsymbol{R}_a'(\boldsymbol{V}_Z^{-1} - \boldsymbol{V}_Z^{-1}\boldsymbol{X}_a\boldsymbol{V}_{\boldsymbol{\delta}_a}\boldsymbol{X}_a'\boldsymbol{V}_Z^{-1})\boldsymbol{R}_a}{2} + \frac{1}{\xi_\sigma^2}\right)^{-(n+1)/2},\end{aligned} \tag{B.2}$$

where P_a is the total number of terminal nodes of tree triplets and $\boldsymbol{V}_{\boldsymbol{\delta}_a} = \left(\boldsymbol{X}_a'\boldsymbol{V}_Z^{-1}\boldsymbol{X}_a + \frac{1}{\tau_a^2\nu^2}\boldsymbol{I}\right)^{-1}$ (Mork et al., 2023). Then the Metropolis-Hastings (MH) ratio for accepting a proposed modifier tree $\mathcal{M}_a^*$ is

$$r = \min\left\{1, \frac{\mathbb{P}(\mathcal{M}_a^*)\mathbb{P}(\boldsymbol{R}_a|\mathcal{M}_a^*, -)\mathbb{P}(\mathcal{M}_a|\mathcal{M}_a^*)}{\mathbb{P}(\mathcal{M}_a)\mathbb{P}(\boldsymbol{R}_a|\mathcal{M}_a, -)\mathbb{P}(\mathcal{M}_a^*|\mathcal{M}_a)}\right\}. \tag{B.3}$$

Similarly for DLM trees, the MH ratio for accepting a proposed DLM tree $\mathcal{T}_{ap}^*$ is

$$r = \min\left\{1, \frac{\mathbb{P}(\mathcal{T}_{ap}^*)\mathbb{P}(\boldsymbol{R}_a|\mathcal{T}_{ap}^*, -)\mathbb{P}(\mathcal{T}_{ap}|\mathcal{T}_{ap}^*)}{\mathbb{P}(\mathcal{T}_{ap})\mathbb{P}(\boldsymbol{R}_a|\mathcal{T}_{ap}, -)\mathbb{P}(\mathcal{T}_{ap}^*|\mathcal{T}_{ap})}\right\} \qquad \text{for } p = 1, 2. \tag{B.4}$$

B.3.2 Modifier and DLM tree update

We use the MH ratios in (B.3) and (B.4) to update the modifier tree and DLM tree structures. For a modifier tree structure proposal, a new modifier tree is proposed with a transition step chosen from grow (0.25), prune (0.25), change (0.4), and swap (0.1). For each DLM tree, a new DLM tree is proposed with a new DLM tree structure grown from the root node with a specified splitting probability prior and is assigned with a randomly chosen component from $\mathcal{E}$. This proposal approach is advantageous because it simplifies the MH ratio calculation in (B.4) and allows for component selection in a mixture exposure setting.

B.3.3 Algorithm with full conditionals

Before running the HDLMM algorithm, we process the exposure data to have a standard deviation of 1 and scale and center the response variable and the continuous covariates to have a mean zero and standard deviation of 1. The HDLMM algorithm is:

Algorithm 1 HDLMM algorithm

Require: Number of MCMC iterations $N > 0$, number of tree triplets $A > 0$

Initialize parameters with prior distributions

for $s \leftarrow 1; s \leq N; s \leftarrow s + 1$ **do**

  for $a \leftarrow 1; a \leq A; a \leftarrow a + 1$ **do** ▷ Ensemble of tree triplets

    Propose $\mathcal{M}_a^*$ from grow, prune, change, and swap ▷ Modifier tree

    Accept or reject $\mathcal{M}_a^*$ with MH ratio in (B.3)

    for $p = 1, 2$ **do** ▷ DLM trees

      Propose $\mathcal{T}_{ap}^*$ by growing a new tree structure from the root node

      Assign $\mathcal{T}_{ap}^*$ a component randomly sampled from $\mathcal{E} = \{E_1, \ldots, E_Q\}$

      Accept or reject $\mathcal{T}_{ap}^*$ with MH ratio in (B.4)

    end for

    $\boldsymbol{\delta}_a|- \sim \text{MVN}(\boldsymbol{V}_{\delta_a}\boldsymbol{X}_a'\boldsymbol{V}_Z^{-1}\boldsymbol{R}_a, \boldsymbol{V}_{\delta_a})$ ▷ Terminal node effect parameters

    $\tau_a^2|- \sim \text{Inv-Gamma}\Big(\frac{P_a+1}{2}, \frac{\boldsymbol{\delta}_a'\boldsymbol{\delta}_a}{2\nu^2\sigma^2} + \frac{1}{\xi_{\tau_a}}\Big)$ ▷ Tree-specific shrinkage

    $\xi_{\tau_a}|- \sim \text{Inv-Gamma}\Big(1, 1 + \frac{1}{\tau_a^2}\Big)$

  end for

  $\boldsymbol{\gamma}|- \sim \text{MVN}(\boldsymbol{V}_\gamma\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{f}), \sigma^2\boldsymbol{V}_\gamma)$ ▷ Regression coefficients

  $\boldsymbol{\psi}|- \sim \text{Dirichlet}\left(\frac{\kappa}{L} + c_{m_1}, \ldots, \frac{\kappa}{L} + c_{m_L}\right)$ ▷ Modifier selection

    where c_{m_l} is the count of modifier m_l used in modifier trees

  $\mathcal{E}|- \sim \text{Dirichlet}(\phi + e_1, \ldots, \phi + e_Q)$ ▷ Component selection

    where e_q is the count of component q used in DLM trees

  $\nu^2|- \sim \text{Inv-Gamma}\Big(\frac{\sum_{a=1}^A P_a+1}{2} + \frac{\sum_{a=1}^A \boldsymbol{\delta}_a'\boldsymbol{\delta}_a}{2\sigma^2\tau_a^2} + \frac{1}{\xi_\nu}\Big)$ ▷ Global shrinkage

  $\xi_\nu|- \sim \text{Inv-Gamma}\big(1, 1 + \frac{1}{\nu^2}\big)$

  $\sigma^2|- \sim \text{Inv-Gamma}\left(\frac{n+\sum_{a=1}^A P_a+1}{2}, \frac{\sum_{a=1}^A \boldsymbol{\delta}_a'\boldsymbol{\delta}_a}{2\nu^2\tau_a^2} + \frac{||\boldsymbol{V}_Z^{-1/2}(\boldsymbol{y}-\boldsymbol{f})||}{2} + \frac{1}{\xi_\sigma}\right)$

  $\xi_\sigma|- \sim \text{Inv-Gamma}\big(1, 1 + \frac{1}{\sigma^2}\big)$

end for

B.4 Sensitivity analysis

B.4.1 Size of the ensemble

Simulation: Scenario 1

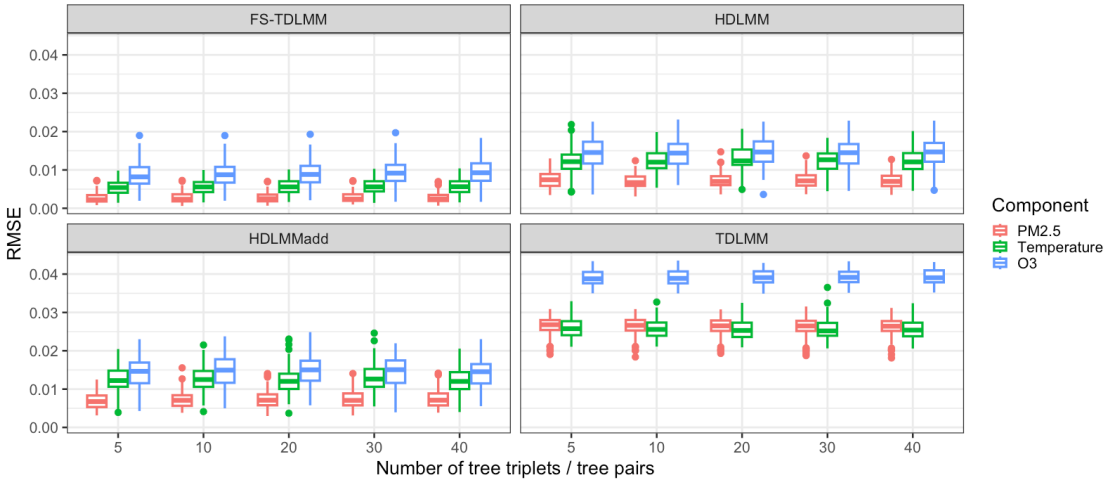


Figure B.3: RMSE for different numbers of tree triplets or tree pairs included in the ensemble.

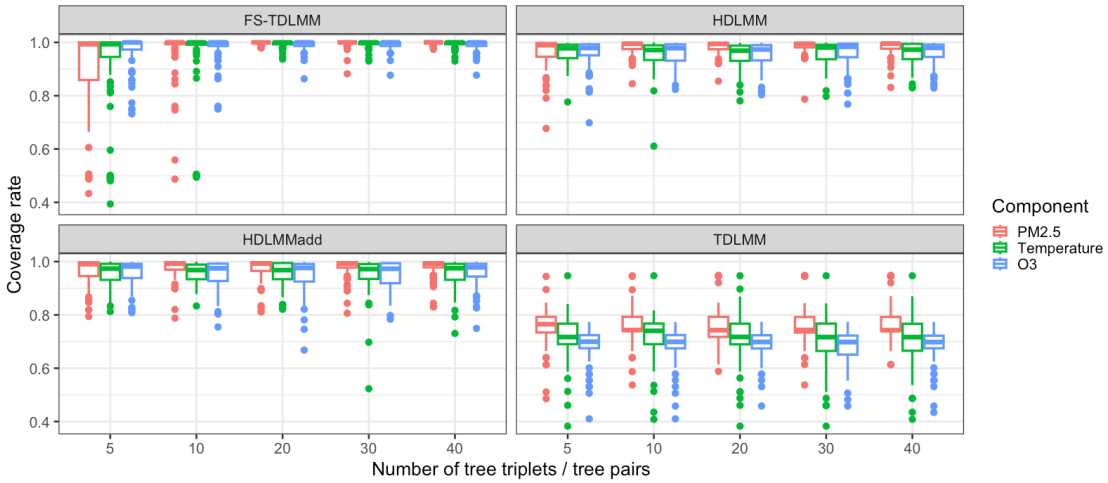


Figure B.4: The coverage rate of 95% credible intervals for different numbers of tree triplets or tree pairs included in the ensemble.

Simulation: Scenario 2

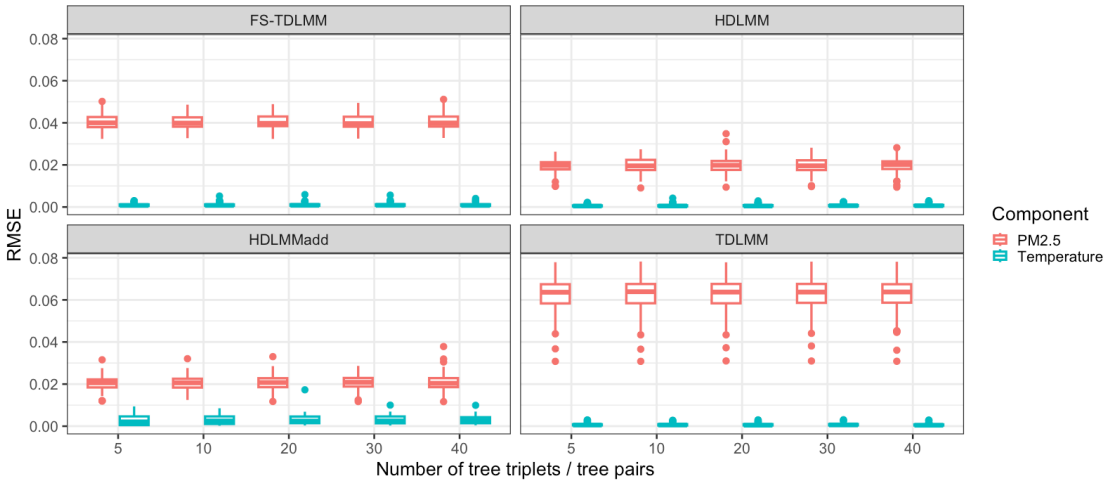


Figure B.5: RMSE for different numbers of tree triplets or tree pairs included in the ensemble.

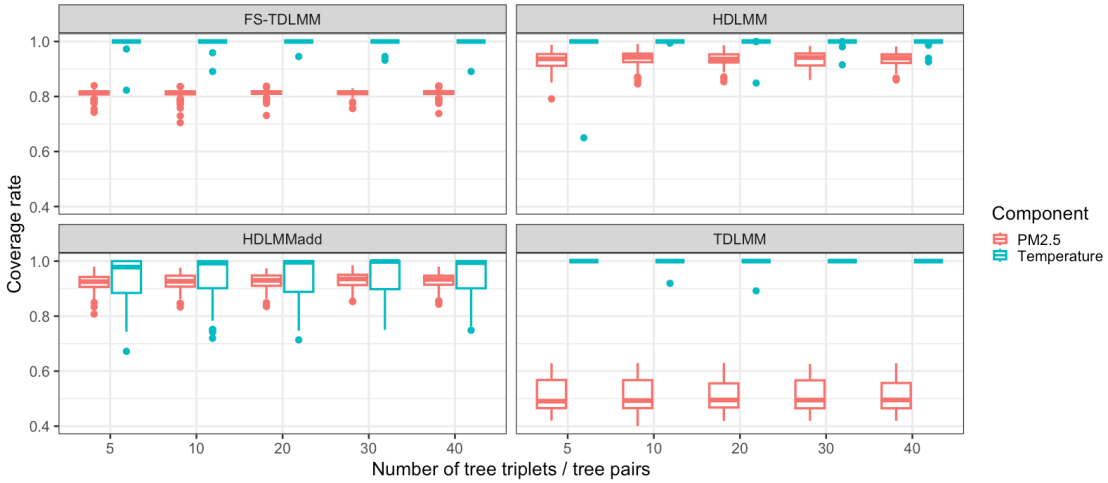


Figure B.6: The coverage rate of 95% credible intervals for different numbers of tree triplets or tree pairs included in the ensemble.

B.4.2 MCMC convergence diagnostics

Simulation: Scenario 1

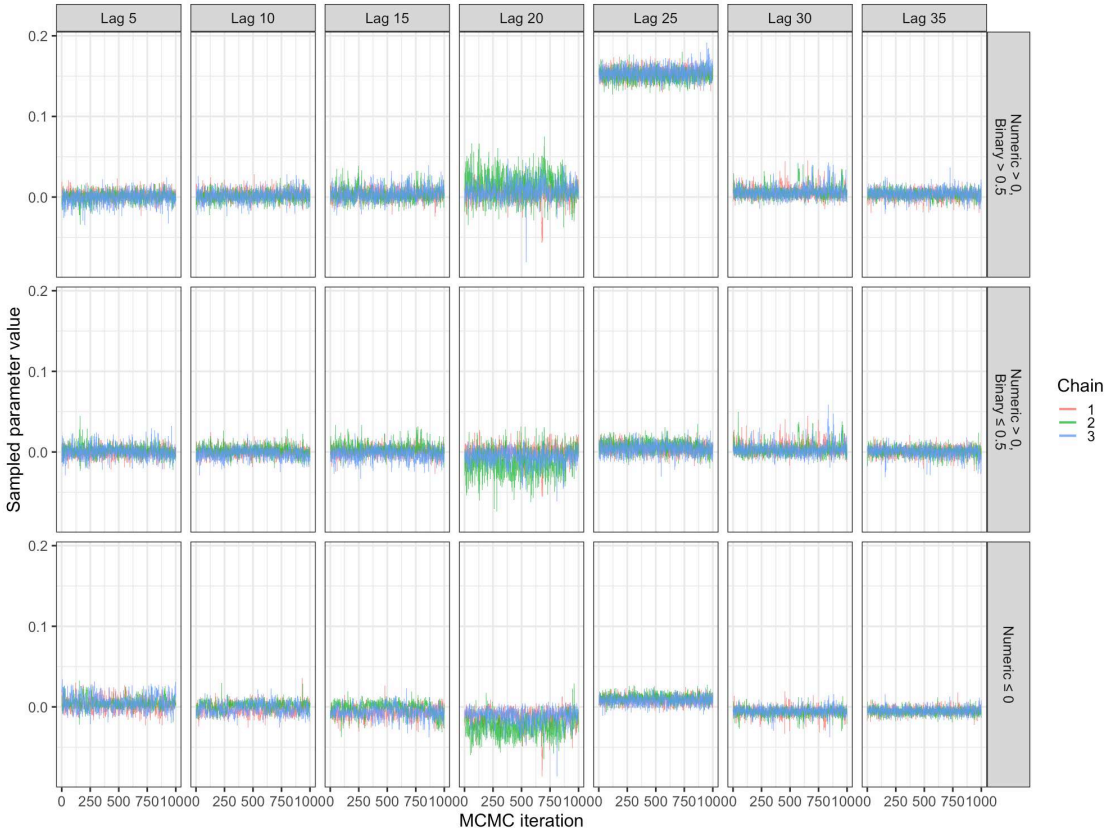


Figure B.7: Traceplots of exposure effect parameter of the first component $\theta_{1t}(\boldsymbol{m}_i)$ of seven selected lags of HDLMMadd fit.

Simulation: Scenario 2

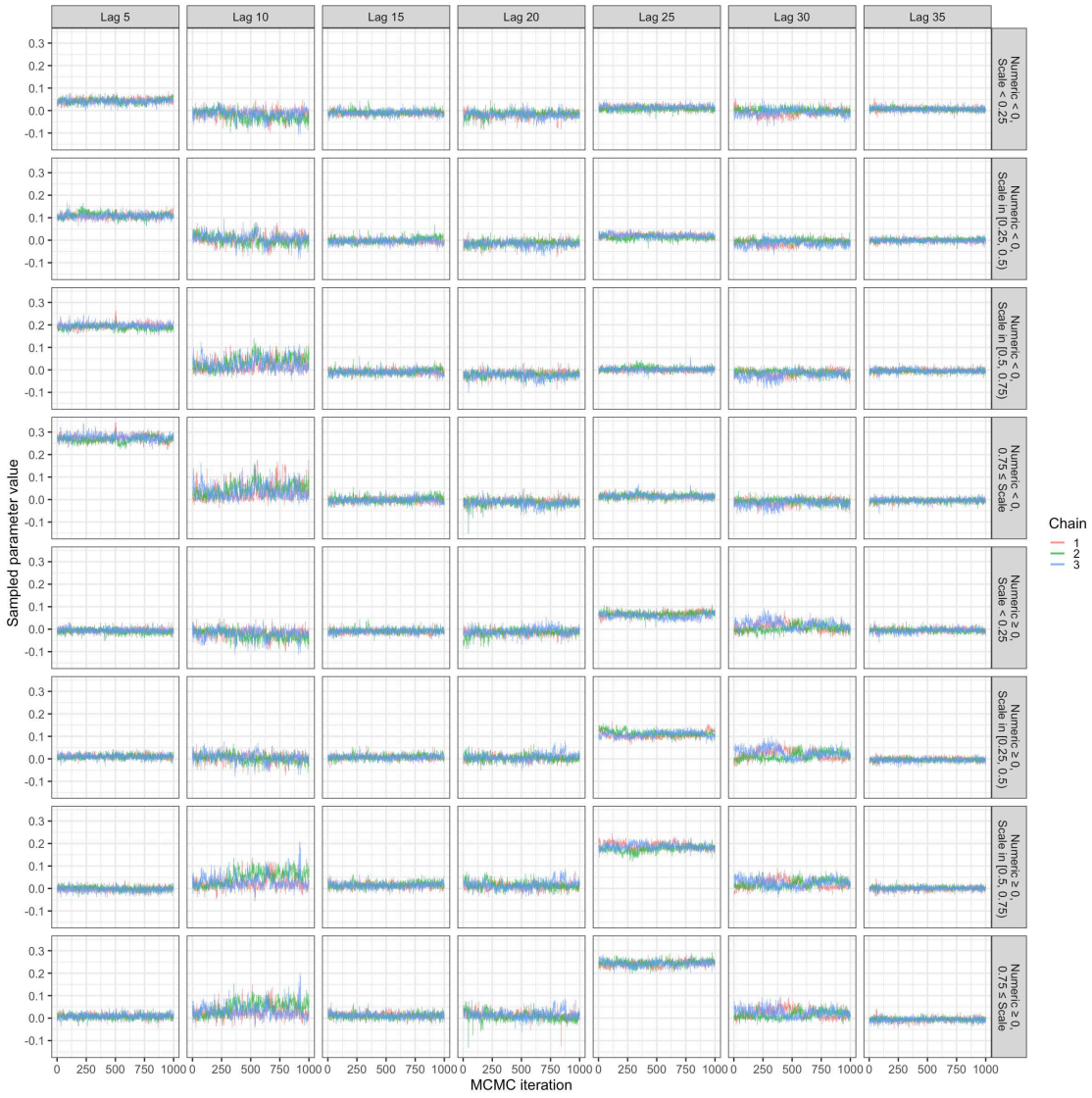


Figure B.8: Traceplots of exposure effect parameters of the first component $\theta_{1t}(\boldsymbol{m}_i)$ of seven selected lags of HDLMM fit.

Data analysis

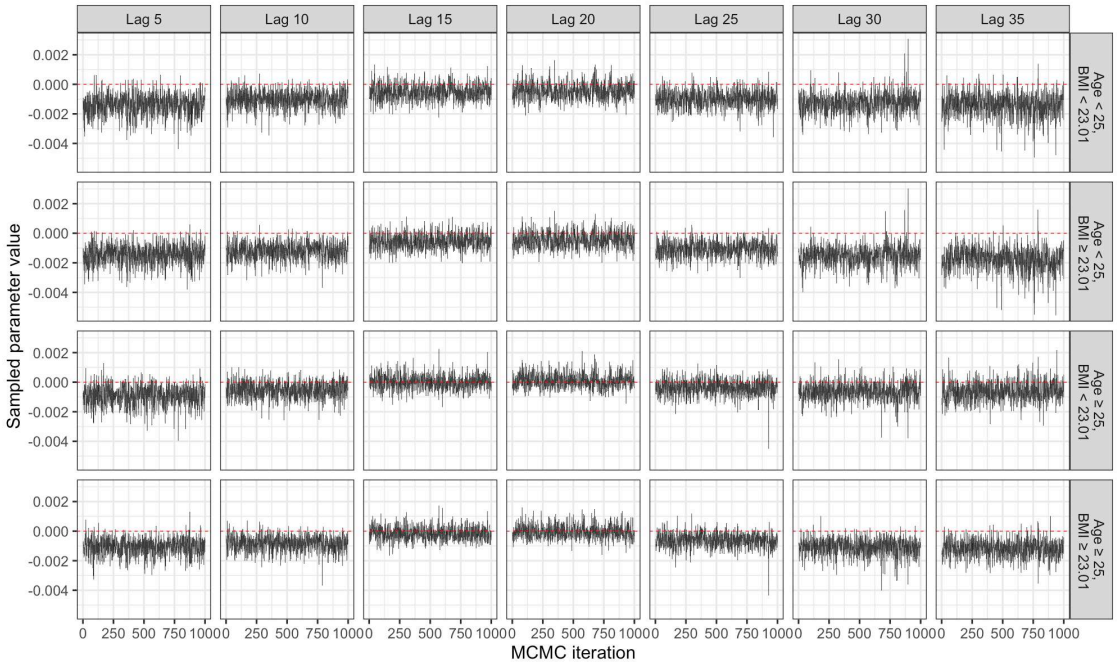


Figure B.9: Traceplots of effect parameters of $PM_{2.5}$ of seven selected lags of HDLMMadd used for data analysis. Each row represents a subgroup grouped by maternal age and BMI. The red line indicates the effect of zero.

Figure B.10: Traceplots of effect parameters of temperature of seven selected lags of HDLMMadd used for data analysis. Each row represents a subgroup grouped by race and Hispanic designation. The red line indicates the effect of zero.

B.5 Performance metrics for simulation studies

For the modifier selection, mPIP is the proportion of MCMC iterations in which a modifier is chosen in a modifier tree at least once in the ensemble of tree triplets. Similarly, for the component selection, cPIP is the proportion of MCMC iterations that a mixture component is assigned to at least one DLM tree in the ensemble of tree triplets.

For exposure effect estimation, we calculated RMSE for each component q to account for the heterogeneous exposure effect on each subgroup. For instance, if observation i is associated with the first component but not the second component, the true exposure effect of the first component $\tilde{\theta}_{i1t}$ is different from that of the second component $\tilde{\theta}_{i2t}$ for all t. We calculated the RMSE of the estimated effect for component q as

$$\text{RMSE}_q = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\sum_{t=1}^{T}(\tilde{\theta}_{iqt} - \hat{\tilde{\theta}}_{iqt})^2/T}, \tag{B.5}$$

where $\tilde{\theta}_{iqt}$ and $\hat{\tilde{\theta}}_{iqt}$ denote the true and estimated exposure effect at time t of component q of observation i, respectively. Similarly, we define the 95% CrI coverage for component q as

$$\text{Coverage}_q = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{I}\left\{\hat{\tilde{\theta}}_{iqt}^{(0.025)} < \tilde{\theta}_{iqt} < \hat{\tilde{\theta}}_{iqt}^{(0.975)}\right\}\right), \tag{B.6}$$

where $\hat{\tilde{\theta}}_{iqt}^{(0.025)}$ and $\hat{\tilde{\theta}}_{iqt}^{(0.975)}$ are 2.5th and 97.5th percentiles of the MCMC posterior samples of $\hat{\tilde{\theta}}_{iqt}$.

For the window of susceptibility identification, we calculated TP, FP, and precision. We define windows of susceptibility as time points where the credible interval for the time-component-specific parameters does not contain zero. We considered a total number of windows of susceptibility for each observation i over all components and defined combined TP and FP as

$$\text{TP} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\sum_{q=1}^{Q}\sum_{t=1}^{T}\mathbb{I}\left\{\hat{\tilde{\theta}}_{iqt}^{(0.025)} > 0, \tilde{\theta}_{iqt} > 0 \vee \hat{\tilde{\theta}}_{iqt}^{(0.975)} < 0, \tilde{\theta}_{iqt} < 0\right\}}{\sum_{q=1}^{Q}\sum_{t=1}^{T}\mathbb{I}\left\{\tilde{\theta}_{iqt} \neq 0\right\}}\right) \tag{B.7}$$

$$\text{FP} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\sum_{q=1}^{Q}\sum_{t=1}^{T}\mathbb{I}\left\{\hat{\tilde{\theta}}_{iqt}^{(0.025)} > 0 \vee \hat{\tilde{\theta}}_{iqt}^{(0.975)} < 0, \tilde{\theta}_{iqt} = 0\right\}}{\sum_{q=1}^{Q}\sum_{t=1}^{T}\mathbb{I}\left\{\tilde{\theta}_{iqt} = 0\right\}}\right), \tag{B.8}$$

and we calculated the combined precision as TP / (FP + TP). We obtained MSPE using 10-fold cross-validation. We calculated MSPE as $n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is a predicted outcome, per fold and averaged across folds for each simulated dataset.

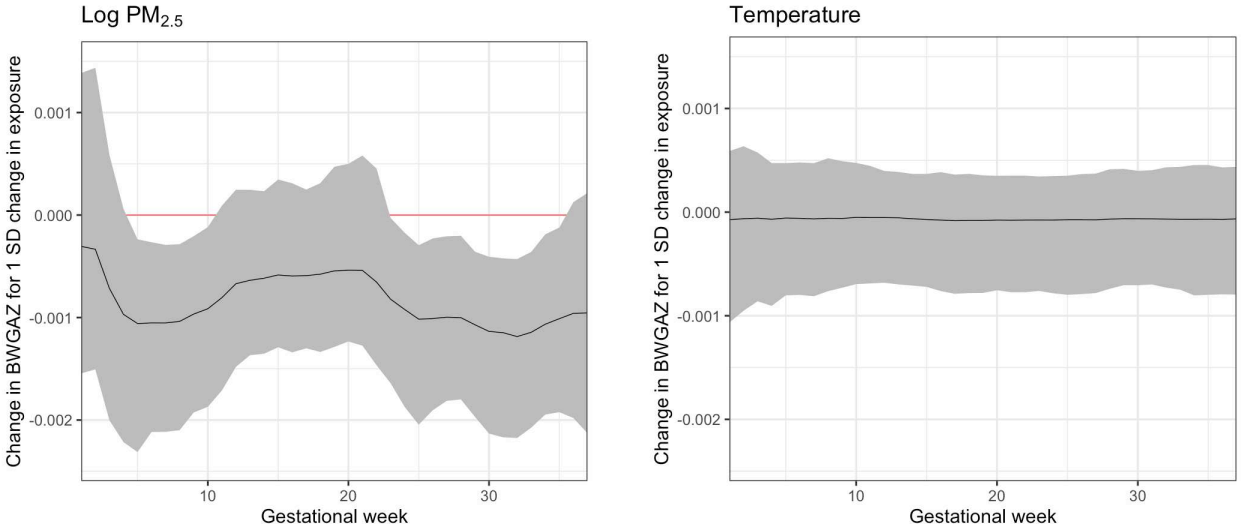
B.6 Additional information for data analysis

B.6.1 Variables included in the data analysis

Table B.1: Description and mPIPs of variables included in the data analysis

Variable	Type	Mean	IQR	Category #	Modifier	Covariate	mPIP
BMI	Continuous	25.8	7.0		✓	✓	1.000
Height (inch)	Continuous	64.4	4.0			✓	
Maternal age	Continuous	28.8	8.0		✓	✓	0.998
Prior weight (lbs)	Continuous	151.8	43.0			✓	
Elevation (feet)	Continuous	5353.7	413.4			✓	
Child sex	Binary			2	✓		0.551
Hispanic designation	Binary			2	✓	✓	1.000
Marital status	Categorical			3	✓	✓	0.511
Prenatal care	Categorical			3	✓	✓	0.546
Smoking habits	Categorical			4	✓	✓	0.551
Maternal residence	Categorical			13		✓	
Race	Categorical			4	✓	✓	0.968
Yearly income	Ordinal			6	✓	✓	0.654
Education attainment	Ordinal			5	✓	✓	0.600
Month of conception	Nominal			12		✓	
Year of conception	Nominal			11		✓	

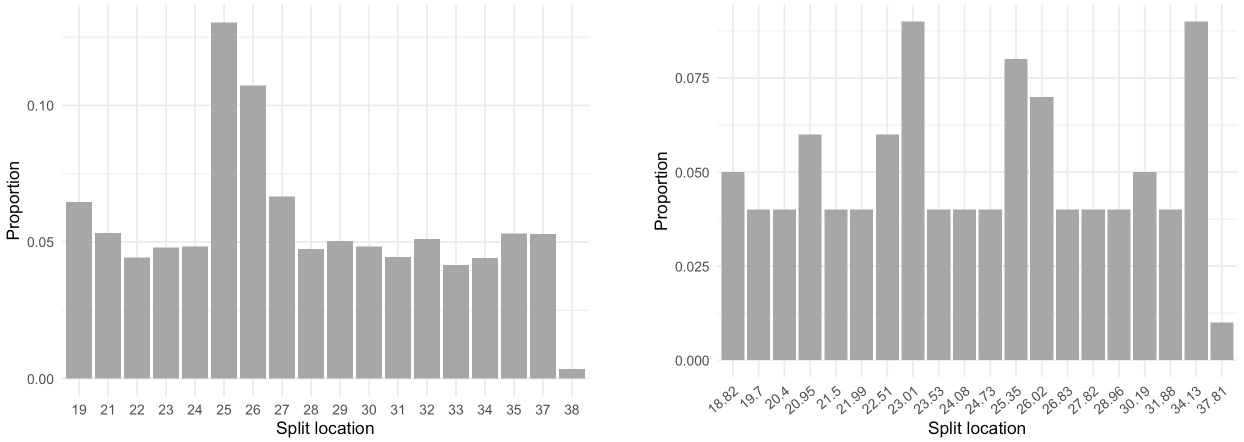
B.6.2 Preliminary analysis assuming no heterogeneity



(a) $PM_{2.5}$ **(b)** Temperature

Figure B.11: Estimated distributed lag effect for each component using the TDLMM method with no heterogeneity after marginalizing out the other co-exposures. The gray area shows the 95% credible interval of the effect. The red line indicates the effect of zero.

B.6.3 Determining split points



(a) Maternal age

(b) BMI

Figure B.12: Proportions of split points of maternal age and BMI used in the modifier tree splitting rules for HDLMMadd.

B.6.4 Additional results of data analysis

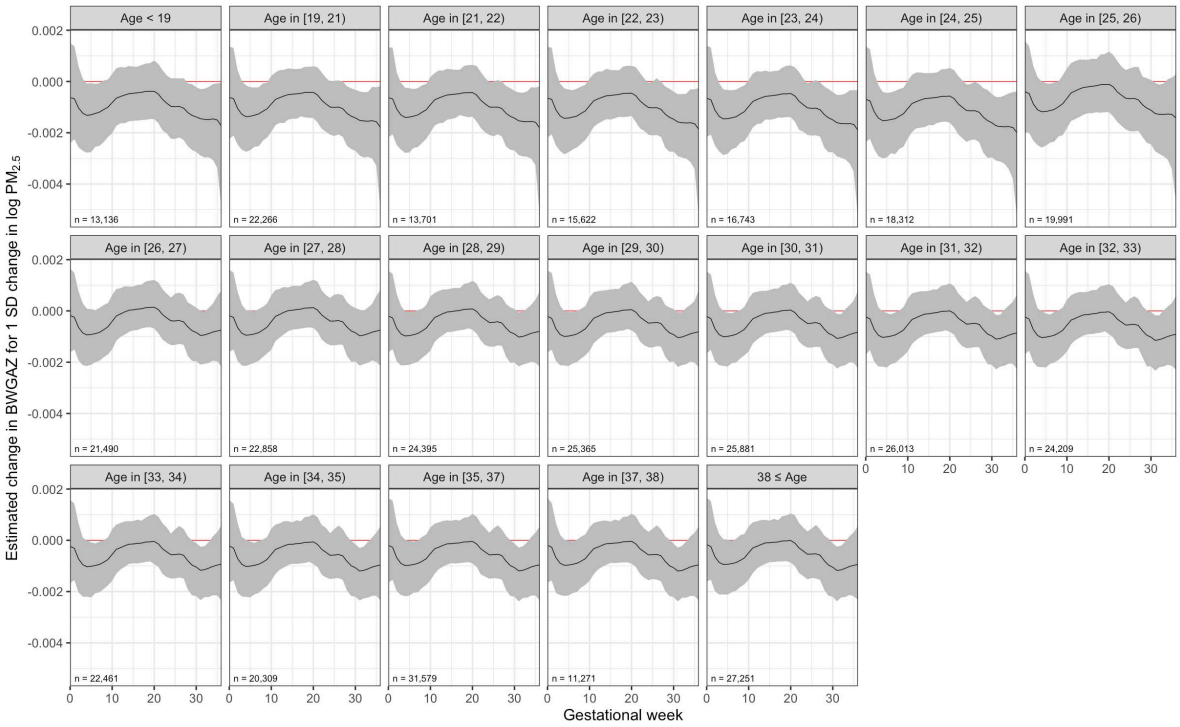


Figure B.13: Estimated GATEs of $PM_{2.5}$ with HDLMMadd for 19 subgroups grouped by maternal age.

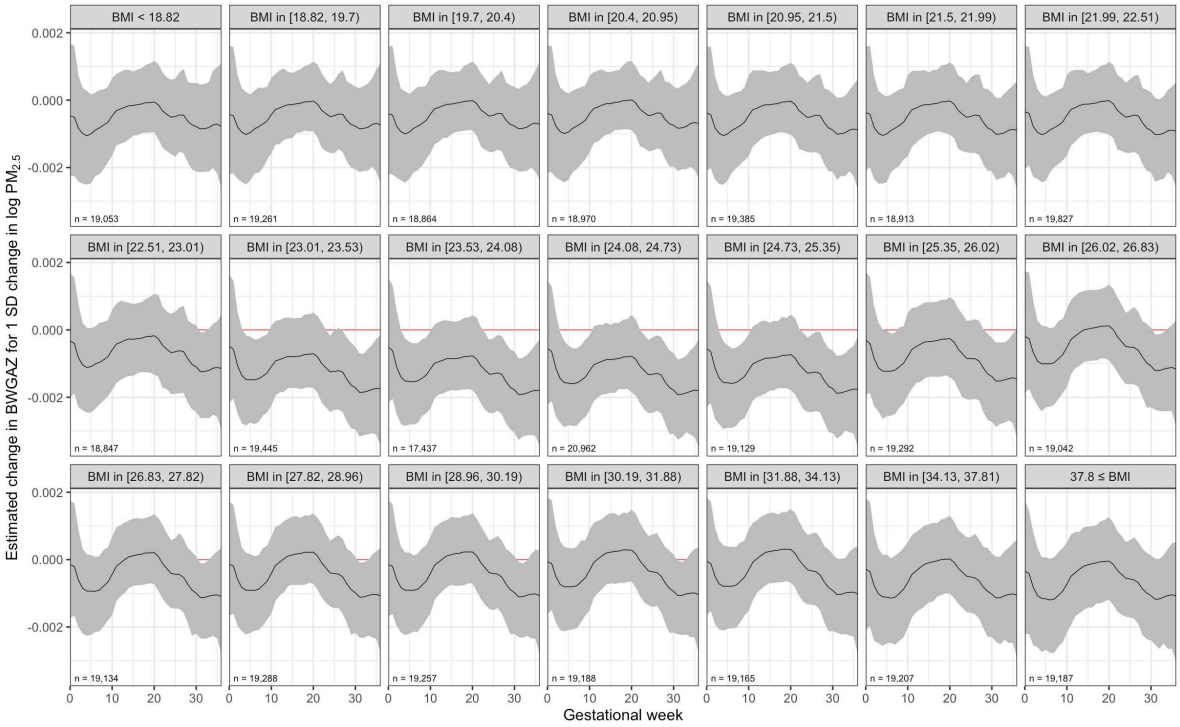
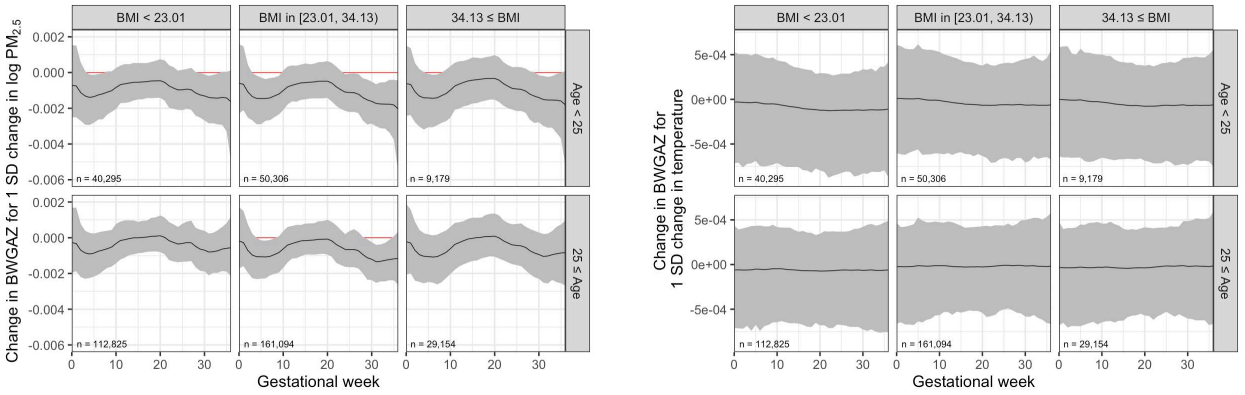
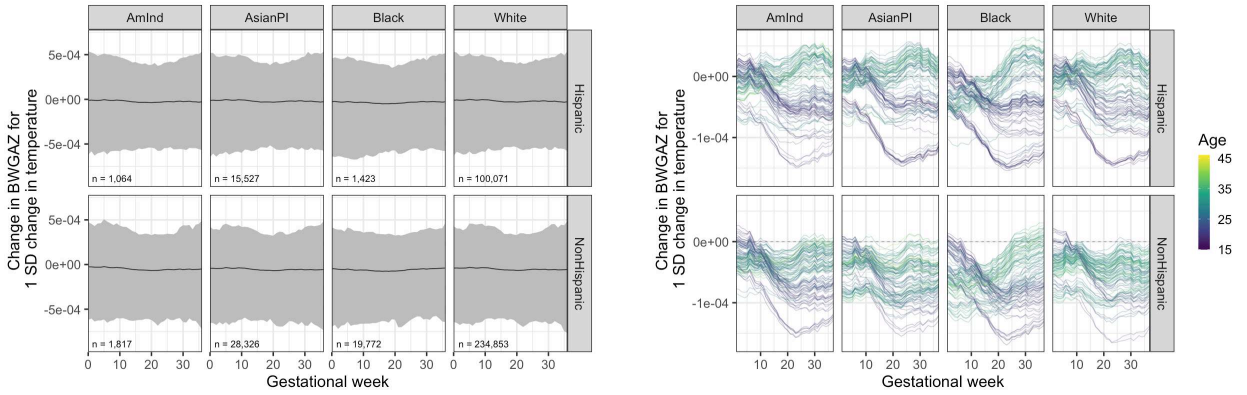


Figure B.14: Estimated GATEs of $PM_{2.5}$ with HDLMMadd for 21 subgroups grouped by mother's BMI.



(a) $PM_{2.5}$ **(b)** Temperature

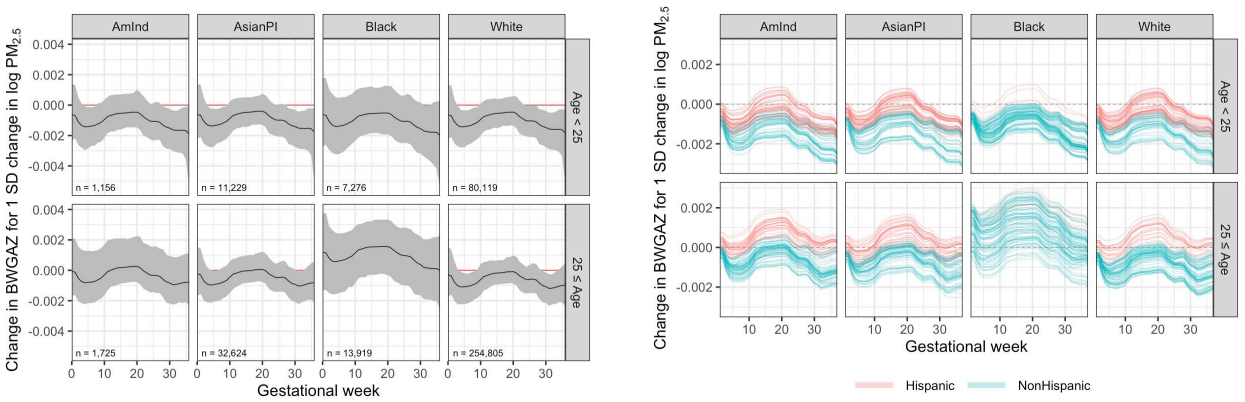
Figure B.15: Estimated GATEs of $PM_{2.5}$ and temperature with HDLMMadd grouped by maternal age and BMI. The gray area shows the 95% credible interval for each effect. The sample size of each subgroup is indicated in the bottom left corner.



(a) Subgroup effect

(b) Individual effect

Figure B.16: Estimated distributed lag effects of maximal daily temperature with HDLMMadd grouped by race and the Hispanic designation. Panel (a) shows the GATEs of temperature where the gray area shows the 95% credible interval for each effect. The sample size of each subgroup is indicated in the bottom left corner. Panel (b) shows the CATEs where each line is colored with the maternal age. Each subgroup includes 100 randomly sampled mothers.



(a) Subgroup effect

(b) Individual effect

Figure B.17: Estimated distributed lag effects of $PM_{2.5}$ with HDLMMadd grouped by race and age. Panel (a) shows the GATEs of $PM_{2.5}$ where the gray area indicates the 95% credible interval. The sample size of each subgroup is indicated in the bottom left corner. Panel (b) shows the CATEs where each line is colored by the mother's Hispanic designation. Each subgroup includes 100 randomly sampled mothers.

Appendix C

Structured Bayesian Regression Tree Models for Estimating Distributed Lag Effects: The R Package dlmtree

C.1 Syntax for TDLM

This section demonstrates the implementation of treed distributed lag model (TDLM). More details can be found in Mork and Wilson (2023).

```
library(dlmtree)
library(dplyr)
set.seed(1)
```

Load data

Simulated data is available on GitHub. It can be loaded with the following code.

```
sbd_dlmtree <- get_sbd_dlmtree()
```

Data preparation

```
# Response and covariates
sbd_cov <- sbd_dlmtree %>%
            select(bwgaz, ChildSex, MomAge, GestAge, MomPriorBMI, Race,
                   Hispanic, MomEdu, SmkAny, Marital, Income,
                   EstDateConcept, EstMonthConcept, EstYearConcept)

# Exposure data
sbd_exp <- list(PM25 = sbd_dlmtree %>% select(starts_with("pm25_")),
                TEMP = sbd_dlmtree %>% select(starts_with("temp_")),
                SO2 = sbd_dlmtree %>% select(starts_with("so2_")),
                CO = sbd_dlmtree %>% select(starts_with("co_")),
                NO2 = sbd_dlmtree %>% select(starts_with("no2_")))
sbd_exp <- sbd_exp %>% lapply(as.matrix)
```

Fitting the model

```
tdlm.fit <- dlmtree(formula = bwgaz ~ ChildSex + MomAge + MomPriorBMI +
                                      Race + Hispanic + SmkAny + EstMonthConcept,
                    data = sbd_cov,
                    exposure.data = sbd_exp[["PM25"]], # A single numeric matrix
                    family = "gaussian", dlm.type = "linear",
                    n.burn = 2500, n.iter = 10000, n.thin = 5)
```

```
#> Preparing data...
#>
#> Running TDLM:
#> Burn-in % complete
#> [0--------25--------50--------75--------100]
#>  ''''''''''''''''''''''''''''''''''''''''''
#> MCMC iterations (est time: 24 seconds)
#> [0--------25--------50--------75--------100]
#>  ''''''''''''''''''''''''''''''''''''''''''
#> Compiling results...
```

Model fit summary

```
tdlm.sum <- summary(tdlm.fit)
tdlm.sum
```

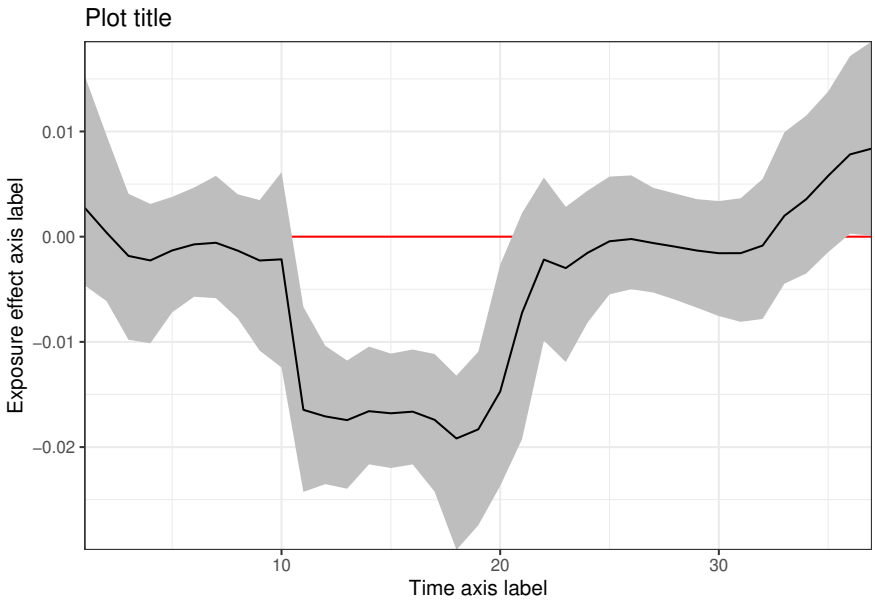
```
#> ---
#> TDLM summary
#>
#> Model run info:
#> - bwgaz ~ ChildSex + MomAge + MomPriorBMI + Race + Hispanic + SmkAny + EstMonthConcept
#> - sample size: 10,000
#> - family: gaussian
#> - 20 trees
#> - 2500 burn-in iterations
#> - 10000 post-burn iterations
#> - 5 thinning factor
#> - 0.95 confidence level
#>
#> Fixed effect coefficients:
#>                         Mean  Lower  Upper
#> *(Intercept)           2.289  2.032  2.542
#> *ChildSexM            -2.105 -2.126 -2.085
#> MomAge                 0.000 -0.001  0.002
```

```
#> *MomPriorBMI         -0.021 -0.022 -0.019
#> RaceAsianPI           0.069 -0.057  0.192
#> RaceBlack             0.078 -0.050  0.205
#> Racewhite             0.059 -0.060  0.181
#> *HispanicNonHispanic  0.255  0.233  0.278
#> *SmkAnyY             -0.403 -0.451 -0.356
#> EstMonthConcept2     -0.049 -0.109  0.010
#> *EstMonthConcept3    -0.145 -0.211 -0.077
#> *EstMonthConcept4    -0.230 -0.295 -0.160
#> *EstMonthConcept5    -0.207 -0.265 -0.147
#> *EstMonthConcept6    -0.205 -0.260 -0.153
#> EstMonthConcept7     -0.032 -0.083  0.023
#> *EstMonthConcept8     0.145  0.081  0.210
#> *EstMonthConcept9     0.393  0.326  0.460
#> *EstMonthConcept10    0.372  0.311  0.437
#> *EstMonthConcept11    0.330  0.271  0.387
#> *EstMonthConcept12    0.129  0.078  0.181
#> ---
#> * = CI does not contain zero
#>
#> DLM effect:
#> range = [-0.019, 0.008]
#> signal-to-noise = 0.021
#> critical windows: 11-20,36-37
#>              Mean  Lower  Upper
#> Period 1    0.003 -0.005  0.015
#> Period 2    0.000 -0.006  0.010
#> Period 3   -0.002 -0.010  0.004
#> Period 4   -0.002 -0.010  0.003
#> Period 5   -0.001 -0.007  0.004
#> Period 6   -0.001 -0.006  0.005
#> Period 7   -0.001 -0.006  0.006
#> Period 8   -0.001 -0.008  0.004
#> Period 9   -0.002 -0.011  0.003
#> Period 10  -0.002 -0.012  0.006
#> *Period 11 -0.016 -0.024 -0.007
#> *Period 12 -0.017 -0.024 -0.010
#> *Period 13 -0.017 -0.024 -0.012
#> *Period 14 -0.017 -0.022 -0.010
#> *Period 15 -0.017 -0.022 -0.011
#> *Period 16 -0.017 -0.022 -0.011
#> *Period 17 -0.017 -0.024 -0.011
#> *Period 18 -0.019 -0.030 -0.013
#> *Period 19 -0.018 -0.027 -0.011
#> *Period 20 -0.015 -0.024 -0.003
```

```
#> Period 21  -0.007 -0.019  0.002
#> Period 22  -0.002 -0.010  0.006
#> Period 23  -0.003 -0.012  0.003
#> Period 24  -0.002 -0.008  0.004
#> Period 25   0.000 -0.005  0.006
#> Period 26   0.000 -0.005  0.006
#> Period 27  -0.001 -0.005  0.005
#> Period 28  -0.001 -0.006  0.004
#> Period 29  -0.001 -0.007  0.004
#> Period 30  -0.002 -0.008  0.003
#> Period 31  -0.002 -0.008  0.004
#> Period 32  -0.001 -0.008  0.005
#> Period 33   0.002 -0.004  0.010
#> Period 34   0.004 -0.003  0.012
#> Period 35   0.006 -0.001  0.014
#> *Period 36  0.008  0.000  0.017
#> *Period 37  0.008  0.000  0.019
#> ---
#> * = CI does not contain zero
#>
#> residual standard errors: 0.004
#> ---
```

Exposure effect

```
plot(tdlm.sum,
     main = "Plot title",
     xlab = "Time axis label",
     ylab = "Exposure effect axis label")
```



C.2 Syntax for TDLNM

This section demonstrates the implementation of treed distributed lag non-linear model (TDLNM). More details can be found in Mork and Wilson (2022).

Fitting the model

```
tdlnm.fit <- dlmtree(formula = bwgaz ~ ChildSex + MomAge + MomPriorBMI +
                                       Race + Hispanic + SmkAny + EstMonthConcept,
                     data = sbd_cov,
                     exposure.data = sbd_exp[["TEMP"]],
                     dlm.type = "nonlinear",
                     family = "gaussian",
                     tdlnm.exposure.splits = 20,
                     n.burn = 2500, n.iter = 10000, n.thin = 5)
#> Preparing data...
#>
#> Running TDLNM:
#> Burn-in % complete
#> [0--------25--------50--------75--------100]
#>  ''''''''''''''''''''''''''''''''''''''''''
#> MCMC iterations (est time: 28 seconds)
#> [0--------25--------50--------75--------100]
#>  ''''''''''''''''''''''''''''''''''''''''''
#> Compiling results...
```

Model fit summary

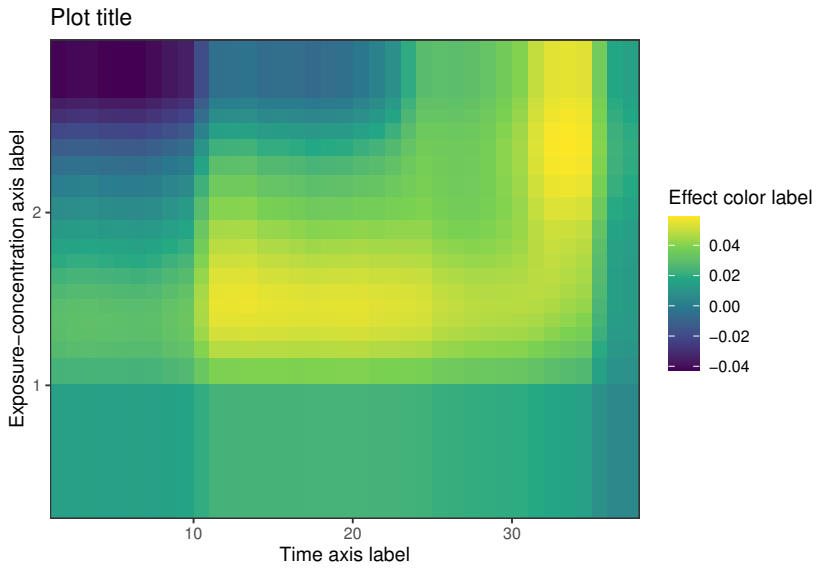
```
tdlnm.sum <- summary(tdlnm.fit)
#> Centered DLNM at exposure value 0
```

```
tdlnm.sum
#> ---
#> TDLNM summary
#>
#> Model run info:
#> - bwgaz ~ ChildSex + MomAge + MomPriorBMI + Race + Hispanic + SmkAny + EstMonthConcept
#> - sample size: 10,000
#> - family: gaussian
#> - 20 trees
#> - 2500 burn-in iterations
#> - 10000 post-burn iterations
#> - 5 thinning factor
#> - 0.95 confidence level
#>
#> Fixed effect coefficients:
#>                          Mean  Lower  Upper
#> (Intercept)             0.160 -0.913  1.180
#> *ChildSexM             -2.105 -2.127 -2.086
#> MomAge                  0.001 -0.001  0.002
#> *MomPriorBMI           -0.021 -0.023 -0.019
#> RaceAsianPI             0.026 -0.103  0.152
#> RaceBlack               0.031 -0.094  0.157
#> Racewhite               0.012 -0.112  0.133
#> *HispanicNonHispanic    0.256  0.233  0.278
#> *SmkAnyY               -0.396 -0.442 -0.351
#> *EstMonthConcept2       0.117  0.035  0.202
#> *EstMonthConcept3       0.231  0.100  0.356
#> *EstMonthConcept4       0.369  0.201  0.533
#> *EstMonthConcept5       0.498  0.318  0.681
#> *EstMonthConcept6       0.452  0.275  0.639
#> *EstMonthConcept7       0.387  0.214  0.573
#> *EstMonthConcept8       0.238  0.072  0.420
#> *EstMonthConcept9       0.264  0.099  0.437
#> *EstMonthConcept10      0.159  0.014  0.308
#> *EstMonthConcept11      0.127  0.013  0.237
#> EstMonthConcept12       0.020 -0.054  0.094
#> ---
#> * = CI does not contain zero
#>
#> DLNM effect:
#> range = [-0.042, 0.059]
```

```
#> signal-to-noise = 0.405
#> critical windows: 4-6,10-34
#>
#> residual standard errors: 0.004
```

Exposure-time surface

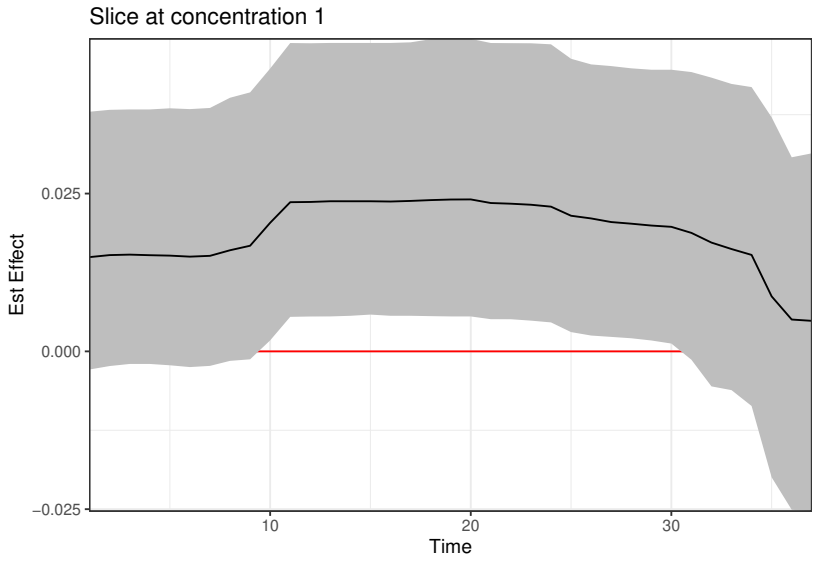
```
plot(tdlnm.sum,
     main = "Plot title",
     xlab = "Time axis label",
     ylab = "Exposure-concentration axis label",
     flab = "Effect color label")
```

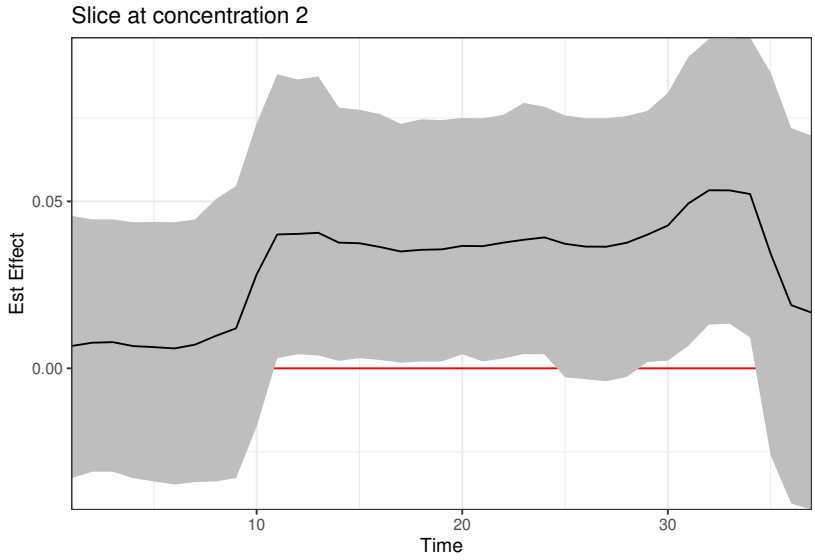


Slicing on exposure-concentration

```
# slicing on exposure-concentration
plot(tdlnm.sum, plot.type = "slice", val = 1, main = "Slice at concentration 1")
```

```
plot(tdlnm.sum, plot.type = "slice", val = 2, main = "Slice at concentration 2")
```

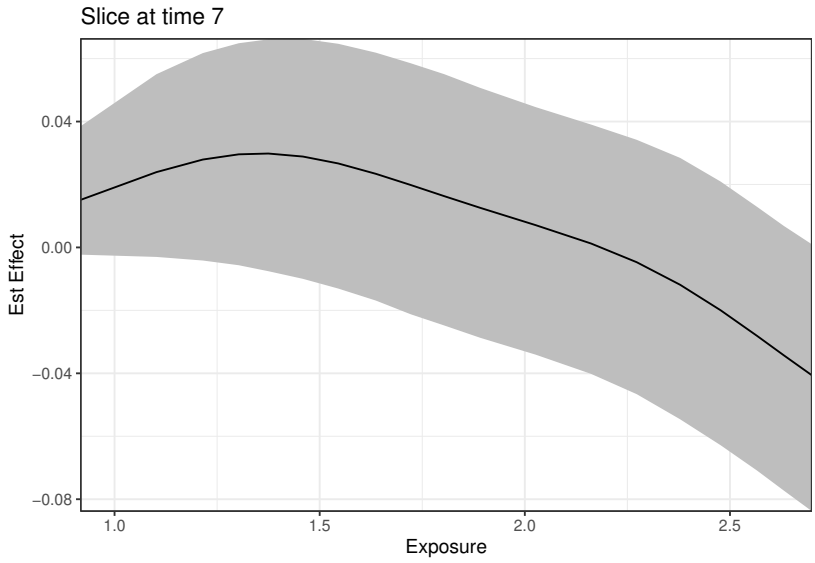


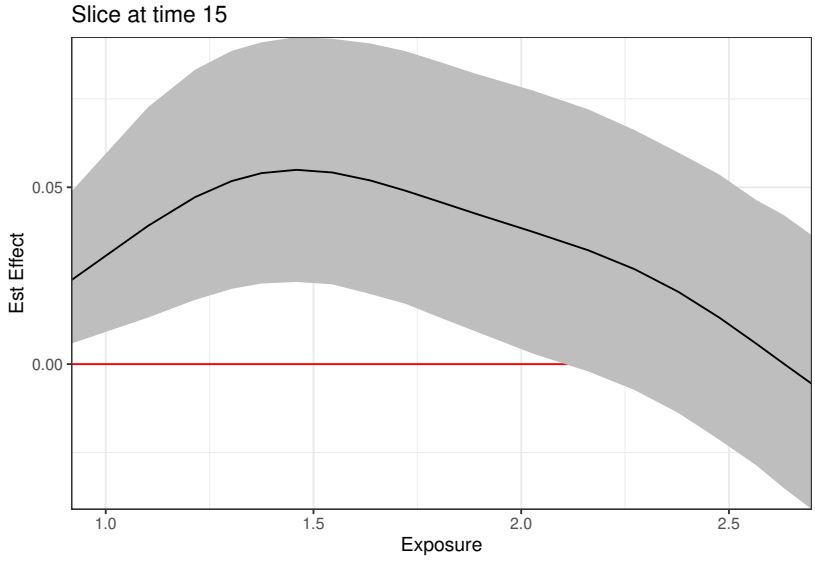


Slicing on time lag

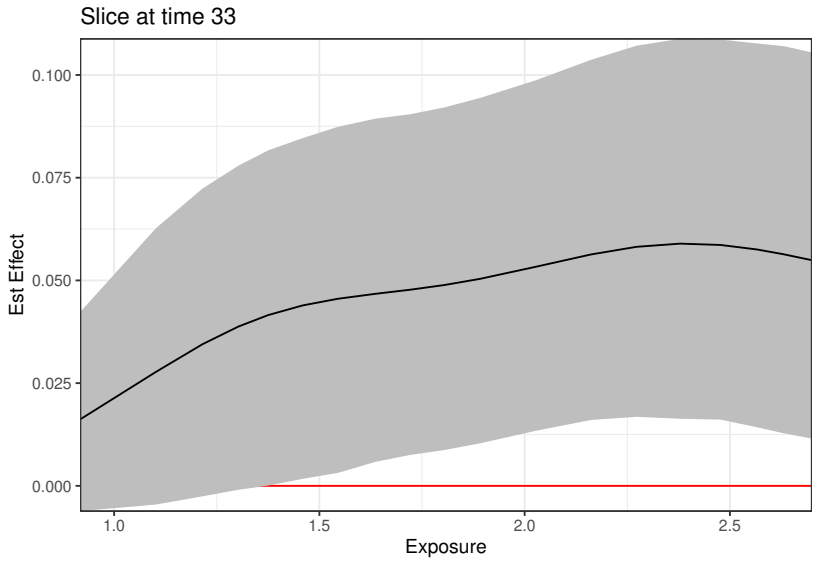
```
# slicing on exposure-concentration
plot(tdlnm.sum, plot.type = "slice", time = 7, main = "Slice at time 7")
```

```
plot(tdlnm.sum, plot.type = "slice", time = 15, main = "Slice at time 15")
```






```r
plot(tdlnm.sum, plot.type = "slice", time = 33, main = "Slice at time 33")
```



C.3 Syntax for TDLMM

This section demonstrates the implementation of treed distributed lag mixture model (TDLMM). More details can be found in Mork and Wilson (2023).

Fitting the model

```
tdlmm.fit <- dlmtree(formula = bwgaz ~ ChildSex + MomAge + MomPriorBMI +
                                       Race + Hispanic + SmkAny + EstMonthConcept,
                     data = sbd_cov,
                     exposure.data = sbd_exp,
                     family = "gaussian", dlm.type = "linear", mixture = TRUE,
                     mixture.interactions = "noself",
                     n.burn = 2500, n.iter = 10000, n.thin = 5)
#> Preparing data...
#>
#> Running TDLMM:
#> Burn-in % complete
#> [0--------25--------50--------75--------100]
#>  ''''''''''''''''''''''''''''''''''''''''''
#> MCMC iterations (est time: 3.4 minutes)
#> [0--------25--------50--------75--------100]
#>  ''''''''''''''''''''''''''''''''''''''''''
#> Compiling results...
```

Model fit summary

```
# Marginalization with co-exposure fixed at exact levels for each exposure
tdlmm.sum.exact <- summary(tdlmm.fit, marginalize = c(3, 2, 1, 2, 1))
#> Reconstructing main effects...
#> Reconstructing interaction effects...
#> 0%...25%...50%...75%...100%
#> Calculating marginal effects...
#> Calculating fixed effects...
```

```
# Marginalization with co-exposure fixed at 25th percentile
tdlmm.sum.percentile <- summary(tdlmm.fit, marginalize = 25)
#> Reconstructing main effects...
#> Reconstructing interaction effects...
#> 0%...25%...50%...75%...100%
#> Calculating marginal effects...
#> Calculating fixed effects...
```

```
# Marginalization with co-exposure fixed at the empirical means (default)
tdlmm.sum <- summary(tdlmm.fit, marginalize = "mean", log10BF.crit = 0.5)
#> Reconstructing main effects...
#> Reconstructing interaction effects...
#> 0%...25%...50%...75%...100%
#> Calculating marginal effects...
#> Calculating fixed effects...
```

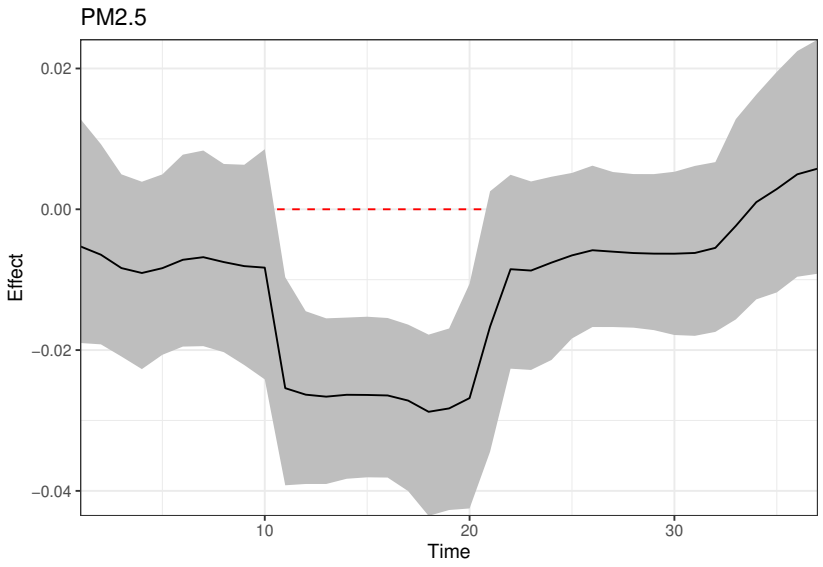
```
tdlmm.sum
#> ---
#> TDLMM summary
#>
#> Model run info:
#> - bwgaz ~ ChildSex + MomAge + MomPriorBMI + Race + Hispanic + SmkAny + EstMonthConcept
#> - sample size: 10,000
#> - family: gaussian
#> - 20 trees (alpha = 0.95, beta = 2)
#> - 2500 burn-in iterations
#> - 10000 post-burn iterations
#> - 5 thinning factor
#> - 5 exposures measured at 37 time points
#> - 10 two-way interactions (no-self interactions)
#> - 1 kappa sparsity prior
#> - 0.95 confidence level
#>
```

```
#> Fixed effects:
#>                         Mean  Lower  Upper
#> *(Intercept)           0.172  0.043  0.307
#> *ChildSexM            -2.063 -2.085 -2.041
#>  MomAge                0.001 -0.001  0.002
#> *MomPriorBMI          -0.020 -0.022 -0.019
#>  RaceAsianPI           0.027 -0.058  0.117
#>  RaceBlack             0.033 -0.063  0.124
#>  Racewhite             0.016 -0.067  0.100
#> *HispanicNonHispanic   0.248  0.224  0.272
#> *SmkAnyY              -0.393 -0.441 -0.346
#>  EstMonthConcept2      0.073 -0.003  0.145
#> *EstMonthConcept3      0.107  0.009  0.211
#> *EstMonthConcept4      0.158  0.038  0.282
#> *EstMonthConcept5      0.255  0.126  0.388
#> *EstMonthConcept6      0.200  0.064  0.333
#> *EstMonthConcept7      0.223  0.084  0.354
#> *EstMonthConcept8      0.199  0.068  0.331
#> *EstMonthConcept9      0.291  0.164  0.418
#> *EstMonthConcept10     0.182  0.070  0.296
#> *EstMonthConcept11     0.135  0.040  0.236
#>  EstMonthConcept12     0.006 -0.062  0.077
#> ---
#> * = CI does not contain zero
#>
#> --
#> Exposure effects: critical windows
#> * = Exposure selected by Bayes Factor
#> (x.xx) = Relative effect size
#>
#>  *PM25 (0.7): 11-20
#>  *TEMP (0.7): 5-19
#>  *SO2 (0.21):
#>  *CO (0.63):
#>  *NO2 (0.26): 23
#> --
#> Interaction effects: critical windows
#>
#>  PM25/TEMP (0.8):
#>  12/6-19
#>  13/6-19
#>  14/6-20
#>  15/6-20
#>  16/6-20
#>  17/6-21
#>  18/5-22
#>  19/5-22
```

```
#>  20/6-21
#> ---
#> residual standard errors: 0.005
```

Main exposure effect

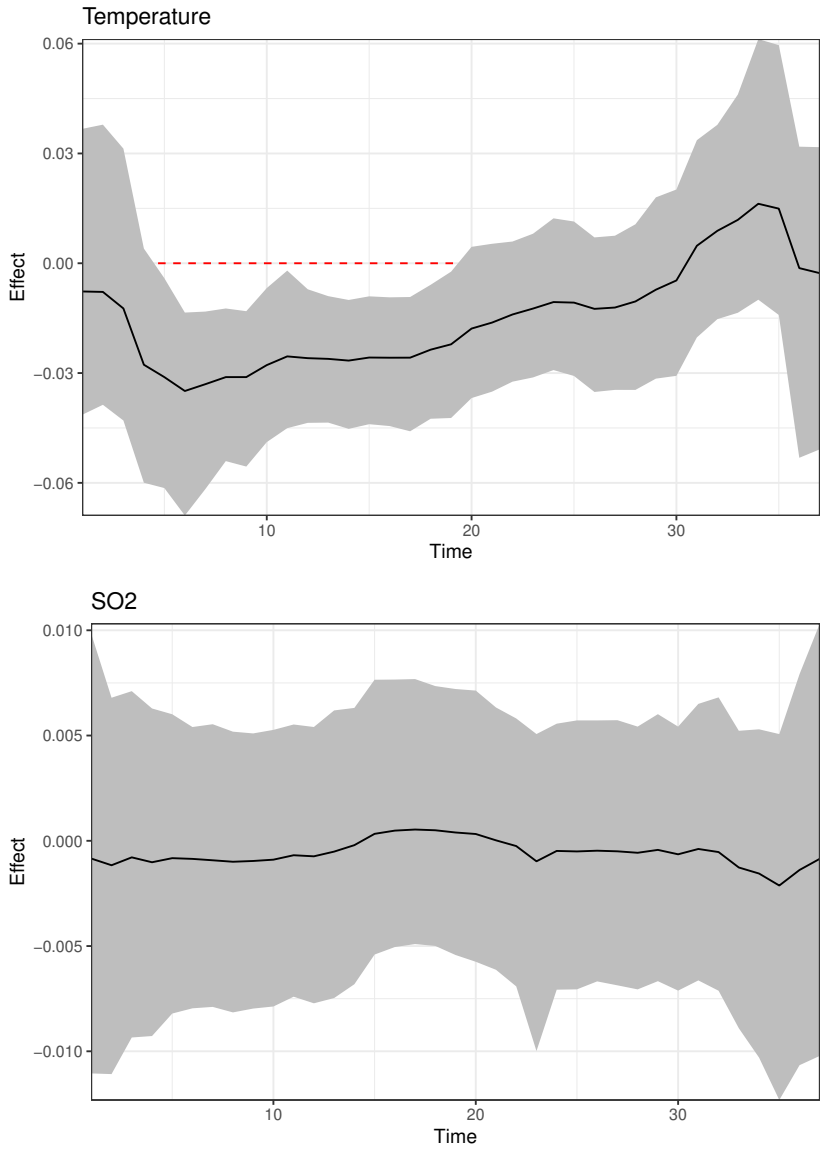
```
p1 <- plot(tdlmm.sum, exposure1 = "PM25", main = "PM2.5")
p2 <- plot(tdlmm.sum, exposure1 = "TEMP", main = "Temperature")
p3 <- plot(tdlmm.sum, exposure1 = "SO2", main = "SO2")

p1
```




```
p2
```

```
p3
```

Lagged interaction effect

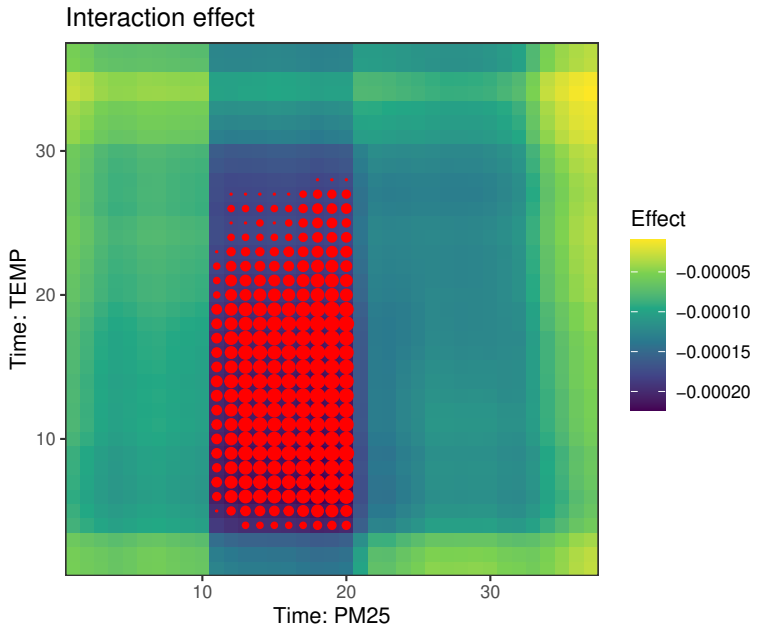



```
plot(tdlmm.sum, exposure1 = "PM25", exposure2 = "TEMP")
```

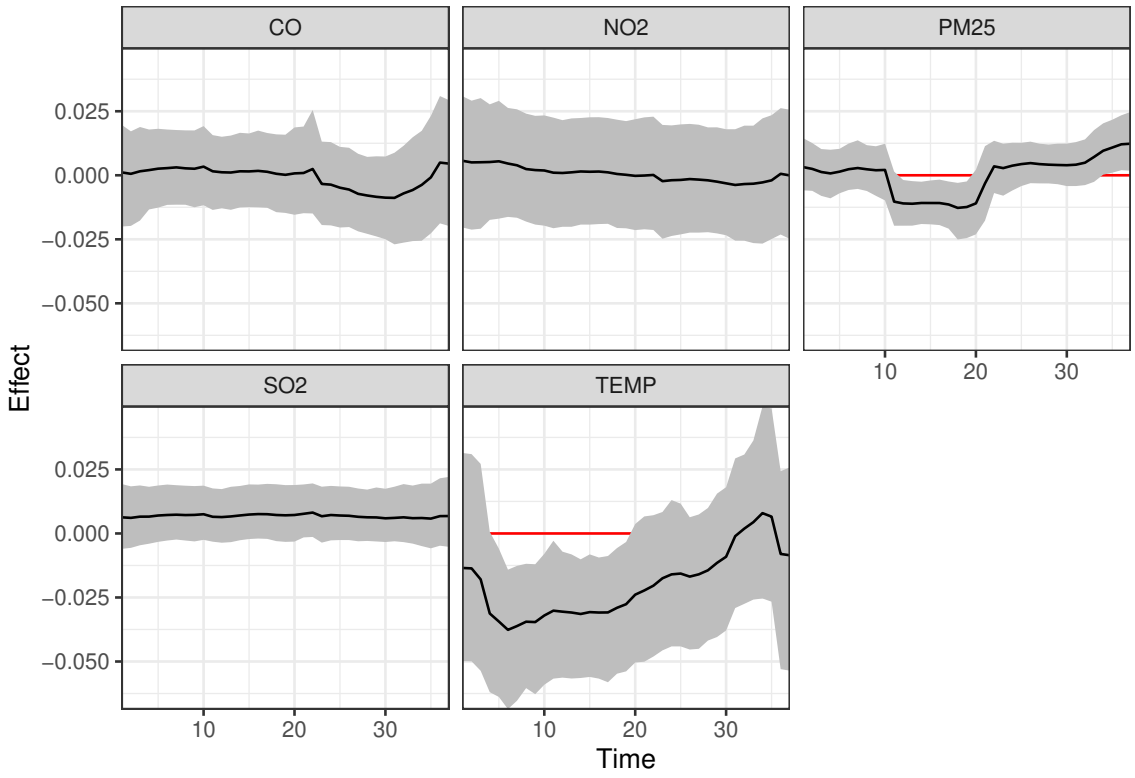
Adjusting for expected changes in co-occurring exposures

Here we consider going from the 25th to the 75th percentile in each exposure while adjusting for the expected changes in other exposures due to their correlations with the exposure of interest.

```
library(ggplot2)
dlm_coexp <- adj_coexposure(sbd_exp, tdlmm.fit,
                            contrast_perc = c(0.25, 0.75), verbose = FALSE)
```




```
ggplot(dlm_coexp, aes(x = Time, y = Effect, ymin = Lower, ymax = Upper)) +
  geom_hline(yintercept = 0, color = "red") +
  geom_ribbon(fill = "grey") +
  geom_line() +
  facet_wrap(~Name) +
  theme_bw() +
  scale_x_continuous(expand = c(0, 0)) +
  scale_y_continuous(expand = c(0, 0))
```



C.4 Syntax for HDLM

This section demonstrates the implementation of heterogeneous treed distributed lag model (HDLM). More details can be found in Mork et al. (2023).

Fitting the model

```
hdlm.fit <- dlmtree(formula = bwgaz ~ ChildSex + MomAge + MomPriorBMI +
                                      Race + Hispanic + SmkAny + EstMonthConcept,
                    data = sbd_cov,
                    exposure.data = sbd_exp[["PM25"]],
                    family = "gaussian", dlm.type = "linear", het = TRUE,
                    hdlm.modifiers = c("ChildSex", "MomAge", "MomPriorBMI", "SmkAny"),
                    hdlm.modifier.splits = 15,
                    n.burn = 2500, n.iter = 10000, n.thin = 5)
#> Preparing data...
#>
#> Running shared HDLM:
#> Burn-in % complete
#> [0--------25--------50--------75--------100]
#>  ''''''''''''''''''''''''''''''''''''''''''
#> MCMC iterations (est time: 2.5 minutes)
```

```
#> [0--------25--------50--------75--------100]
#>  ''''''''''''''''''''''''''''''''''''''''''
#> Compiling results...
```

Model fit summary

```
hdlm.sum <- summary(hdlm.fit)
hdlm.sum
#> ---
#> HDLM summary
#>
#> Model run info:
#> - bwgaz ~ ChildSex + MomAge + MomPriorBMI + Race + Hispanic + SmkAny + EstMonthConcept
#> - sample size: 10,000
#> - family: gaussian
#> - 20 trees
#> - 2500 burn-in iterations
#> - 10000 post-burn iterations
#> - 5 thinning factor
#> - 0.5 modifier sparsity prior
#> - 0.95 confidence level
#>
#> Fixed effects:
#>                         Mean  Lower  Upper
#> *(Intercept)           1.272  0.891  1.650
#>  ChildSexM             0.080 -0.365  0.525
#>  MomAge                0.001 -0.004  0.007
#> *MomPriorBMI          -0.021 -0.023 -0.019
#>  RaceAsianPI           0.046 -0.075  0.174
#>  RaceBlack             0.056 -0.065  0.185
#>  Racewhite             0.035 -0.081  0.161
#> *HispanicNonHispanic   0.254  0.231  0.277
#> *SmkAnyY              -0.405 -0.452 -0.356
#>  EstMonthConcept2     -0.047 -0.102  0.006
#> *EstMonthConcept3     -0.133 -0.197 -0.072
#> *EstMonthConcept4     -0.203 -0.263 -0.143
#> *EstMonthConcept5     -0.194 -0.246 -0.142
#> *EstMonthConcept6     -0.201 -0.253 -0.149
#>  EstMonthConcept7     -0.037 -0.091  0.015
#> *EstMonthConcept8      0.146  0.086  0.207
#> *EstMonthConcept9      0.388  0.324  0.452
#> *EstMonthConcept10     0.389  0.326  0.449
#> *EstMonthConcept11     0.342  0.288  0.397
#> *EstMonthConcept12     0.143  0.095  0.191
#> ---
#> * = CI does not contain zero
```

```
#>
#> Modifiers:
#>               PIP
#> ChildSex    1.000
#> MomAge      1.000
#> MomPriorBMI 0.306
#> SmkAny      0.107
#> ---
#> PIP = Posterior inclusion probability
#>
#> residual standard errors: 0.004
#> ---
#> To obtain exposure effect estimates, use the 'shiny(fit)' function.
```

Launching Shiny app

```
shiny(hdlm.fit)
```

C.5 Syntax for HDLMM

This section demonstrates the implementation of heterogeneous treed distributed lag mixture model (HDLMM).

Fitting the model

```
# Gaussian
hdlmm.fit <- dlmtree(formula = bwgaz ~ ChildSex + MomAge + MomPriorBMI +
                                       Race + Hispanic + SmkAny + EstMonthConcept,
                     data = sbd_cov,
                     exposure.data = sbd_exp,
                     family = "gaussian",
                     dlm.type = "linear", mixture = TRUE, het = TRUE,
                     hdlm.modifiers = c("ChildSex", "MomAge", "MomPriorBMI", "SmkAny"),
                     hdlm.modifier.splits = 15,
                     n.burn = 2500, n.iter = 10000, n.thin = 5)
```

```
#> Preparing data...
#>
#> Running HDLMM:
#> Burn-in % complete
#> [0--------25--------50--------75--------100]
```

```
#>  ''''''''''''''''''''''''''''''''''''''''
#> MCMC iterations (est time: 7.2 minutes)
#> [0--------25--------50--------75--------100]
#>  ''''''''''''''''''''''''''''''''''''''''
#> Compiling results...
```

Model fit summary

```
hdlmm.sum <- summary(hdlmm.fit)
hdlmm.sum
```

```
#> ---
#> HDLMM summary
#>
#> Model run info:
#> - bwgaz ~ ChildSex + MomAge + MomPriorBMI + Race + Hispanic + SmkAny + EstMonthConcept
#> - family: gaussian
#> - 20 trees
#> - 2500 burn-in iterations
#> - 10000 post-burn iterations
#> - 5 thinning factor
#> - 5 exposures measured at 37 time points
#> - 10 two-way interactions (no-self interactions)
#> - 0.5 modifier sparsity prior
#> - 1 exposure sparsity prior
#> - 0.95 confidence level
#>
#> Fixed effects:
#>                         Mean  Lower  Upper
#> *(Intercept)           1.463  0.878  2.101
#>  ChildSexM            -0.462 -1.613  0.442
#>  MomAge                0.000 -0.003  0.003
#> *MomPriorBMI          -0.020 -0.023 -0.017
#>  RaceAsianPI           0.027 -0.099  0.152
#>  RaceBlack             0.037 -0.092  0.168
#>  Racewhite             0.017 -0.106  0.142
#> *HispanicNonHispanic   0.254  0.232  0.275
#> *SmkAnyY              -0.392 -0.444 -0.330
#> *EstMonthConcept2      0.125  0.052  0.197
#> *EstMonthConcept3      0.232  0.122  0.334
#> *EstMonthConcept4      0.323  0.176  0.467
#> *EstMonthConcept5      0.411  0.234  0.573
#> *EstMonthConcept6      0.386  0.199  0.559
#> *EstMonthConcept7      0.407  0.230  0.573
```

```
#> *EstMonthConcept8     0.386  0.222  0.537
#> *EstMonthConcept9     0.455  0.309  0.588
#> *EstMonthConcept10    0.327  0.216  0.435
#> *EstMonthConcept11    0.220  0.141  0.302
#>  EstMonthConcept12    0.044 -0.017  0.103
#> ---
#> * = CI does not contain zero
#>
#> Modifiers:
#>                PIP
#> ChildSex    1.0000
#> MomAge      0.9560
#> MomPriorBMI 0.9885
#> SmkAny      0.3080
#> ---
#> PIP = Posterior inclusion probability
#>
#> residual standard errors: 0.009
#> ---
#> To obtain exposure effect estimates, use the 'shiny(fit)' function.
```

Launching Shiny app

```
shiny(hdlmm.fit)
```