Distribution Agreement

In presenting this dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this dissertation. I retain all ownership rights to the copyright of the dissertation. I also retain the right to use in future works (such as articles or books) all or part of this dissertation.

_____ _____

Jacob R. Englert Date

Bayesian Tree-Based Methods for Environmental Health Research

By

Jacob R. Englert
Doctor of Philosophy

Biostatistics

Howard Chang, Ph.D.
Advisor

David Benkeser, Ph.D.
Committee Member

Stefanie Ebelt, Sc.D.
Committee Member

Lance Waller, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Bayesian Tree-Based Methods for Environmental Health Research

By

Jacob R. Englert
B.S., Northern Kentucky University, 2019
M.S., Emory University, 2023

Advisor: Howard Chang, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2025

Abstract

Bayesian Tree-Based Methods for Environmental Health Research
By Jacob R. Englert

Bayesian nonparametric models are widely used for estimating complex relationships and functional forms among predictors in regression settings. Within this class of models, Bayesian Additive Regression Trees (BART) is frequently cited for its strong performance and flexibility in a wide variety of statistical problems. This dissertation extends BART to three modeling frameworks commonly used to measure associations between environmental exposures and health outcomes.

In the first aim, a varying coefficient BART model is introduced to estimate heterogeneous short-term associations between acute exposure and health outcomes within the case-crossover design. This approach is applied to examine trends in emergency department visits among patients with Alzheimer's disease during heat waves in California. The proposed method allows individual responses to heat waves to vary based on chronic comorbid conditions such as chronic kidney disease and hypertension, thus providing a more nuanced understanding of heat-related vulnerability in this population.

For the second aim, a soft version of BART is applied to model count-based health outcomes in environmental mixtures studies. The approach approximates a smooth exposure-risk surface for daily asthma-related emergency department visits in the Metropolitan Atlanta area, modeling risk as a function of temperature and an air pollution mixture consisting of ozone, fine particulate matter, nitrogen dioxide, and carbon monoxide. Existing BART implementations for count outcomes require complex prior specifications, making it difficult to incorporate other useful model components such as spatial random effects and population offsets. To address this, we use latent random variables to model the risk surface. We further describe the use of accumulated local effects for summarizing exposure-risk surfaces composed of correlated continuous exposures.

In the third aim, an extension of the quantile g-computation framework for studying heterogeneous effects of environmental mixtures is proposed. When data arise from a large geographical study region, it may be unreasonable to expect a common mixture effect due to variation in the composition of the mixture or nonlinearity in the true exposure-response function. The proposed method leverages a recently developed varying coefficient BART model to explore spatially varying mixture effects describing the association between air pollution mixtures and reduced birth weight in Georgia.

Bayesian Tree-Based Methods for Environmental Health Research

By

Jacob R. Englert
B.S., Northern Kentucky University, 2019
M.S., Emory University, 2023

Advisor: Howard Chang, Ph.D.

A dissertation submitted to the Faculty of the
Emory College of Arts and Sciences of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2025

Acknowledgments

I would like to begin by acknowledging my advisor, Dr. Howard Chang, for his mentorship and support throughout this process. I am constantly amazed by your vast wealth of knowledge about statistics and environmental epidemiology, and I always left our meetings feeling like so much was accomplished. I am also thankful to my committee members for their insights, particularly Dr. Stefanie Ebelt for her helpful comments while attempting to publish my work. Thank you Dr. Lance Waller for being a source of joy not only for myself, but for the entire department. I appreciate the opportunities I had to serve as a teaching assistant for Dr. Waller, Dr. Chang, and Dr. Robert Lyles.

I am also incredibly grateful for all who guided me through my journey to prepare me for this endeavor. I extend my heartfelt thanks to the late Mr. Gary Rice for not only sitting me next to the cute girl in class, but also for his unwavering commitment to his students. Your dedication to mathematics education was unmatched, and your legacy continues to inspire me. Thank you to all of my peers and professors in the Department of Mathematics and Statistics at NKU for instilling in me a love for statistics.

Beginning a Ph.D. program in a fully remote environment was not what I had envisioned, but I couldn't be more grateful for my Track 1 "COVID cohort", including Delante Moore, Sydney Busch, and Sam Yin. Our occasional time together on campus made those challenging first years more manageable. The three of you, along with our honorary cohort inductee Wyatt Madden and an amazing TA and friend Thomas Hsiao, provided the glimmers of hope I needed in those pre-candidacy years.

I would like to express my sincere gratitude to my office mates. I appreciate your willingness to always listen to me talk through something, school-related or otherwise. Thank you Dr. Amita Manatunga for recruiting Emily Wu and myself to work on the same project—she has become a close friend and I have learned so much from her. I am grateful for Lindsey Schader, who has been not only a great friend, but also an excellent role model as a Ph.D. student and early-career statistician.

I would be remiss not to mention my Atlanta soccer community. Michael and Ryan—our intramural championship may have eluded us, but I'll always appreciate the great games and even better company. To my Brook Run weekend pickup group—thank you for five years of friendship, football, and making Atlanta feel like home.

To my family back home: the most difficult part of this journey was spending extended time away from all of you. I am the person I am today because of you. Thank you Mom, Dad, Jen, Bryan, and Amy for the unconditional love you have shown and sacrifices you made over the years. Thank you Aaron, Lauren, Rachel, Ryan, Mark, and Aleah, for giving me something to smile about when nothing else seemed to be going right.

And to my wife, Rachel—who followed me down to Georgia in the middle of a global pandemic so that I could pursue this opportunity—your steadfast love, patience, and kindness have carried me through this journey. I would not be here without you, and I look forward to what comes next for us.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Organization

The research presented in this dissertation is divided into three aims, each exploring the utility of Bayesian additive regression trees (BART) for estimating complex relationships between one or more environmental exposures and health outcomes. The chapters corresponding to the three aims share the same outline: (1) Introduction, (2) Data, (3) Methods, (4) Simulation Study, (5) Application, and (6) Discussion. Each chapter concludes with Supplementary Materials section containing relevant additional tables, figures, and derivations.

Before presenting the three aims, Chapter 2 provides background on the technical details of BART. This chapter covers the original model formulation, estimation procedures, and recent methodological advances relevant to the methods and applications discussed in this dissertation. It also introduces model interpretation strategies that are used extensively in later chapters.

Chapter 3 addresses the first aim. In this chapter, we develop a novel extension of the popular case-crossover study design for estimating heterogeneous exposure-response relationships using BART. This work is motivated by the growing interest in identifying subpopulations more vulnerable to environmental exposures. We apply the proposed method

to study the impact of heat waves on people with Alzheimer's disease in California from 2005-2015. We examine effect modification by other chronic conditions such as hypertension and chronic kidney disease. Through this application, we illustrate strategies for interpreting heterogeneous odds ratios through variable importance, partial dependence, and lower-dimensional summaries.

Chapter 4 addresses the second aim. In this chapter, we demonstrate how to incorporate a soft version of BART into a negative binomial regression model to approximate smooth exposure-risk surfaces for count outcomes. The proposed approach enables flexible modeling of mixtures of potentially correlated environmental exposures which may interact with each other. We apply this method to estimate associations between air pollution mixtures, temperature and asthma-related emergency department visits during the warm season in Atlanta, Georgia from 2011-2018. Additionally, we use a strategy known as accumulated local effects to extract meaningful insights into the association between the mixture of interest and asthma-related morbidity.

Chapter 5 addresses the third and final aim of this dissertation. In this chapter, we explore the ability of a recently developed varying coefficient BART model to estimate spatially heterogeneous mixture effects within the quantile g-computation summary index framework. After reviewing spatially varying coefficient models and quantile g-computation, we demonstrate the advantages of varying coefficient BART through simulation. We then apply this model to analyze associations between multiple ambient air pollutants and birth weight in Georgia from 2005-2016.

Finally, Chapter 6 provides concluding remarks and discusses potential future opportunities for integrating BART into statistical modeling frameworks for applications in environmental health. This chapter also describes software that has been developed to perform these analyses.

1.2 Description of Data Sources

The applications discussed in Chapters 3, 4, and 5 involve the analysis of large health administrative and exposure datasets which pull from different sources. While the specific data sources used for each application are described in their respective chapters, this section serves to briefly introduce each data source and provide insight how they were combined.

1.2.1 Health Data Sources

California Department of Health Care Access and Information

Patient-level records for all emergency department visits in the state of California from 2005 to 2015 were obtained from the California Department of Health Care Access and Information (formerly the California Office of Statewide Health Planning and Development). The raw data is available by request at `https://hcai.ca.gov/data/request-data/`. These data contained the admission date, the residential ZIP code for the patient, and demographic information. Each visit record also contained a list of diagnosis codes based on the International Classification of Diseases, ninth and tenth revisions (ICD-9 and ICD-10). From these codes we determined primary and secondary diagnoses of Alzheimer's disease and other chronic conditions. For the analysis in Chapter 3, these records are filtered to only include those visits with a primary or secondary diagnosis of Alzheimer's disease (ICD-9 code 331.0; ICD-10 codes G30.0, G30.1, G30.8, and G30.9).

Georgia Hospital Association

Patient-level billing records for emergency department visits to hospitals in the metropolitan Atlanta area from 2011 to 2018 were obtained from the Georgia Hospital Association (now available through the Healthcare Cost and Utilization Project (HCUP) at `https://hcup-us.ahrq.gov/tech_assist/centdist.jsp`). These data included admission date, billing address, ICD-9/10 discharge diagnosis codes, and patient demographic

information. For the analysis in Chapter 4, these records are filtered to only include visits with a primary or secondary asthma diagnosis (ICD-9 code 493; ICD-10 code J45).

Office of Health Indicators for Planning, Georgia Department of Public Health

Birth records were obtained from the Office of Health Indicators for Planning, Georgia Department of Public Health (available by request via the Public Health Information Portal at `https://dph.georgia.gov/phip-data-request`). These data covered births to mothers residing in any of the 159 counties in Georgia from January 1st, 2005 to December 31st, 2017. Each record contained information about the birth such as the date of birth, estimated conception date and gestational age, birth weight, sex, and plurality. The data also contain demographic information for the mother, including race, ethnicity, level of educational attainment, marital status, and self-reported use of alcohol or tobacco.

Emergency Department Visits vs. Hospitalizations

Both sets of emergency department visit data we use include hospitalizations. The decision to include all emergency department visits provides a greater sample size for studying health outcomes with simple fixes (e.g., administration and distribution of albuterol for asthmatic patients). Limiting the analyses to only hospitalizations would greatly reduce the power to detect associations between exposures and health outcomes.

1.2.2 Exposure Data Sources

Daymet

Daily estimates of minimum air temperature, maximum air temperature, and water vapor pressure were obtained from Daymet Version 4 R1 [104] (available for download at `https://doi.org/10.3334/ORNLDAAC/2129`). This data product covers North America and has a 1km x 1km spatial resolution and daily temporal resolution beginning with January 1st, 1980. Daymet pools from many weather stations across the continental North America, Hawaii,

and Puerto Rico, and interpolates using multiple algorithms to obtain near-surface estimates at a fine spatial resolution.

Air Quality Data

Daily estimates of concentrations of various air pollutants are obtained from the Community Multiscale Air Quality Modeling System (CMAQ) and the Environmental Protection Agency's Air Quality System (AQS) database (available for download at `https://www.epa.gov/cmaq` and `https://www.epa.gov/aqs`). CMAQ produces estimates of a wide variety of pollutants using cutting edge air quality models. The AQS compiles measurements from many air quality networks across the United States.

Specifically, the air pollution data used in Chapters 4 and 5 was produced using the data fusion model described by Senthilkumar et al. [93], which utilized chemical transport model simulations from CMAQ and monitoring data from AQS. This model has been shown to reduce spatial bias in the CMAQ estimates for many common air pollutants by incorporating land use variables and census tract level population data. This data product is available at a 12km x 12km gridded spatial resolution.

1.2.3 Linking Exposures to Health Data

Linking exposure data to patient billing records requires some decision making. Geocoding the residential addresses attached to the billing records allows for matching records to ZIP codes, census tracts, and counties. Matching gridded air quality and meteorology data to these areal units is a bit more involved, as it requires area-weighted spatial averaging. This is done by overlaying the gridded data with the areas of interests and, for each areal unit, calculating a weighted average of all grid estimates which overlap the area, where the weights are determined by the percentage of the area covered by each grid cell.

For each of the applications discussed in this dissertation, daily area-weighted averages are calculated at the ZIP code level for all temperature and air pollution data. These exposure

values are then linked to the health data by ZIP code and date. For the analysis of birth weight in Chapter 5, the assigned daily exposures are also averaged over the duration of the pregnancy.

This strategy for merging temperature and air quality data with health administrative datasets is common in the environmental epidemiology literature, but is not without flaw. While gridded estimates are improving in quality thanks to tools such as those mentioned above, they do contain varying levels of measurement error. Also, Geographic Information Systems (GIS) tools do not always geocode addresses correctly. Certain areas may not be well-covered, and even a perfect GIS geocoding instrument will fail if the supplied address is incorrect. Even for addresses which are successfully geocoded, or for aggregate data reported by areal units, the assigned exposure may not be accurate for all individuals. Unmeasured factors such as lifestyle behaviors or occupation can cause even next-door neighbors to experience very different levels of exposures. Additionally, for some individuals their home address may not even be the most important location. For example, if one were studying wildfire smoke, a firefighter who resides in an area with no wildfires may experience health outcomes similar to an individual exposed to wildfire smoke.

1.3 Notation

The following notation is used throughout the document to denote different types of data and common parameters across chapters. Additional notation specific to each chapter is defined where it is introduced.

1.3.1 Data

- N: total number of observations in a dataset.
- $\mathcal{D}$: collection of observed data for all observations.

- $\mathbf{Y}$: $N \times 1$ random outcome vector.
  - $\boldsymbol{y}$: $N \times 1$ observed vector (used in the data likelihood).
- Y_i: random outcome for observation i.
  - y_i: observed outcome for observation i (used in the data likelihood).
- $\mathbf{W}$: $N \times P_w$ design matrix of confounders.
- $\boldsymbol{w}_i$: $P_w \times 1$ vector of confounders for observation i.
- $\mathbf{X}$: $N \times P_x$ matrix of exposures.
- $\boldsymbol{x}_i$: $P_x \times 1$ vector of exposures for observation i.
- $\mathbf{Z}$: $N \times P_z$ matrix of exposure moderators.
- $\boldsymbol{z}_i$: $P_z \times 1$ vector of exposure moderators for observation i.

1.3.2 Densities and Distributions

- $p(\mathcal{D} \mid \theta)$: data likelihood, in a Bayesian context.
- $\pi(\theta)$: prior distribution of the parameter θ.
- $\pi(\theta \mid \mathcal{D})$: posterior distribution of θ given data $\mathcal{D}$.
- $\pi(\theta) = f(\theta \mid \cdot) \Leftrightarrow \theta \sim f(\cdot)$
  - E.g., $\pi(\theta) = \text{Normal}(\theta \mid \mu, \sigma^2)$ is synonymous with $\theta \sim \text{Normal}(\mu, \sigma^2)$.

1.3.3 Parameters

The following parameters are consistent in their use throughout this dissertation:

- $\boldsymbol{\gamma}$: $P_w \times 1$ vector of regression coefficients for confounders.

- β_p: regression coefficient of the p^{th} exposure. p omitted implies there is only one exposure.
- $\beta_p(\boldsymbol{z}_i)$: heterogeneous regression coefficient of the p^{th} exposure, which depends on $\boldsymbol{z}_i$. p omitted implies there is only one exposure.
- $\mathcal{T}$: collection of decision rules that define a binary tree structure.
- $\mathcal{M}$: collection of scalar-valued terminal ("leaf") node parameters associated with a tree structure $\mathcal{T}$.
- μ_{tl}: a single terminal ("leaf") node parameter associated with leaf l of tree t.

Note on Causation vs. Association

Throughout this dissertation, certain parameters (e.g., β_p) are sometimes referred to as exposure *effects*. While this term may imply a causal relationship – suggesting that modifying the associated exposure would directly or indirectly change the outcome – we use it synonymously with the more general *association*. Although biological evidence supports the causal role of air pollution and extreme heat in various health outcomes, the statistical models presented in this work are not designed to explicitly target causal parameters. Similarly, *effect modification* or *effect moderation* simply refer to heterogeneity in the estimated parameter across demographic subgroups, space, etc.

Chapter 2

Review of Bayesian Additive Regression Trees

2.1 Bayesian Additive Regression Trees

The methods developed and applied in this dissertation are based on a flexible fully Bayesian machine learning approach known as *Bayesian Additive Regression Trees* (BART) [25]. Since BART was introduced in the late 2000s, researchers have applied BART in numerous settings and extended the original methodology to model various types of data and estimate different parameters of interest. BART is widely known for its excellent empirical performance in prediction, classification, and causal inference settings when compared to or used in conjunction with other state of the art methods [25, 52, 31]. Because BART is integral to the work described in this dissertation, this chapter has been provided as a review of the method insofar as it relates to the work described herein. Additionally, the reader may be interested in recently published reviews of BART included in Hill et al. [51] and Tan and Roy [102]. A textbook treatment is available in Daniels et al. [27].

2.1.1 The Original BART Model

BART is a Bayesian ensemble approach most closely related to *boosting* in the machine learning literature, where the individual base learners are Bayesian classification and regression trees (BCART) [24]. The original BCART model is a semiparametric regression model for Gaussian outcome data written as in (2.1).

$$Y \sim \text{Normal}\left(f(\boldsymbol{x}), \sigma^2\right) \tag{2.1}$$

$$f(\boldsymbol{x}) = \text{Tree}(\boldsymbol{x}; \mathcal{T}, \mathcal{M}), \tag{2.2}$$

where $\mathcal{T}$ is a binary tree structure with L terminal, or *leaf*, nodes, and $\mathcal{M} = \{\mu_1, \ldots, \mu_L\}$ is the set of scalar-valued leaf node parameters associated with tree $\mathcal{T}$. Further, $\text{Tree}(\boldsymbol{x}; \mathcal{T}, \mathcal{M})$ is the function which deterministically maps a vector of covariates $\boldsymbol{x} = (x_1, x_2, \ldots, x_{P_x})^T$ to a single leaf node l of $\mathcal{T}$ and assigns it the corresponding scalar parameter $\mu_l \in \mathcal{M}$. An example of such a binary tree is provided in Figure 2.1.

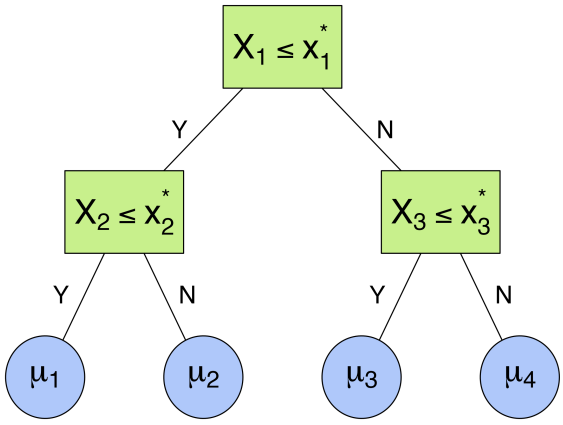

Figure 2.1: Example Binary Tree

Note how the binary tree in Figure 2.1 partitions the covariate space into four distinct groups defined by the covariates X_1, X_2, and X_3, and a set of selected cut points x_1^*, x_2^*, and x_3^*. Each observation is mapped to exactly one of the four leaf nodes depending on the values of its covariates. The three internal nodes depicted as rectangles are referred to as *branch*

nodes (i.e., nodes that split). Henceforth $\mathcal{B}(\mathcal{T})$ will be used to refer to the set of branch nodes and $\mathcal{L}(\mathcal{T})$ will be used to refer to the set of leaf nodes for a tree structure $\mathcal{T}$.

Conceptually, BCART should enjoy several benefits over a frequentist classification and regression tree (CART) [15]. For one, BCART has the ability to fit more data generating processes than CART. This is because CART is "greedy" in the sense that it always selects the best available split based on some scoring criteria, such as the Gini impurity for classification or root mean squared error (RMSE) for prediction. BCART on the other hand randomly samples splitting rules using a Markov chain Monte Carlo (MCMC) algorithm. This allows BCART to explore tree structures that CART cannot. For instance, if the splitting rule based on X_1 in Figure 2.1 was associated with the greatest improvement in scoring criteria out of all splitting rules, then it would always be chosen as the first split in a CART model. Due to the stochastic nature of BCART, this split could appear further down in the tree, or not even be used at all in certain iterations of the MCMC algorithm. Since BCART is Bayesian, the end result of the MCMC algorithm is a posterior distribution of regression trees. Point estimates for predictions are typically taken as the average prediction across posterior samples, which might be viewed as an ensemble-like model averaging approach. Additionally, one can obtain natural estimates of uncertainty for any and all predictions via the posterior distribution. This can be more convenient than, say, bootstrapping a CART model.

Unfortunately, BCART has been shown to have poor mixing, often requiring multiple restarts or many MCMC chains to obtain reliable samples from the posterior distribution of trees [24]. CART also has its limitations, including to but not limited to those discussed above. For this reason, when the true prediction function is complex, algorithms using ensembles of CART models, such as *gradient boosted trees* and *random forests*, are more widely used than just a single CART model. Similarly, we can use BART, which is an ensemble of BCART models, to achieve better performance.

BART extends the BCART approach to the "sum-of-trees" model which replaces the

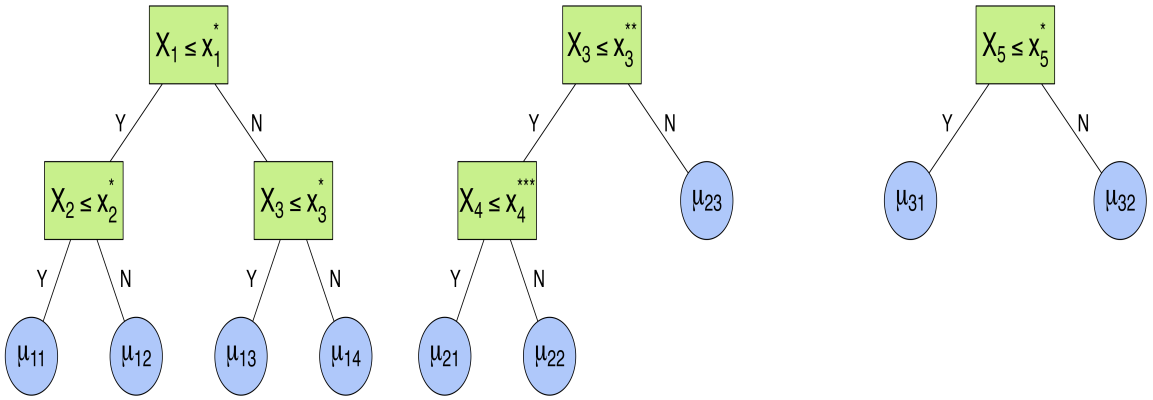

Figure 2.2: Example State of a 3-tree BART Ensemble

mean model in (2.2) with (2.3),

$$f(\boldsymbol{x}) = \sum_{t=1}^{T} \text{Tree}(\boldsymbol{x}; \mathcal{T}_t, \mathcal{M}_t) \tag{2.3}$$

where t is used to index the each of the T binary tree structures ($\mathcal{T}_t$) and their corresponding sets of leaf node parameters ($\mathcal{M}_t$). Each individual tree provides only a small contribution to the overall ensemble prediction, which allows for more efficient exploration of the parameter space. This behavior generally results in BART being more efficient in sampling from the posterior distribution than BCART, as well as being more readily capable of approximating complex regression functions. It is common for BART ensembles to include as many as 50 or 200 trees, where each individual tree may use different splitting criteria based on the same or different predictors. See Figure 2.2 for an example of what a posterior sample of a BART ensemble of $T = 3$ trees might look like.

2.1.2 Parameter Estimation

The tree ensemble portion of any BART model is typically referred to as nonparametric, however this simply means that there is a vast (and variable) number of parameters in the model which require estimation. The parameters which must be estimated in a BART

ensemble are the tree structures $\{\mathcal{T}_t\}_{t=1}^T$ and their corresponding sets of leaf node parameters $\{\mathcal{M}_t\}_{t=1}^T$. This estimation is carried out using MCMC to draw samples of all parameters from their joint posterior distribution. The posterior distribution for any BART model can be written as in (2.4):

$$\pi\left((\mathcal{T}_1, \mathcal{M}_1), \ldots, (\mathcal{T}_T, \mathcal{M}_T), \boldsymbol{\eta} \mid \mathcal{D}\right), \tag{2.4}$$

where $\boldsymbol{\eta}$ represents any additional parameters which might be present in the larger semi-parametric model (e.g., σ^2 in (2.1)), and $\mathcal{D} = \{y_i, \boldsymbol{x}_i\}_{i=1}^N$ is the observed data. Define the sum-of-trees prediction for a single observation i as $\lambda_i \equiv \sum_{t=1}^T \text{Tree}(\boldsymbol{x}_i; \mathcal{T}_t, \mathcal{M}_t)$, and the vector containing the prediction for all N observations as $\boldsymbol{\lambda} \equiv (\lambda_1, \ldots, \lambda_N)^T$. It is also helpful to define the partial residuals $\lambda_i^t \equiv y_i - \sum_{k \neq t} \text{Tree}(\boldsymbol{x}_i; \mathcal{T}_k, \mathcal{M}_k)$, which represent the residual of the overall BART fit excluding the contribution from $\mathcal{T}_t$, and the corresponding vector $\boldsymbol{\lambda}^t \equiv (\lambda_i^t, \ldots, \lambda_N^t)^T$. The MCMC algorithm for sampling from the BART posterior is provided in Algorithm 2.1.

Algorithm 2.1 One MCMC Iteration of the Original BART Algorithm

1: **Input:** $\mathcal{D}, \{\mathcal{T}_t, \mathcal{M}_t\}_{t=1}^T, \boldsymbol{\eta}$
2: **for** $t = 1, \ldots, T$ **do**
3: Compute the partial residuals $\boldsymbol{\lambda}^t$.
4: Sample $\mathcal{T}_t \sim \pi(\mathcal{T}_t \mid \{\mathcal{T}_k, \mathcal{M}_k\}_{k \neq t}, \boldsymbol{\eta}, \mathcal{D}) \equiv \pi(\mathcal{T}_t \mid \boldsymbol{\lambda}^t)$.
5: Sample $\mathcal{M}_t \sim \pi(\mathcal{M}_t \mid \mathcal{T}_t, \{\mathcal{T}_k, \mathcal{M}_k\}_{k \neq t}, \boldsymbol{\eta}, \mathcal{D}) \equiv \pi(\mathcal{M}_t \mid \mathcal{T}_t, \boldsymbol{\lambda}^t)$ using Gibbs sampling.
6: **end for**
7: Update any parameters included in $\boldsymbol{\eta}$, conditional on $\mathcal{D}$ and $\{\mathcal{T}_t, \mathcal{M}_t\}_{t=1}^T$.

In Algorithm 2.1, the tree structures and their corresponding sets of leaf node parameters are updated sequentially using Bayesian backfitting [49]. The first step to updating the t^{th} tree involves computing the vector of partial residuals $\boldsymbol{\lambda}^t$ given the most recent values of the other $T-1$ trees. The ensuing updates of $\mathcal{T}_t$ and $\mathcal{M}_t$ only depend on the data and other trees through the partial residual $\boldsymbol{\lambda}^t$, which serves as the outcome in the Gaussian likelihood calculation. Thus, the t^{th} regression tree is being trained on the partial residuals, as is done in traditional boosting.

Updating Tree Structures

In step 4 of Algorithm 2.1, a new tree is sampled from its marginal distribution. The first step in doing this is to perturb the existing tree structure using, for example, a GROW, PRUNE, or CHANGE move. The type of move is chosen at random given some prior probabilities, and conditional on a selected move type, the choice of which node(s) are affected and the decision rule is also chosen randomly (see Section 2.1.2). Figure 2.3 illustrates an example of how one might arrive at the ensemble depicted in Figure 2.2 after four MCMC iterations using these three moves. The first tree grows at each iteration, while the second tree grows at iteration 2, changes the decision rule at iteration 3, and then grows again at iteration 4. Finally, the third tree grows at iteration 2, is pruned at iteration 3, and grows again at iteration 4, this time using a new decision rule.

While all of the trees in Figures 2.3 change with each iteration, it is also possible for the tree structure to remain unchanged after an iteration. The decision of whether or not to accept the proposed tree structure at any given iteration is made using a Metropolis-Hastings (M-H) step [78, 50]. Because the dimension of the parameter space may change with a GROW or PRUNE move, this step technically requires reversible jump MCMC (RJMCMC) [44, 43]. Brief overviews of sampling from the posterior using the M-H procedure and RJMCMC are provided in Appendixes A.2.1 and A.2.2. The RJMCMC M-H acceptance ratio for the proposal of a new tree structure $\mathcal{T}_t'$ from an existing tree structure $\mathcal{T}_t$ is given by Equation (2.5):

$$\begin{aligned} r_{\mathcal{T}} &= \frac{\pi\left(\mathcal{T}', \mathcal{M}^* \mid \mathcal{D}\right)}{\pi\left(\mathcal{T}, \mathcal{M} \mid \mathcal{D}\right)} \times \frac{q\left(\mathcal{T}, \mathcal{M} \mid \mathcal{T}', \mathcal{M}^*\right)}{q\left(\mathcal{T}', \mathcal{M}^* \mid \mathcal{T}, \mathcal{M}\right)} \\ &= \frac{\pi\left(\mathcal{T}', \mathcal{M}^* \mid \mathcal{D}\right)}{\pi\left(\mathcal{T}, \mathcal{M} \mid \mathcal{D}\right)} \times \frac{q_1\left(\mathcal{T} \mid \mathcal{T}'\right)}{q_1\left(\mathcal{T}' \mid \mathcal{T}\right)} \times \frac{q_2\left(\mathcal{M} \mid \mathcal{M}^*\right)}{q_2\left(\mathcal{M}^* \mid \mathcal{M}\right)} \end{aligned} \tag{2.5}$$

where $q(\cdot \mid \cdot)$ is the proposal distribution for $\mathcal{T}'$ and $\mathcal{M}^*$, the latter of which represents the values we would need to propose for $\mathcal{M}$ under the new tree structure. It is typically the case that the proposal distribution q may be factored into a proposal for the tree structure

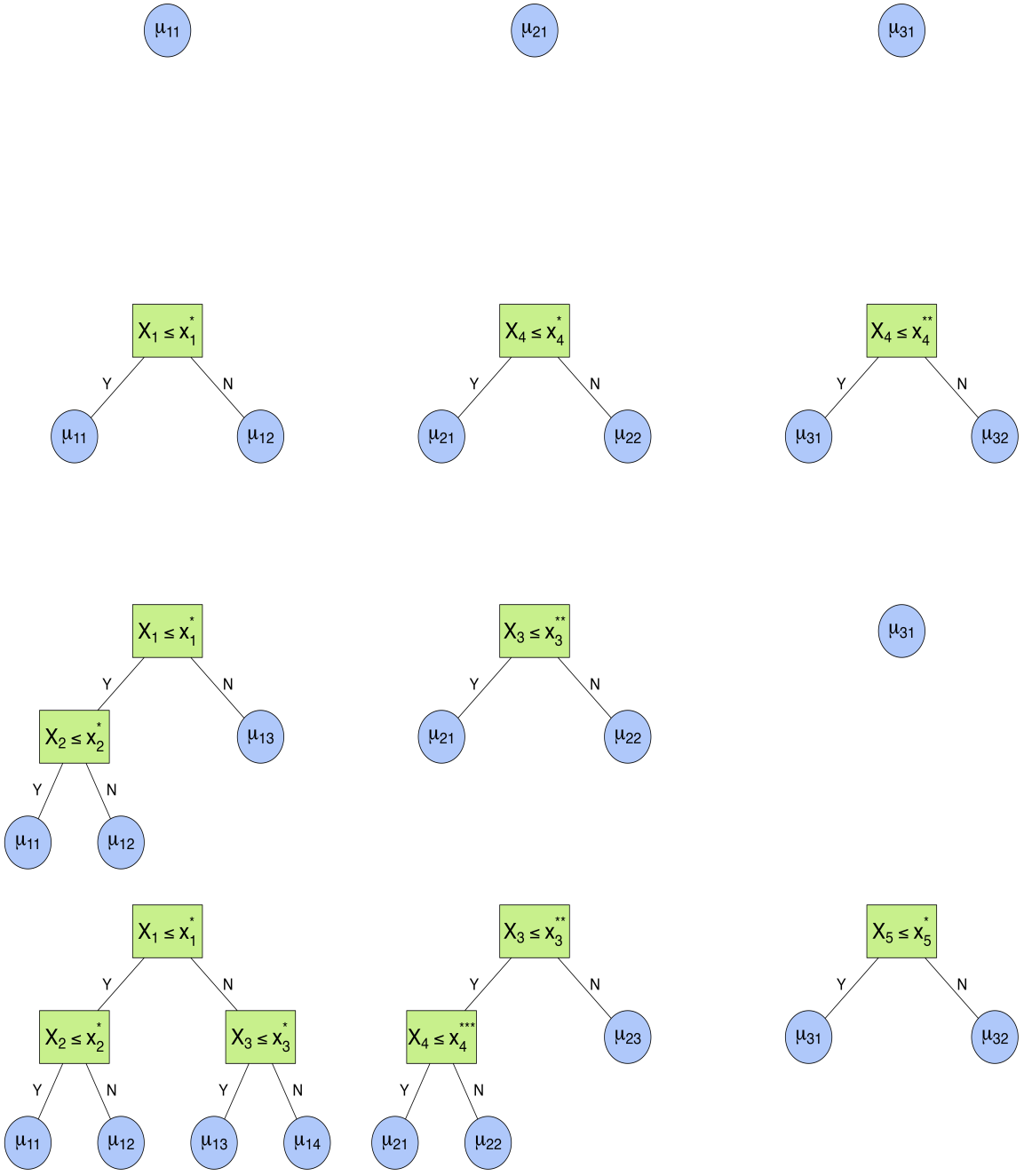

Figure 2.3: Example Branching Process over 4 iterations for a 3-tree BART ensemble

(q_1) and a proposal for the the interim leaf node parameters (q_2). Note that if q_2 is specified to be the full conditional distribution for $\mathcal{M}$, then cancellation occurs with the posterior distribution and the acceptance ratio simplifies to that in (2.6).

$$\begin{aligned} r_{\mathcal{T}} &= \frac{\pi\left(\mathcal{T}', \mathcal{M}^* \mid \mathcal{D}\right)}{\pi\left(\mathcal{T}, \mathcal{M} \mid \mathcal{D}\right)} \times \frac{q_1\left(\mathcal{T} \mid \mathcal{T}'\right)}{q_1\left(\mathcal{T}' \mid \mathcal{T}\right)} \times \frac{\pi\left(\mathcal{M} \mid \mathcal{T}, \mathcal{D}\right)}{\pi\left(\mathcal{M}^* \mid \mathcal{T}', \mathcal{D}\right)} \\ &= \frac{\pi\left(\mathcal{M}^* \mid \mathcal{T}', \mathcal{D}\right)}{\pi\left(\mathcal{M} \mid \mathcal{T}, \mathcal{D}\right)} \times \frac{\pi\left(\mathcal{T}' \mid \mathcal{D}\right)}{\pi\left(\mathcal{T} \mid \mathcal{D}\right)} \times \frac{q_1\left(\mathcal{T} \mid \mathcal{T}'\right)}{q_1\left(\mathcal{T}' \mid \mathcal{T}\right)} \times \frac{\pi\left(\mathcal{M} \mid \mathcal{T}, \mathcal{D}\right)}{\pi\left(\mathcal{M}^* \mid \mathcal{T}', \mathcal{D}\right)} \\ &= \frac{\pi\left(\mathcal{T}' \mid \mathcal{D}\right)}{\pi\left(\mathcal{T} \mid \mathcal{D}\right)} \times \frac{q_1\left(\mathcal{T} \mid \mathcal{T}'\right)}{q_1\left(\mathcal{T}' \mid \mathcal{T}\right)}. \end{aligned} \tag{2.6}$$

The result is that the acceptance ratio $r_{\mathcal{T}}$ does not depend on $\mathcal{M}$ or $\mathcal{M}^*$ at all, and so step 4 can be conducted independently of and prior to sampling $\mathcal{M}$ in step 5 of Algorithm 2.1. This simplification is only possible when the full conditional distribution of $\mathcal{M}$, $\pi(\mathcal{M} \mid \mathcal{T}, \mathcal{D})$, is available in closed form. This generally only occurs when the prior over leaf nodes is conditionally conjugate to the outcome model. In the case of the original BART model (2.3), the prior for the leaf nodes is chosen to be $\mu_{tl} \sim \text{Normal}(\mu_\mu, \sigma_\mu^2)$. Due to the conjugacy with the Gaussian outcome regression, $\mathcal{M}_t$ may be integrated out of the full conditional distribution for $\mathcal{T}_t$, allowing for the simplification from (2.5) to (2.6). Algorithm 2.2 formalizes this process and expands upon step 4 in Algorithm 2.1.

Algorithm 2.2 Sampling Tree Structures in the Original BART Algorithm

1: **Input:** $\mathcal{T}_t, \boldsymbol{\lambda}^t$
2: Propose a new tree structure $\mathcal{T}_t'$ from the current state $\mathcal{T}_t$ using, e.g., a GROW, PRUNE, or CHANGE move.
3: Compute $r_{\mathcal{T}}$, the Metropolis-Hastings acceptance ratio in (2.6).
4: Set $\mathcal{T}_t \leftarrow \mathcal{T}_t'$ with probability $\min(1, r_{\mathcal{T}})$.

The GROW, PRUNE, and CHANGE proposals introduced by Chipman et al. [24] are the most commonly used proposals for tree structures in BART models. The decision rules used for these moves are traditionally of the form $X \leq x^*$, where all observations having $X \leq x*$ move into the left child node, and all observations having $X > x*$ into the right. These moves

tend to perform well in a variety of settings, but some researchers have introduced other types of moves. For instance, Deshpande [29] developed BIRTH and CHANGE proposal decision rules which are more suitable to categorical covariates where rules of the form $X \leq x^*$ don't apply. This work includes rules for graph-structured covariates, which is explored further in Chapter 5. Others have developed new proposal mechanisms entirely in an attempt to more efficiently sample from the posterior distribution [90, 77].

Updating Leaf Node Parameters

In step 5 of Algorithm 2.1, new values of the leaf node parameters $\mathcal{M}_t$ may be sampled sequentially from their full conditional distribution $\pi(\mathcal{M}_t \mid \mathcal{T}_t, \boldsymbol{\lambda}^t)$. Once again, the dependence on the structures and parameters of the other $T-1$ trees is entirely captured by the partial residual $\boldsymbol{\lambda}^t$. When $\pi(\mathcal{M}_t \mid \mathcal{T}_t, \boldsymbol{\lambda}^t)$ is available in closed form, as it is in the original BART model, Gibbs sampling may be used. It is worth mentioning that this update occurs at every MCMC iteration, regardless of whether or not the proposed tree structure is accepted in the previous step.

Regularization Priors

The specification of priors over trees in the BART model is crucial to the performance of the model and to prevent over-fitting. Assuming priors on individual leaf nodes are independent conditional on their tree structure, the prior distribution for the BART model (2.3) may be factored as in (2.7).

$$\pi\left(\{\mathcal{T}_t, \mathcal{M}_t\}_{t=1}^T\right) = \prod_{t=1}^T \pi(\mathcal{T}_t)\pi(\mathcal{M}_t \mid \mathcal{T}_t) = \prod_{t=1}^T \pi(\mathcal{T}_t) \prod_{l \in \mathcal{L}(\mathcal{T}_t)} \pi(\mu_{tl} \mid \mathcal{T}_t) \tag{2.7}$$

The computation of $r_{\mathcal{T}}$ in (2.6) requires a prior distribution over trees, which we denote $\pi(\mathcal{T})$. This is typically taken to be the *branching process* of Chipman et al. [25]. The tree prior, which is assumed to be the same for all trees, has three main components:

1. Depth penalty: the probability that a given node at depth d splits is given by:

$$\alpha_{\mathcal{T}}(1+d)^{-\beta_{\mathcal{T}}}, \qquad \alpha_{\mathcal{T}} \in (0,1), \beta_{\mathcal{T}} \in [0,\infty) \tag{2.8}$$

   Here, $\alpha_{\mathcal{T}}$ controls the probability that the root node splits, and higher values of $\beta_{\mathcal{T}}$ penalizes deeper trees. Chipman et al. [25] recommend default values of $(\alpha_{\mathcal{T}}, \beta_{\mathcal{T}}) = (0.95, 2)$, which set the prior probability of trees with 1, 2, 3, 4, and over 5 leaf nodes at 0.05, 0.55, 0.28, 0.09, and 0.03, respectively.

2. Choice of splitting covariate: the probabilities of selecting a variable to split upon in a decision rule. This is typically taken to be uniform over the available covariates, though Linero [70] suggests an extension for high-dimensional settings (see Section 2.1.3).

3. Choice of splitting rule: given a splitting covariate, the probability of selecting the decision rule. This is typically taken to be uniform over the available cut points.

The depth penalty is the most important feature of the branching process prior when considering the regularization features, as it encourages shallower trees. This in turn results in BART favoring additive relationships and lower-order interactions in the estimated response surface.

Recall that the prior distribution for the leaf nodes is typically taken to be $\pi(\mu_{tl}) = \text{Normal}(\mu_{tl} \mid \mu_\mu, \sigma_\mu^2)$. This corresponds to a prior distribution for a prediction from the entire sum-of-trees ensemble to be $\text{Normal}(T\mu_\mu, T\sigma_\mu^2)$. We often let $\mu_\mu = 0$, which is reasonable when scaling the outcome to have mean zero prior to fitting the model. To prevent the prior for predictions from being dependent on the number of trees in the ensemble, and to further constrain the contributions of individual trees to be small, we usually set $\sigma_\mu = \sigma_\mu^*/\sqrt{T}$. Prior scaling of the outcome can inform the value chosen for σ_μ^*. More details surrounding this technique and alternative methods using prior distributions on σ_μ are described in Appendix A.5.

Updating Additional Parameters

Once all of the parameters related to the tree ensemble have been updated, any remaining parameters can be updated as per usual. In the original BART model 2.1, this only refers to the outcome variance σ^2, but in more complicated BART models $\boldsymbol{\eta}$ may refer to many different types of parameters and/or hyperparameters, including but not limited to a vector of regression coefficients, random effects, or even parameters from another BART ensemble. Whether or not a closed form Gibbs sampler exists for $\boldsymbol{\eta}$ depends on the parameters included in model being estimated and whether conditional conjugacy holds.

2.1.3 Extensions

Different Outcomes

BART was originally proposed for Gaussian outcomes, and by extension binary classification via a probit link function [25]. More generally, extending BART to any outcome distribution that admits a conditionally conjugate prior distribution is relatively straightforward, though it does require thoughtful consideration to preserve the effect of the regularization priors discussed in the previous section. A few recent BART developments include nonparametric extensions for multinomial logistic and count regression models [82], heteroskedastic log-normal and gamma hurdle models [74], and survival analysis models [14, 96].

Recently, Linero [71] published a generalized BART method based on RJMCMC that theoretically allows using BART to estimate any parameter in any model without the need for conditional conjugacy. For this method, it is helpful to define the partial residuals $\lambda_i^{(t)} \equiv \sum_{k \neq t} \text{Tree}(\boldsymbol{x}_i; \mathcal{T}_k, \mathcal{M}_k)$, which represents the contribution to λ_i by all trees except for $\mathcal{T}_t$, and the corresponding vector $\boldsymbol{\lambda}^{(t)} \equiv \left(\lambda_i^{(t)}, \ldots, \lambda_N^{(t)}\right)^T$.

One of the main difficulties with extending BART to new types of models is designing regularization, or *shrinkage*, priors for the leaf node parameters which are conjugate to the outcome distribution. In general, BART makes use of the integrated likelihood (2.9) when

calculating $r_{\mathcal{T}}$ to update the tree structures in Algorithms 2.1 and 2.2. When (2.9) is not tractable, the original algorithm falls apart.

$$\Lambda(\mathcal{T}_t) = p(\mathcal{D} \mid \mathcal{T}_t) = \prod_{l \in \mathcal{L}(\mathcal{T}_t)} \int \pi(\mu) \left(\prod_{i:\boldsymbol{x}_i \mapsto l} p\left(y_i \mid \lambda_i^{(t)} + \mu, \boldsymbol{\eta}\right) \right) d\mu \tag{2.9}$$

This generalized BART approach of Linero [71] is outlined in Algorithm 2.3.

Algorithm 2.3 One MCMC Iteration of the RJMCMC BART Algorithm

1: **Input:** $\mathcal{D}, \{\mathcal{T}_t, \mathcal{M}_t\}_{t=1}^T, \boldsymbol{\eta}$
2: **for** $t = 1, \ldots, T$ **do**
3: Compute the partial residuals $\boldsymbol{\lambda}^{(t)}$.
4: Propose a new tree structure $\mathcal{T}_t'$ from the current state $\mathcal{T}_t$ using, e.g., a GROW, PRUNE, or CHANGE move.
5: Propose interim values for $\mathcal{M}_t^*$.
6: Compute $r_{\mathcal{T}}$, the RJMCMC M-H acceptance ratio in (2.5) for the move from $(\mathcal{T}_t, \mathcal{M}_t)$ to $(\mathcal{T}_t', \mathcal{M}^*)$.
7: Set $\mathcal{T}_t \leftarrow \mathcal{T}_t'$ with probability $\min(1, r_{\mathcal{T}})$.
8: Sample $\mathcal{M}_t$ targeting its full conditional distribution using, say, slice sampling [84] or adaptive rejection sampling [42].
9: **end for**
10: Update any parameters included in $\boldsymbol{\eta}$, conditional on $\mathcal{D}$ and $\{\mathcal{T}_t, \mathcal{M}_t\}_{t=1}^T$.

Algorithm 2.3 differs in several ways from Algorithm 2.1. Notably, step 5 is not required in the original algorithm, and step 8 now requires advanced sampling techniques because a closed-form Gibbs sampler is not available. The lack of conjugacy also requires using $\boldsymbol{\lambda}^{(t)}$ instead of $\boldsymbol{\lambda}^t$ in steps 5-8.

To see this more closely, notice that Algorithm 2.1 could be written in terms of $\boldsymbol{\lambda}^{(t)}$, but the reverse is not true. This is because the BART parameter in the integrated likelihood used behind the scenes in Algorithm 2.3 is essentially offset by $\boldsymbol{\lambda}^{(t)}$, as opposed to using $\boldsymbol{\lambda}^t$ as the outcome. In other words, the likelihood component used in the M-H update of $\mathcal{T}_t$ for the original BART model can be changed from $p\left(\lambda_i^t \mid \mu\right)$ to $p\left(y_i \mid \lambda_i^{(t)} + \mu\right)$ (where μ is the parameter being estimated by $\mathcal{T}_t$) with no change in result since $\lambda_i^t = y_i - \lambda_i^{(t)}$. This won't always be the case, with the most obvious example being when we want to use BART to estimate individual-level parameters other than predictions. When there is no observed

outcome with which to define the partial residual λ_i^t, we will still have access to $\lambda_i^{(t)}$. The generalized BART framework allows for modeling of any parameter in a statistical model, not just predicting the observed outcome.

This approach is indeed very flexible, and is promising for the future of BART. In Chapter 3, this methodology is used to allow for a heterogeneous regression coefficient within a conditional logistic regression model. The downside to this approach is that proposing and sampling values of $\mathcal{M}^*$ in step 5 and $\mathcal{M}$ in step 8 can be challenging and time consuming. For step 8, sampling techniques such as slice sampling or adaptive rejection sampling may be used [84, 42]. A brief overview of adaptive rejection sampling is provided in Appendix A.2.4.

Dirichlet Additive Regression Trees

The BART algorithm will naturally tend towards trees with decision rules based on predictors that are most important in the true data generating process since those rules will be associated with the greatest M-H acceptance ratios. Nevertheless, the algorithm will at times choose to split on unimportant predictors due to the innate randomness of the approach. This is particularly an issue in high-dimensional settings where only a fraction of the predictors are actually useful. Linero [70] proposed using an additional hyperprior on the selection of predictors to use within a splitting rule. The original model uniformly selects a single predictor from the available list of predictors which is to be used for a splitting rule. In other words, $S_p = \frac{1}{P}$, where P is the number of predictors. Linero [70] instead specifies a hierarchical prior distribution on predictor selection probabilities as in (2.10)–(2.12):

$$\{u_p\}_{p=1}^P \sim \text{Multinomial}\left(N_b, \{S_p\}_{p=1}^P\right) \tag{2.10}$$

$$\{S_p\}_{p=1}^P \sim \text{Dirichlet}\left(\frac{\alpha}{P}\right) \tag{2.11}$$

$$\frac{\alpha}{\alpha + P} \sim \text{Beta}(1, 0.5) \tag{2.12}$$

where u_p is the number of splits in the ensemble based on X_p, and N_b is the total number of splits in the ensemble in the current iteration. This hierarchical formulation leads to the conjugate update for $\{S_p\}_{p=1}^P$ provided in (2.13).

$$\{S_p\}_{p=1}^P \mid \{u_p\}_{p=1}^P \sim \text{Dirichlet}\left(\frac{\alpha}{P} + u_1, \ldots, \frac{\alpha}{P} + u_P\right) \tag{2.13}$$

This proposed Dirichlet prior more effectively filters out unimportant predictors by encouraging the algorithm to split on predictors which have previously been used. In practice this update is often initiated half-way through the burn-in period to allow the ensemble to reach a "good" point before tuning variable selection probabilities.

Soft Bayesian Additive Regression Trees

BART, like other tree-based ensemble machine learning methods, can be used to learn all types of response surfaces. However, the usage of binary trees forces the resulting fit to be rigid. When enough trees are used, the model might be able to approximate a smooth function well enough, but the resulting fit will never truly be smooth.

Recently, Linero and Yang [73] proposed an modification to the algorithm coined *soft* BART. This modification replaces the deterministic decision trees traditionally used within BART ensembles with soft decision trees [122]. In a soft decision tree, the prediction for an observation is a weighted average of all of the leaf node parameters within the tree, rather than just a single leaf node parameter. The weights are defined as the probability that the observation is mapped to each leaf node as determined by, say, a logistic gating function [73]. When compared to the deterministic predictions obtained from a traditional decision tree, this has the effect of smoothing over the otherwise rigid decision rules that form the binary tree. Mathematically, $\boldsymbol{x}$ is mapped to leaf $l \in \mathcal{L}(\mathcal{T})$ with probability (2.14)

$$\phi(\boldsymbol{x}; \mathcal{T}, l) = \prod_{b \in A(l)} \psi(\boldsymbol{x}; \mathcal{T}, b)^{1-R_b} \left\{1 - \psi(\boldsymbol{x}; \mathcal{T}, b)\right\}^{R_b}, \tag{2.14}$$

where $A(l)$ is the set of interior nodes ancestral to leaf node l and $R_b = 1$ if the path to l goes right at branch node $b \in \mathcal{B}(\mathcal{T})$. The choice of ψ is important, and the original authors suggest (2.15)

$$\psi(\boldsymbol{x}; \mathcal{T}, b) = \psi\left(\frac{C_b - x_p}{\tau_b}\right), \tag{2.15}$$

where C_b is the cut point for variable X_p used in the split for branch node b, τ_b is a bandwidth probability with higher values resulting in a smoother fit, and $\psi(x) = (1 + \exp(-x))^{-1}$ is the logistic gating function previously mentioned. More information regarding this approach, including details for how to update and specify τ_b, can be found at the original source [73]. The primary drawback to using soft BART is the greatly increased computational burden that comes as a result of having to complete traversals over the entire tree for all observations.

This formulation results in predictions for observations with values close to, but on opposite sides of a split C_b to have more similar predictions than observations further away from one another in the covariate space. When working with environmental data, it is not often the case that small changes in an exposure should result in sudden, large changes in the exposure-response surface, so soft BART will prove useful in this setting.

2.2 Interpretable BART

BART is similar to other machine learning methods in that the ultimate result is a so-called "black-box" from which it can be difficult to explain how the fitted model is generating predictions from its inputs. To combat this, researchers use a suite of tools to quantify the relative importance, as well as join and marginal effects of input variables on the ensemble output.

2.2.1 Variable Importance

One common strategy for summarizing a fitted BART model is to tabulate the usage of each input variable in the ensemble. For any given MCMC iteration, one might count the number

of splits across all trees based on each covariate, and report the proportions of all splits based on each covariate. This can be done for all posterior samples, providing an estimate of uncertainty to go along with a posterior mean. Similarly, one might report posterior inclusion proportions, calculated as the proportion of all posterior samples which use each variable within the ensemble. For large T, both approaches will struggle due to unimportant variables entering the model unless a sparsity inducing prior such as the one described in Section 2.1.3. This approach tends to favor continuous variables, as they admit more valid split values.

2.2.2 Partial Dependence

When it comes to describing marginal effects of one variable, or joint effects of multiple variables, a common tool in the machine learning literature is *partial dependence*. Partial dependence statistics were originally introduced by Friedman [36] and serve as a way to quantify the impact of shifting one or more predictors on the estimated prediction function $\hat{f}$. The partial dependence function for a single predictor X_p evaluated at x_p is given by (2.16)

$$f_{p,PD}(x_p) \equiv \mathbb{E}\left[\ \hat{f}\left(x_p, \mathbf{X}_{-p}\right)\right], \tag{2.16}$$

where $\mathbf{X}_{-p}$ represents all but the p^{th} predictor. It's important to reiterate that partial dependence functions are defined in terms of the estimated prediction function, $\hat{f}$, and not the true function f. Thus, this approach is simply a way to better understand a fitted model, and it is not a tool to be used for determining whether the model is performing well. This quantity is estimated by averaging evaluations of $\hat{f}$ over the sample as in (2.17).

$$\hat{f}_{p,PD}(x_p) = \frac{1}{N}\sum_{i=1}^{N} \hat{f}\left(x_p, \boldsymbol{x}_{i,-p}\right) \tag{2.17}$$

The estimator tracks the average prediction in the sample as if all observations were to have $X_p = x_p$. This quantity is typically calculated for a range of values of X_p, most commonly running from the minimum observed value of X_p, $x_{min,p}$, to the maximum observed

value, $x_{max,p}$. For continuous X_p, these values are traditionally chosen to be either equally spaced in this range or perhaps to align with quantiles of the observations. The result may be plotted as a "dose-response" or "exposure-response" curve. For discrete X_p, one might compute (2.17) for each distinct value in X_p, or for the difference between two values.

Poor Man's Partial Dependence

A major drawback of using partial dependence is the need to evaluate $\hat{f}$ for all observations, for all values of X_p under consideration. This can be time consuming for large datasets or for models that require more time to generate predictions (e.g., time spent traversing many tree structures across many posterior samples for BART). Alternatively, one might fix the other variables $\mathbf{X}_{-p}$ to a typical value, such as their respective medians. This approximation to the partial dependence method is sometimes referred to as poor man's partial dependence, or fixed value partial dependence. The estimand is given by (2.18)

$$f_{p,Fixed}(x_p) \equiv \mathbb{E}\left[\ \hat{f}\left(x_p, \boldsymbol{x}^*_{-p}\right)\right], \tag{2.18}$$

where $\boldsymbol{x}^*_{-p}$ represents the observed medians (or some other fixed value) for each predictor except for X_p. The quantity is estimated by (2.19).

$$\hat{f}_{p,Fixed}(x_p) = \hat{f}\left(x_p, \boldsymbol{x}^*_{-p}\right) \tag{2.19}$$

Observe that (2.19) is much faster to compute than (2.17) since it only requires one evaluation of $\hat{f}$ for each value of X_p, whereas the number of computations required for (2.17) scales linearly with N. The downside with the fixed values approach is that the result does not average over the predictors that aren't of interest.

2.2.3 Accumulated Local Effects

A undesirable feature of both the partial dependence and fixed values approaches is that the resulting estimates may involve extrapolating the prediction function. An alternative approach that addresses these concerns and bridges the computation gap between the two is known as *accumulated local effects*, or ALE [6]. The true uncentered ALE main effect for X_p evaluated at x_p is defined as the quantity in (2.20).

$$f_{p,ALE}(x_p) \equiv \int_{x_{min,p}}^{x_p} \mathbb{E}\left[\frac{\partial \hat{f}}{\partial X_p}(x_p, \mathbf{X}_{-p}) \mid X_p = x'_p\right] dx'_p \tag{2.20}$$

The idea behind this estimand is threefold:

1. Isolate the partial effect of shifting X_p by targeting the partial derivative of $\hat{f}$ with respect to X_p.

2. Avoid extrapolation in the evaluation of $\hat{f}$ by taking the expectation over the conditional distribution of X_p.

3. Visualize by accumulating local changes in $\hat{f}$ corresponding to small incremental changes in X_p.

To estimate the quantity in (2.20), we first rewrite the estimand using the limit definition of the derivative as in (2.21).

$$f_{p,ALE}(x_p) \equiv \lim_{K\to\infty} \sum_{k=1}^{k_p^K(x_p)} \mathbb{E}\left[\hat{f}\left(x_{k,p}^K, \mathbf{X}_{-p}\right) - \hat{f}\left(x_{k-1,p}^K, \mathbf{X}_{-p}\right) \mid X_p \in \left(x_{k-1,p}^K, x_{k,p}^K\right]\right] \tag{2.21}$$

Here, K is the number of intervals in the support of X_p over which local effects are estimated, and $x_{k-1,p}^K$ and $x_{k,p}^K$ are the lower and upper bounds of the k^{th} interval. The

estimator of the uncentered ALE main effect is given by (4.18)

$$\hat{f}_{p,ALE}(x_p) = \sum_{k=1}^{k_p^K(x_p)} \frac{1}{N_k} \sum_{i:x_{i,p} \in \left(x_{k-1,p}^K, x_{k,p}^K\right]} \left[\hat{f}\left(x_{k,p}^K, \boldsymbol{x}_{i,-p}\right) - \hat{f}\left(x_{k-1,p}^K, \boldsymbol{x}_{i,-p}\right)\right], \tag{2.22}$$

where N_k is the number of observations having $X_p \in \left(x_{k-1,p}^K, x_{k,p}^K\right]$. In practice we compute estimates of local effects for a fixed number of intervals (K). The ALE is also typically centered vertically by overall average prediction. This results in the interpretation of the estimates to become relative to the average prediction, as opposed to the prediction for the lowest observed value of X_p.

Accumulated Local Effects: A Brief Illustration

Suppose two correlated variables, X_1 and X_2, are observed and the true bivariate response surface is given by $f(X_1, X_2) = -X_2(X_1 - 0.5)^2$. The resulting surface might resemble Figure 2.4.

Overlaid on the surface in 2.4 are 25 observations. The data have been simulated so that there is a strong positive correlation between X_1 and X_2. This is the exact type of scenario where ALE shines. Figure 2.5 walks through the steps necessary to compute and plot the ALE.

Figure 2.6 plots the ALE main effect of X_1 on $\hat{f}$ alongside the partial dependence and fixed values (median) approach for estimating the main effect of X_1 on $\hat{f}$. The latter two approaches are similar to one another, but very different from the ALE. For small values of X_1, the ALE main effect estimate is much higher. We can see why this occurs by examining Figure 2.4. For this same lower range of values of X_1, from about 0 to 0.25, the surface value is relatively higher in the region where data was observed. During the estimation process, the partial dependence approach extrapolates to the upper left quadrant and the fixed values approach extrapolates to the median value of X_2 (≈ 0.50). The result is that the main effect estimate for X_1 in this lower range has a downward bias when using either of these

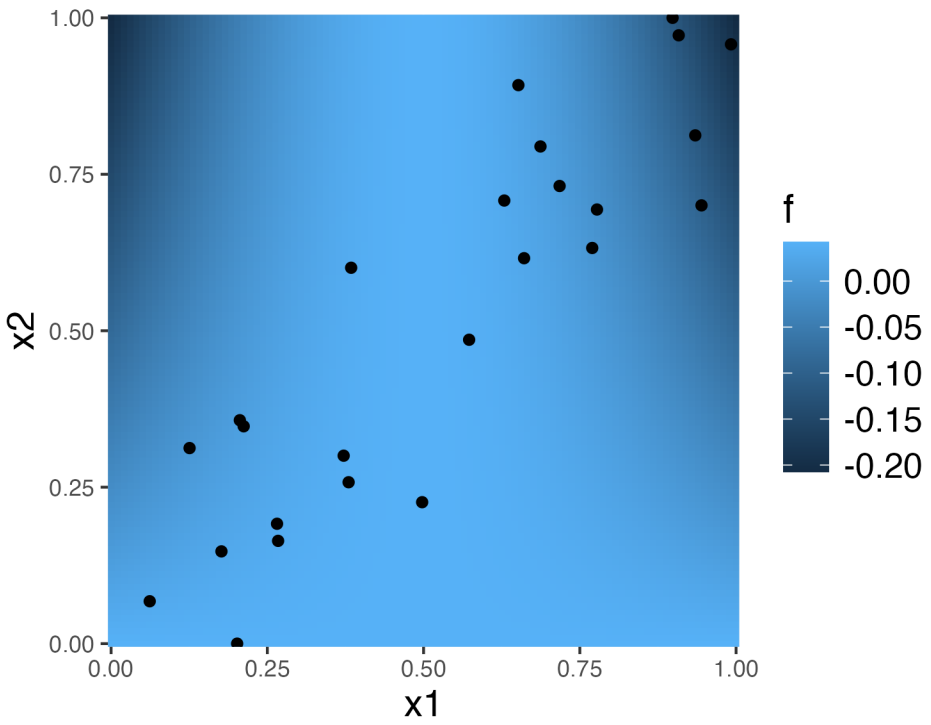

Figure 2.4: True Bivariate Response Surface for ALE Illustration

approaches. The reverse is true for the upper range of X_1 (from about 0.75 to 1). Unlike partial dependence and fixed values approach, the ALE only computes the main effect for X_1 using realistic values of X_2, providing the result with a more reliable interpretation.

Partial dependence and the fixed values approach are easily extendable to more than one exposure of interest, while ALE requires a bit more work (see Apley and Zhu [6] for more details on this). For continuous variables, it is not common to estimate any of these quantities for more than 2 variables since interpretation and visualization becomes very complicated. Regardless of the method chosen for estimating partial effects, the accuracy of the result hinges entirely on the estimated prediction function $\hat{f}$ capturing the true response surface f.

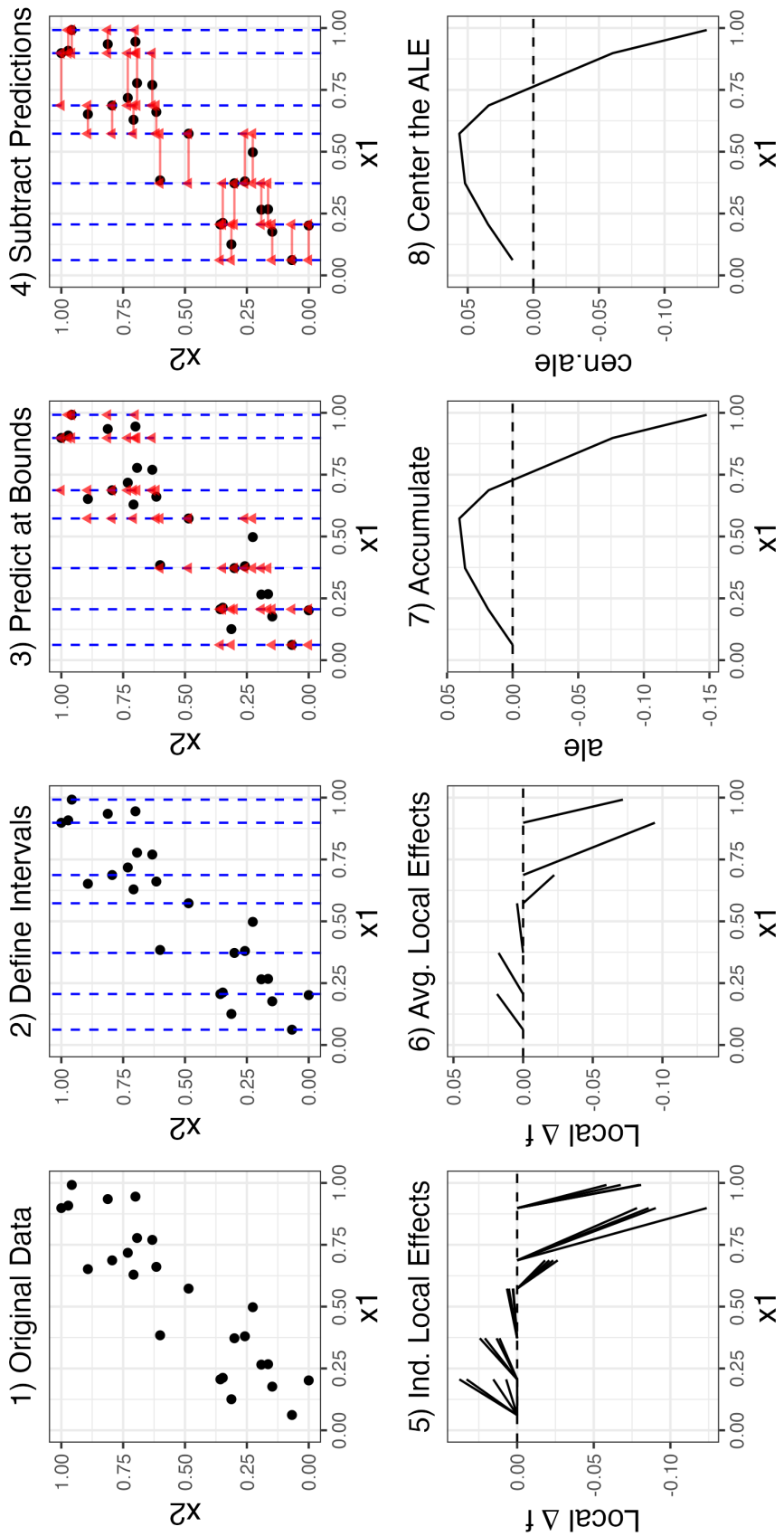

Figure 2.5: ALE Computation Walkthrough

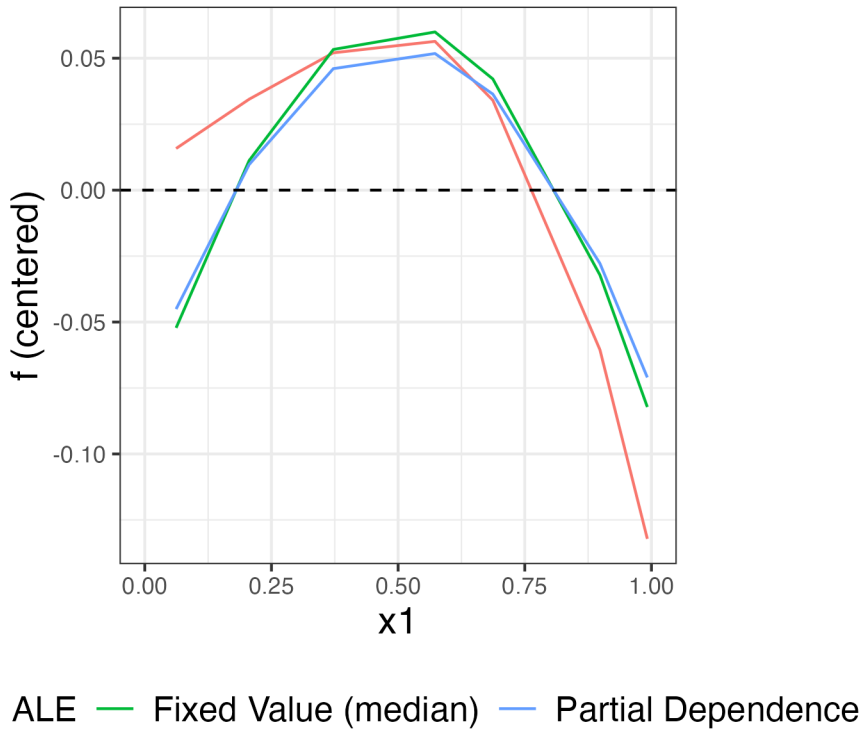

Figure 2.6: Comparison of Approaches for Estimating the Partial Effect of X_1 on $\hat{f}$

Chapter 3

Estimating Heterogeneous Exposure Effects in the Case-Crossover Design using BART

3.1 Introduction

In the United States, Alzheimer's disease (AD) affected 6.7 million people aged 65 and older in 2023, with that number projected to more than double by 2060. AD is the most common cause of dementia, entirely or partially responsible for 60-80% of all cases. People with AD often struggle to communicate and complete tasks in their daily life due to a host of symptoms headlined by forgetfulness, lethargy, and confusion. An estimated 1.3% of emergency department (ED) visits involve people with AD and related dementia, and within this population the number of ED visits per 1,000 Medicare beneficiaries increased 28% from 2008 to 2018 - outpacing cancer, stroke, and heart failure [3].

In recent years, extreme heat has been associated with elevated risk of ED visit, hospitalization, and death among people with AD and dementia in Spain, Australia, Germany, and the United States [26, 118, 37, 109, 120]. Fritze [37] reported that the number of comorbid

conditions is associated with increased risk of mortality among people with dementia.

There are several potential explanations for why people with AD are more affected by extreme heat. People with AD may have elevated core body temperature due to disturbed circadian rhythms responsible for thermoregulation [106, 91, 48, 59]. Alternatively, people with AD tend to wander or get lost, resulting in prolonged exposure to extreme temperatures [3]. Another possibility is that these individuals may struggle to communicate their heat-related discomfort in certain situations with their caregivers [105].

These explanations may not apply to the entire AD population. People with AD and related dementia are 2.7 times more likely to have 4 or more additional chronic conditions compared to people without AD or dementia; in the United States, 56% have hypertension, 46% have chronic kidney disease (CKD), 37% have diabetes, 34% have congestive heart failure (CHF), and 20% have chronic obstructive pulmonary disease (COPD) [3]. Additionally, an estimated 12.7% of people with AD have clinically diagnosed depression [23]. Given the variety of concomitant diagnoses these individuals tend to have, it is possible that heterogeneity exists in the exposure-response relationship describing heat wave-related morbidity. Thus, studying modifiers of this relationship is of interest.

Some of the aforementioned studies model heterogeneous exposure effects via stratification or by interacting covariates with the exposure [37, 120], while others directly target these effects using the case-only approach of Armstrong [7] [118]. In the extreme temperature literature, stratification has been mostly applied for demographic characteristics (e.g., age, race, and sex), while the case-only approach has also been used for chronic conditions, socioeconomic status, and various census tract characteristics [92, 119, 117, 76]. These approaches to heterogeneous effects estimation are limited by the need for expert knowledge regarding which factors are important before model-fitting, and they are not readily capable of identifying complex interactions among potential effect moderators.

We propose an extension of the popular case-crossover study design to estimate heterogeneous exposure effects using Bayesian additive regression trees (BART) [25]. As it is typically

applied, this design is limited to the previously mentioned strategies for heterogeneous effects estimation. The proposed method, CL-BART, uses BART within the case-crossover design to flexibly learn potentially complicated heterogeneous exposure-response relationships during the model-fitting process, with minimal prespecification required. In Section 3.2 we introduce the data for the motivating application. In Section 3.3 we review the case-crossover design, conditional logistic regression, and BART. We then develop CL-BART, focusing on the reversible jump portion of the estimation algorithm. In Section 3.3.3 we describe strategies for drawing posterior inference from the proposed model. In Section 3.4, we conduct two simulations illustrating the performance of CL-BART, and in Section 3.5 we apply CL-BART to estimate the effects of heat waves on ED visits among people with AD in California. Finally, in Section 3.6 we summarize our findings, discuss the limitations of the approach, and suggest possibilities for future improvements.

3.2 Data

3.2.1 Health Data

The data for our motivating application includes all ED visits among people with AD in California occurring from 2005 to 2015. These data were obtained from the California Office of Statewide Health Planning and Development (now California Department of Health Care Access and Information), and include patients' visit date, sex, age, race, ethnicity, residential ZIP code, and diagnosis codes based on the International Classification of Diseases. We restrict the ED visit records to include only those who had either a primary or secondary diagnosis of AD. Diagnoses of comorbid conditions were also based on the presence of any diagnosis code for CHF, CKD, COPD, depression, diabetes, hypertension, and hyperlipidemia (see the Chapter 3 Supplementary Materials for a list of codes).

3.2.2 Exposure Data

Meteorology data were obtained from Daymet [104]. The 1km x 1km data product was spatially averaged within each ZIP code, and linked to the ED visit data by both date and ZIP code. Specifically, we use the daily average temperature (°C) and dew-point temperature (°C). The former is calculated as the arithmetic mean of the daily minimum and maximum temperature, and the latter is derived from water vapor pressure using the Magnus formula presented in Sonntag [95]. The exposure of interest, heat wave, is defined as any sequence of two or more days at or above the ZIP code-specific 95th percentile of daily average temperature (excluding the first day of such a sequence to better reflect sustained heat exposure). Daily average temperature, dew-point temperature, and a US federal holiday indicator were also treated as potential confounders in the health model.

3.3 Methods

3.3.1 Model Development

Review of the Case-Crossover Design

In environmental epidemiology studies, it is common to only observe *cases*, or events associated with some health outcome. For instance, we might observe visits to an emergency department or hospitalizations, but we generally do not observe anything for healthy individuals, or on days when events do not occur. Because of this, clever matching schemes are often employed to select *controls*, or observations for which no event was recorded.

Case-crossover designs are frequently used to analyze the effects of short-term exposure on health outcomes in large environmental epidemiology studies when only cases are available [18]. This design allows for estimation of associations between an outcome and time-varying exposures of interest while avoiding the need to adjust for confounding by time-invariant covariates that may be difficult or impossible to measure. In the case-crossover design, each

observed case is matched to a set of controls within a *referent window* to create a stratum. There are many options for selecting referent windows, but the most popular approach is the *time-stratified* design (see Appendix A.1 for further details regarding the various options). This strategy matches each case to the 3-4 other dates in a calendar month with the same day of the week. A crucial assumption of the time-stratified approach is that the observed cases are independent and rare enough such that an individual would not experience the event twice within the referent window. The time-stratified design is a popular choice for selecting referent windows because it is both localizable and ignorable, thus providing unbiased estimation of regression coefficients when using conditional logistic regression [53].

Review of Conditional Logistic Regression

Suppose we observe N cases. Each of the $i = 1, \ldots, N$ individuals is assigned a referent window, $\mathcal{W}_i$, of time points based on the time-stratified design. We note whether a case was observed for the individual at each time point $j \in \mathcal{W}_i$ and record it as either $Y_{ij} = 1$ (case) or $Y_{ij} = 0$ (no case). Then the true data generating model might be represented as:

$$Y_{ij} \sim \text{Bernoulli}(p_{ij}), \qquad j \in \mathcal{W}_i \tag{3.1}$$

$$\text{logit}(p_{ij}) = \boldsymbol{v}_i^T \boldsymbol{\alpha} + \boldsymbol{w}_{ij}^T \boldsymbol{\gamma} + \beta x_{ij}, \tag{3.2}$$

where Y_{ij} is the outcome for individual i at time j, p_{ij} is the probability of observing $Y_{ij} = 1$, and $\mathcal{W}_i$ is the referent window containing observation times j for individual i. The time-varying primary exposure is denoted by x_{ij}, while $\boldsymbol{v}_i$ and $\boldsymbol{w}_{ij}$ represent column vectors which include time-invariant and time-varying confounders of the exposure-response relationship, respectively. For the AD example, x_{ij} is a binary heat wave indicator, $\boldsymbol{v}_i$ includes time-invariant demographic information or other unmeasured quantities, and $\boldsymbol{w}_{ij}$ includes daily average temperature, dew-point temperature, and a federal holiday indicator. The parameters $\boldsymbol{\alpha}, \boldsymbol{\gamma}$, and β represent log odds ratios quantifying the association between the

various predictor variables and the outcome.

Integral to the case-crossover design is the assumption that $\sum_{j \in \mathcal{W}_i} Y_{ij} = 1$ for $i = 1, \ldots, N$ (i.e., individuals experience exactly one event within their referent window). Given this information, the conditional probability of observing a case at time point j for observation i is given by (3.3).

$$\begin{aligned} p_{ij}^c &= \Pr\left(Y_{ij} = 1 \mid \sum_{j \in \mathcal{W}_i} Y_{ij} = 1\right) \\ &= \frac{\exp\left(\boldsymbol{v}_i^T \boldsymbol{\alpha} + \boldsymbol{w}_{ij}^T \boldsymbol{\gamma} + \beta x_{ij}\right)}{\sum_{j \in \mathcal{W}_i} \exp\left(\boldsymbol{v}_i^T \boldsymbol{\alpha} + \boldsymbol{w}_{ij}^T \boldsymbol{\gamma} + \beta x_{ij}\right)} \\ &= \frac{\exp\left(\boldsymbol{w}_{ij}^T \boldsymbol{\gamma} + \beta x_{ij}\right)}{\sum_{j \in \mathcal{W}_i} \exp\left(\boldsymbol{w}_{ij}^T \boldsymbol{\gamma} + \beta x_{ij}\right)}. \end{aligned} \tag{3.3}$$

Notably, all time-invariant covariates ($\boldsymbol{v}_i$) have been conditioned out entirely. This is the primary benefit of working with the conditional likelihood, but it comes at the expense of being able to produce unconditional probability predictions. The conditional likelihood contribution for a single individual corresponds to a multinomial distribution with probabilities given by (3.3). With this in mind, the conditional likelihood for the observed data is is expressed as:

$$\begin{aligned} p(\boldsymbol{y} \mid \beta, \boldsymbol{\gamma}) &= \prod_{i=1}^{N} p(\boldsymbol{y}_i \mid \beta, \boldsymbol{\gamma}) \\ &= \prod_{i=1}^{N} \prod_{j \in \mathcal{W}_i} \left(p_{ij}^c\right)^{y_{ij}} \\ &= \prod_{i=1}^{N} \prod_{j \in \mathcal{W}_i} \left(\frac{\exp\left(\boldsymbol{w}_{ij}^T \boldsymbol{\gamma} + \beta x_{ij}\right)}{\sum_{j \in \mathcal{W}_i} \exp\left(\boldsymbol{w}_{ij}^T \boldsymbol{\gamma} + \beta x_{ij}\right)}\right)^{y_{ij}} \\ &= \prod_{i=1}^{N} \frac{\exp\left(\boldsymbol{w}_{ij_i}^T \boldsymbol{\gamma} + \beta x_{ij_i}\right)}{\sum_{j \in \mathcal{W}_i} \exp\left(\boldsymbol{w}_{ij}^T \boldsymbol{\gamma} + \beta x_{ij}\right)}, \end{aligned} \tag{3.4}$$

where $\boldsymbol{y}_i$ is the observed vector of outcomes within the i^{th} individual's referent window $\mathcal{W}_i$, and $\boldsymbol{y}$ is a vector containing all $\boldsymbol{y}_i$, $i = 1, \ldots, N$. For the final form of $p(\boldsymbol{y} \mid \beta, \boldsymbol{\gamma})$ in (3.4), we

implicitly include the outcome through the subscript ij_i, which represents the index time point of the i^{th} case. This notation relies on the fact that only one of the y_{ij} is equal to 1 for each individual. Maximum likelihood estimation of (3.4) results in unbiased log odds ratios (ORs) for the confounders ($\boldsymbol{\gamma}$) and the primary exposure (β). Alternatively, the parameters may be estimated using Markov chain Monte Carlo sampling in a Bayesian framework.

The data model in (3.4) assumes a homogeneous exposure effect, β, across all individuals. To examine how the association between the exposure and outcome varies across individuals, researchers may specify subgroup analyses ahead of time, defining the subgroups using demographic characteristics like sex and age. This requires some knowledge of the outcome and exposure to be able to identify which subgroups should be considered prior to analyzing the data.

CL-BART and the Exposure Moderating Function $\beta(\cdot)$

We propose extending this modeling framework to allow for estimation of heterogeneous exposure effects within a study population. Specifically, we use BART [25] to model the exposure effect as a function of individual-level covariates that were previously conditioned out. We start by defining a more general version of the conditional likelihood in (3.5), which we will refer to as the conditional logistic BART (CL-BART) likelihood.

$$p\left(\boldsymbol{y} \mid \beta(\cdot), \boldsymbol{\gamma}\right) = \prod_{i=1}^{N} p\left(\boldsymbol{y}_i \mid \beta(\boldsymbol{z}_i), \boldsymbol{\gamma}\right) = \prod_{i=1}^{N} \frac{\exp\left(\boldsymbol{w}_{ij_i}^T \boldsymbol{\gamma} + \beta(\boldsymbol{z}_i) x_{ij_i}\right)}{\sum_{j \in \mathcal{W}_i} \exp\left(\boldsymbol{w}_{ij}^T \boldsymbol{\gamma} + \beta(\boldsymbol{z}_i) x_{ij}\right)}. \tag{3.5}$$

Here we have simply replaced β with $\beta(\boldsymbol{z}_i)$, suggesting that the increase in the log odds of an ED visit due to a unit increase in the primary exposure may differ across individuals. The contents of $\boldsymbol{z}_i$ may overlap with $\boldsymbol{v}_i$ in (3.2), but the two are not required to be identical. In the AD example, $\boldsymbol{z}_i$ includes comorbid conditions, such as diabetes and chronic kidney disease, as well as sex and age.

We place a nonparametric BART prior on the exposure moderating function $\beta(\cdot)$ as in

(3.6).

$$\beta(\boldsymbol{z}_i) = \sum_{t=1}^{T} \text{Tree}(\boldsymbol{z}_i; \mathcal{T}_t, \mathcal{M}_t). \tag{3.6}$$

The BART prior represents $\beta(\boldsymbol{z}_i)$ as a sum of T weak learners - in this case, Bayesian regression trees [24]. Each tree is composed of a tree structure $\mathcal{T}$ defined by a series of binary splits based on covariates $\boldsymbol{z}$, a set of terminal or *leaf* nodes $\mathcal{L}(\mathcal{T})$, and a set of scalar-valued leaf node parameters $\mathcal{M} = \{\mu_l\}_{l \in \mathcal{L}(\mathcal{T})}$. In (3.6), "Tree" is the function that makes a prediction for covariate vector $\boldsymbol{z}$ by mapping $\boldsymbol{z}$ to a single leaf node in the given tree.

In the simplest setting, $\boldsymbol{z}$ consists only of a series of P_z binary effect moderators and the maximum number of unique values of $\beta(\boldsymbol{z})$ is 2^{P_z}, regardless of sample size. When $T = 1$, CL-BART simplifies to a treed conditional logistic regression, where the confounder effects are shared across leaf nodes. Including continuous covariates is a straightforward extension, and allows for modeling more complex high-order interactions and nonlinearities among exposure effect moderators. We do not consider time-varying covariates in $\boldsymbol{z}$, as this would result in individual strata being allocated to multiple leaf nodes, thus violating the case-crossover design.

3.3.2 Estimation

Generalized BART

BART was originally designed with Gaussian outcomes in mind, relying heavily on the conditional conjugacy between the outcome model and the prior distribution on the leaf node parameters. This allows for a Metropolis-Hastings (M-H) proposal for the tree structure to be conducted separately from the Gibbs update of the leaf node parameters through marginalization, resulting in a simple and efficient Markov chain Monte Carlo (MCMC) algorithm [25]. BART has since been extended to other outcome regressions, including survival, log-linear, and gamma models [96, 82, 74], but such extensions require extensive modification of the original algorithm and existing software. BART has also been used to

model varying coefficients, but these leverage conditional conjugacy as well [30, 46]. As mentioned in 2.1.3, Linero [71] recently proposed a general strategy based on reversible jump MCMC (RJMCMC) [44, 43] as a promising alternative for adapting BART to more complicated likelihoods. This approach is appealing because it avoids the need for conjugate priors altogether, and so we use it for CL-BART. We now provide a brief overview of this approach and our implementation, but refer the reader to the source for further detail.

Data Likelihood

It is first helpful to rewrite the data likelihood in terms of the tree structure. The likelihood for a single tree $\mathcal{T}_t$ can be represented as in (3.7).

$$p(\boldsymbol{y} \mid \mathcal{T}_t, \mathcal{M}_t) = \prod_{l \in \mathcal{L}(\mathcal{T}_t)} \prod_{i : \boldsymbol{z}_i \mapsto l} p\left(\boldsymbol{y}_i \mid \mu_l\right). \tag{3.7}$$

For CL-BART, we may substitute the likelihood given in (3.4), where μ_l represents the prediction from $\mathcal{T}_t$ (i.e., the exposure effect) for strata having $\boldsymbol{z}_i$ mapped ($\mapsto$) to leaf node l. Since $\boldsymbol{\gamma}$ is shared across all leaf nodes, we omit it in (3.7) to lighten the notation.

Prior Distribution

The unknown quantities for each tree in CL-BART include the leaf node parameters $\mathcal{M}_t = \{\mu_l\}_{l \in \mathcal{L}(\mathcal{T}_t)}$ and the tree structure itself $\mathcal{T}_t$. By imposing independence on the former, we may factor the joint prior distribution for a single tree as in (3.8).

$$\pi\left(\mathcal{T}_t, \mathcal{M}_t\right) \sim \pi(\mathcal{T}_t)\pi(\mathcal{M}_t \mid \mathcal{T}_t) \sim \pi(\mathcal{T}_t) \prod_{l \in \mathcal{L}(\mathcal{T}_t)} \pi\left(\mu_l\right). \tag{3.8}$$

The μ_l are given i.i.d. Normal$(0, \sigma_\mu^2)$ priors, but this is not a requirement since conjugacy with the likelihood is no longer a concern. While one may have some intuition regarding the range of log odds ratio values to expect for $\beta(\cdot)$, generally this will be unknown. For

this reason, we follow Linero [71] and specify a half-Cauchy hyperprior $\sigma_\mu \sim \mathcal{C}_+\left(0, k/\sqrt{T}\right)$ to help learn the range of appropriate predictions. Here, k is a fixed hyperparameter, and the division by $\sqrt{T}$ ensures predictions are made on the same general scale regardless of the number of trees used. Mathematical details for this update are provided in Appendix A.5.2.

For the tree structure $\mathcal{T}_t$, we use the *branching process* prior described in Chipman et al. [25], where each node in $\mathcal{T}_t$ is split with probability $\rho_d = \alpha_{\mathcal{T}}(1+d)^{-\beta_{\mathcal{T}}}$ (here d is the depth of the node in $\mathcal{T}_t$). We use the default values of $(\alpha_{\mathcal{T}}, \beta_{\mathcal{T}}) = (0.95, 2)$, but note that in the heterogeneous effects setting there have been several proponents for stronger regularization [46, 17]. We make one departure from the traditional branching process by further placing a Dirichlet hyperprior on the covariate selection probabilities as suggested in Linero [70] (see Section 2.1.3 for more detail). This modification is particularly helpful in settings with many covariates that each have many available values upon which to split.

Tree Proposals and the Posterior Distribution

New tree structures are proposed and accepted with a M-H step. We consider three types of proposals: GROW, PRUNE, and CHANGE. Both the GROW and PRUNE moves involve jumping between parameter spaces of differing dimensions, and thus require modification of the traditional M-H acceptance ratio. The general form for this ratio is given in (3.9).

$$r_{\mathcal{T}} = \underbrace{\frac{\pi\left(\mathcal{T}', \mathcal{M}'\right)}{\pi\left(\mathcal{T}, \mathcal{M}\right)}}_{\text{Prior Ratio}} \times \underbrace{\frac{p\left(\boldsymbol{y} \mid \mathcal{T}', \mathcal{M}'\right)}{p\left(\boldsymbol{y} \mid \mathcal{T}, \mathcal{M}\right)}}_{\text{Likelihood Ratio}} \times \underbrace{\frac{q\left(\mathcal{T}, \mathcal{M} \mid \mathcal{T}', \mathcal{M}'\right)}{q\left(\mathcal{T}', \mathcal{M}' \mid \mathcal{T}, \mathcal{M}\right)}}_{\text{Proposal Ratio}}. \tag{3.9}$$

The prior term may be factored as in Section 3.3.2, while the proposal term may be factored into two parts: a structural component and a proposal for the new leaf node parameter(s) based on some distribution G. We use a normal distribution based on a Laplace approximation for G, as suggested by [71] (see the Chapter 3 Supplementary Materials for details). Each type of proposal is summarized below, where $\text{NOG}(\mathcal{T})$ is defined as the set of nodes in $\mathcal{T}$ that are parents of two terminal nodes - that is, internal nodes which have no

grandchildren.

- **GROW**: a random leaf node $l \in \mathcal{L}(\mathcal{T})$ is selected. Subsequently, a splitting covariate Z_p and cut-point z_p^* based on the observed values of Z_p are selected. Then node l is split into lL and lR, where strata having $Z_p \leq z_p^*$ are fed into lL and strata having $Z_p > z_p^*$ are fed into lR. For l of depth d, the modified RJMCMC M-H acceptance ratio is given by (3.10).

$$\begin{aligned} r_{\mathcal{T}}^{GROW} = \frac{\rho_d(1-\rho_{d+1})^2}{(1-\rho_d)} \times \frac{\pi(\mu'_{lL} \mid 0, \sigma_\mu^2) \times \pi(\mu'_{lR} \mid 0, \sigma_\mu^2)}{\pi(\mu_l \mid 0, \sigma_\mu^2)} \\ \times \frac{\prod_{i:\boldsymbol{z}_i \mapsto lL} p(\boldsymbol{y}_i \mid \mu'_{lL}) \times \prod_{i:\boldsymbol{z}_i \mapsto lR} p(\boldsymbol{y}_i \mid \mu'_{lR})}{\prod_{i:\boldsymbol{z}_i \mapsto l} p(\boldsymbol{y}_i \mid \mu_l)} \\ \times \frac{p_{PRUNE}(\mathcal{T}')|\text{NOG}(\mathcal{T}')|^{-1}}{p_{GROW}(\mathcal{T})|\mathcal{L}(\mathcal{T})|^{-1}} \times \frac{G_{PRUNE}(\mu_l)}{G_{GROW}(\mu'_{lL}, \mu'_{lR})}. \end{aligned} \tag{3.10}$$

- **PRUNE**: a random node branch node $b \in \text{NOG}(\mathcal{T}_t)$ is selected. Leaf nodes bL and bR are removed from the tree, along with the decision rule that defined the branch. For b of depth d, the modified RJMCMC M-H acceptance ratio is given by (3.11).

$$\begin{aligned} r_{\mathcal{T}}^{PRUNE} = \frac{(1-\rho_d)}{\rho_d(1-\rho_{d+1})^2} \times \frac{\pi(\mu'_l \mid 0, \sigma_\mu^2)}{\pi(\mu_{lL} \mid 0, \sigma_\mu^2) \times \pi(\mu_{lR} \mid 0, \sigma_\mu^2)} \\ \times \frac{\prod_{i:\boldsymbol{z}_i \mapsto b} p(\boldsymbol{y}_i \mid \mu'_b)}{\prod_{i:\boldsymbol{z}_i \mapsto bL} p(\boldsymbol{y}_i \mid \mu_{bL}) \times \prod_{i:\boldsymbol{z}_i \mapsto bR} p(\boldsymbol{y}_i \mid \mu_{bR})} \\ \times \frac{p_{GROW}(\mathcal{T}')|\mathcal{L}(\mathcal{T}')|^{-1}}{p_{PRUNE}(\mathcal{T})|\text{NOG}(\mathcal{T})|^{-1}} \times \frac{G_{GROW}(\mu_{bL}, \mu_{bR})}{G_{PRUNE}(\mu'_b)}. \end{aligned} \tag{3.11}$$

- **CHANGE**: a random branch node $b \in \text{NOG}(\mathcal{T}_t)$ is selected. The criteria for further splitting into leaf nodes bL and bR are exchanged for another covariate and/or cut-point. Since the general tree structure is unchanged, the structural components of the prior

and proposal ratios cancel out. The M-H acceptance ratio is given by (3.12).

$$
\begin{aligned}
r_{\mathcal{T}}^{CHANGE} = & \frac{\pi(\mu'_{lL} \mid 0, \sigma_\mu^2) \times \pi(\mu'_{lR} \mid 0, \sigma_\mu^2)}{\pi(\mu_{lL} \mid 0, \sigma_\mu^2) \times \pi(\mu_{lR} \mid 0, \sigma_\mu^2)} \\
& \times \frac{\prod_{i:\boldsymbol{z}_i \mapsto bL} p(\boldsymbol{y}_i \mid \mu'_{bL}) \times \prod_{i:\boldsymbol{z}_i \mapsto bR} p(\boldsymbol{y}_i \mid \mu'_{bR})}{\prod_{i:\boldsymbol{z}_i \mapsto bL} p(\boldsymbol{y}_i \mid \mu_{bL}) \times \prod_{i:\boldsymbol{z}_i \mapsto bR} p(\boldsymbol{y}_i \mid \mu_{bR})} \\
& \times \frac{G_{CHANGE}(\mu_{bL}, \mu_{bR})}{G_{CHANGE}(\mu'_{bL}, \mu'_{bR})}.
\end{aligned} \tag{3.12}
$$

At each iteration, one type of proposal is made for each tree in the ensemble. We set the prior probability of each proposal type to $p_{GROW} = 0.3$, $p_{PRUNE} = 0.3$, and $p_{CHANGE} = 0.4$. The trees are cycled through using a generalized version of Bayesian backfitting [49, 71]. Essentially, this involves offsetting the likelihood calculation in the M-H acceptance ratio for the update of tree $\mathcal{T}_t$ by the sum of the predictions from the remaining $T - 1$ trees. Mathematically, we swap (3.7) with (3.13),

$$
p(\boldsymbol{y} \mid \mathcal{T}_t, \mathcal{M}_t) = \prod_{l \in \mathcal{L}(\mathcal{T}_t)} \prod_{i:\boldsymbol{z}_i \mapsto l} p\left(\boldsymbol{y}_i \mid \mu_l + \lambda_i^{(t)}\right), \tag{3.13}
$$

where $\lambda_i^{(t)} = \sum_{k \neq t} \text{Tree}(\boldsymbol{z}_i; \mathcal{T}_k, \mathcal{M}_k)$. The M-H acceptance ratios presented in this section only depend on the likelihood within the affected leaf nodes, and so the inner product term of (3.13) can be used wherever the likelihood component is evaluated in (3.10), (3.11), and (3.12).

Thus far for tree t, the proposed values $\mathcal{M}_t$ have been used solely to update $\mathcal{T}_t$. Once $\mathcal{T}_t$ has been updated, we propose new values for all $\mu_l \in \mathcal{M}_t$ sequentially from their full conditional distribution via adaptive rejection sampling [42].

Prior to the BART update, we update $\boldsymbol{\gamma}$ using a traditional random-walk M-H step. For the proposal distribution, we use a multivariate normal distribution centered at the current value of $\boldsymbol{\gamma}$ and with covariance matrix $\sigma_\gamma^2 \mathbf{V}_{\boldsymbol{\gamma}}$, where we initialize $\mathbf{V}_{\boldsymbol{\gamma}}$ as the confounder portion of the covariance matrix of $\hat{\boldsymbol{\gamma}}$ from the fit of a conventional conditional logistic regression as

in (3.4), and $\sigma^2_{\boldsymbol{\gamma}}$ is initially set to unity but tuned throughout the burn-in phase to achieve an optimal acceptance rate. Note that the proposal for $\{\mathcal{T}_t, \mathcal{M}_t\}$ is also offset by the confounders, in addition to the fits of other $T-1$ trees. An outline for the CL-BART MCMC algorithm is provided in Algorithm 3.4.

Algorithm 3.4 One MCMC Iteration of CL-BART

Input: $\mathcal{D} = \{\boldsymbol{y}, \mathbf{W}, \mathbf{X}, \mathbf{Z}\}, \boldsymbol{\gamma}, \{\mathcal{T}_t, \mathcal{M}_t\}_{t=1}^{T}, \alpha_{\mathcal{T}}, \beta_{\mathcal{T}}, \{S_p\}_{p=1}^{P_z}, a, \sigma^2_{\mu}, \sigma^2_{\boldsymbol{\gamma}}, \mathbf{V}_{\boldsymbol{\gamma}}$
Update $\boldsymbol{\gamma}$ (via M-H step with multivariate normal prior).
Set $\lambda_i \leftarrow \sum_{t=1}^{T} \text{Tree}(\boldsymbol{z}_i; \mathcal{T}_t, \mathcal{M}_t)$ for $i = 1, \ldots, N$.
for $t = 1$ to T **do**
    Set $\lambda_i^{(t)} \leftarrow \lambda_i - \text{Tree}(\boldsymbol{z}_i; \mathcal{T}_t, \mathcal{M}_t)$ for $i = 1, \ldots, N$.
    Propose $\mathcal{T}_t'$ from $\mathcal{T}_t$ using a GROW, PRUNE, or CHANGE step.
    Propose $\mathcal{M}_t^*$ from G based on a Laplace approximation [71].
    Compute $r_{\mathcal{T}}$, the (modified) M-H acceptance ratio for $\{\mathcal{T}_t', \mathcal{M}_t^*\}$.
    Set $\mathcal{T}_t \leftarrow \mathcal{T}_t'$ with probability $\min(1, r_{\mathcal{T}})$.
    Update $\mathcal{M}_t \mid \mathcal{T}_t, \left\{\lambda_i^{(t)}\right\}_{i=1}^{N}$ using adaptive rejection sampling [42].
    Set $\lambda_i \leftarrow \lambda_i^{(t)} + \text{Tree}(\boldsymbol{z}_i; \mathcal{T}_t, \mathcal{M}_t)$ for $i = 1, \ldots, N$.
end for
Update $\{S_p\}_{p=1}^{P_z} \sim \text{Dirichlet}(\frac{a}{P_z} + u_1, \ldots, \frac{a}{P_z} + u_{P_z})$, where u_p is the number of times Z_p is split upon.
Update a (via discrete step described in Linero [70]).
Update σ_{μ} (via M-H step with half-Cauchy prior - see Appendix A.5).

3.3.3 Posterior Inference

As with any Bayesian model, point estimates and posterior credible intervals may be obtained for the confounder coefficients and other scalar parameters. To summarize the estimated heterogeneous exposure effects, we introduce estimands similar to those presented in the BART for causal inference literature [52, 46], with the two main differences being that we are working on the log odds ratio scale, and that we have not laid out a formal statistical framework allowing us to assert the presented quantities have causal interpretations.

Initially, we estimate the average conditional exposure effect for a unit increase in the exposure as $\bar{\beta} = \frac{1}{N}\sum_{i=1}^{N} \hat{\beta}(\boldsymbol{z}_i)$. This may also be exponentiated if an odds ratio interpretation is desired. Perhaps of greater interest are the individual conditional exposure effects $\beta(\boldsymbol{z}_i)$,

$i = 1, \ldots, N$. These are numerous, so it is helpful to have strategies for summarizing them. We can easily obtain point estimates and posterior credible intervals of $\beta(\boldsymbol{z})$ for any desired set of exposure modifiers $\boldsymbol{z}$. However, the individual-level quantities can be noisy, and so it can be beneficial to instead report partial averages of conditional exposure effects, such as the partial dependence functions introduced in Friedman [36] and described in Section 2.2.2.

Define $\boldsymbol{z} = (z^1, z^2, \ldots, z^{P_z})^T$ as a $P_z \times 1$ vector of observed potential effect moderators, $\boldsymbol{z}_{\mathbf{p}}$ as the $\mathbf{p}^{th}$ component of $\boldsymbol{z}$, and $\boldsymbol{z}_{-\mathbf{p}}$ as all but the $\mathbf{p}^{th}$ component of $\boldsymbol{z}$. Note that either or both of $\boldsymbol{z}_{\mathbf{p}}$ and $\boldsymbol{z}_{-\mathbf{p}}$ might represent multiple effect moderators. The corresponding observations made on individual i are $\boldsymbol{z}_i$, $\boldsymbol{z}_{i,\mathbf{p}}$, and $\boldsymbol{z}_{i,-\mathbf{p}}$, respectively. The partial average exposure effect corresponding to setting $\mathbf{Z_p} = \boldsymbol{z}^*_{\mathbf{p}}$ is estimated as in (3.14).

$$\bar{\beta}_{\mathbf{p},PD}(\boldsymbol{z}^*_{\mathbf{p}}) = \frac{1}{N}\sum_{i=1}^{N} \hat{\beta}(\boldsymbol{z}^*_{\mathbf{p}}, \boldsymbol{z}_{i,-\mathbf{p}}). \tag{3.14}$$

One might select multiple settings of $\boldsymbol{z}^*_{\mathbf{p}}$ for comparison, where only a subset of $\boldsymbol{z}$ need be included in $\boldsymbol{z}^*_{\mathbf{p}}$, and calculate (3.14) for each setting. The resulting estimates (or any function of the estimates) can be compared across the posterior distribution. The simplest case is to select a single binary covariate, say Z_p, compute (3.14) for both levels of the covariate, and then calculate the difference in the partial dependence functions as in (3.15). The corresponding estimate represents the average difference in the (log) exposure effect due to having $Z_p = 1$ versus having $Z_p = 0$.

$$\bar{\beta}_{p,DPD} = \bar{\beta}_{p,PD}(1) - \bar{\beta}_{p,PD}(0) = \frac{1}{N}\sum_{i=1}^{N} \left[\hat{\beta}(1, \boldsymbol{z}_{i,-p}) - \hat{\beta}(0, \boldsymbol{z}_{i,-p})\right] \tag{3.15}$$

It may be difficult or computationally infeasible to perform an exhaustive comparison of all partial average exposure effects. To identify covariate values to fix during the partial averaging in (3.14), we suggest creating a lower-dimensional summary using, say, a single classification and regression tree (CART) as described by Woody et al. [116]. This involves using some subset of the input covariates $\mathbf{Z}$ to "predict" the posterior mean individual

exposure effects. We can then compute (3.14) for the combinations of covariates leading to each leaf node in the resulting CART summary.

3.3.4 Model Diagnostics

Sometimes it is helpful to have a "quick and dirty" method for establishing variable importance. One option is to check the frequencies with which the BART portion of the model splits on each of the effect moderators [25]. In general, we expect the model to favor splits on covariates which are essential to the true data generating process. However, as the size of the ensemble increases, spurious splits will be included. The sparse branching process prior of Linero [70] help to alleviate this issue in many cases. Additionally, it is important to consider correlation between the covariates, and that there may be more than one path to a good model.

Since unconditional predictions are not available when using conditional logistic regression, cross-validation based on model selection criteria that involve the outcome (e.g., RMSE or similar) do not apply. However, models can still be evaluated using likelihood-based criteria. We suggest referencing the Widely Applicable Information Criterion (WAIC), which approximates leave-one-out cross-validation [108, 40]. The WAIC uses the entire posterior distribution and all of the available data to evaluate and penalize models, and can conveniently be computed during model-fitting. This metric is useful for comparing CL-BART models with different hyperparameter specifications, such as the number of trees. Details for computing WAIC are provided in Appendix A.3.1.

Lastly, it is essential to monitor the convergence of the Markov chain samples. Since the RJMCMC algorithm performs both model selection and parameter estimation, posterior chains of individual exposure effects may not have well-mixed trace plots due to the possibility of jumping between different parameter spaces. For this reason we suggest monitoring trace plots for global parameters, such as $\bar{\beta}$, $\boldsymbol{\gamma}$, σ^2_μ, and other quantities such as the log-likelihood or average number of nodes across trees. We did not find it was necessary to run multiple

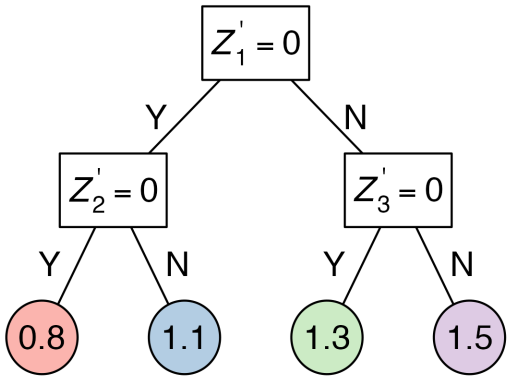

Figure 3.1: True Conditional Odds Ratios for CART Simulation

chains for the analyses described in the simulation study and application.

3.4 Simulation Study

In this section we design a simulation to mimic the case-crossover design. We follow 10,000 individuals for three years, and generate their shared exposure time-series as in (3.16).

$$X_j \sim \text{Normal}\left(\sin\left(\frac{2\pi j \times 3}{1096}\right), 1\right), \quad j = 1, \ldots, 1096. \tag{3.16}$$

Five time-varying covariates are generated as $W_1, \ldots, W_5 \overset{i.i.d.}{\sim} \text{Uniform}(0, 1)$, with odds ratios 0.5, 0.8, 1.0, 1.2, and 2.0. The probability of individual i experiencing the event at time j is calculated as $p_{ij} = \text{expit}\left(\alpha + \boldsymbol{w}_{ij}^T\boldsymbol{\gamma} + \beta(\boldsymbol{z}_i)x_j\right)$, where we set $\alpha = -8$ to ensure rare events, and the true $\beta(\boldsymbol{z}_i)$ for each individual i is specified under two deterministic scenarios:

1. **CART**: 10 binary covariates are generated as $[Z_1, Z_2, \ldots, Z_{10}]^T \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ has an AR-1 structure (i.e., $\boldsymbol{\Sigma}_{p,p'} = 0.6^{|p-p'|}$). Three of these 10 covariates are randomly selected (Z_1', Z_2', Z_3') and odds ratio $\exp[\beta(\boldsymbol{z}_i)]$ is given one of four values according to the tree diagram in Figure 3.1.

2. **Friedman**: 10 continuous covariates are generated as $Z_1, \ldots, Z_{10} \overset{i.i.d.}{\sim} \text{Uniform}(0, 1)$,

and $\beta(\boldsymbol{z}_i) = [f(\boldsymbol{z}_i) - 14]/15$, where f (3.17) is the benchmark function proposed in Friedman [35]. We have scaled f to approximately have a mean of zero and standard deviation of one-third, thus restricting the majority of potential odds ratios to be between 0.5 and 2.

$$f(\mathbf{Z}) = 10\sin(\pi Z_1 Z_2) + 20(Z_3 - 0.5)^2 + 10Z_4 + 5Z_5 \tag{3.17}$$

As individuals are followed throughout the study period, cases are noted and the time-stratified case-crossover design is implemented. In both scenarios, approximately 4500 cases are typically observed.

For Scenario 1, we compare 1, 5, 10, 25, and 50 tree ensembles. For Scenario 2, due to the presence of many continuous predictors, we explore larger ensembles of 5, 10, 25, 50, and 100 trees. For both scenarios, we set $(k, \alpha_{\mathcal{T}}, \beta_{\mathcal{T}}) = (1, 0.95, 2)$ and run 10,000 total MCMC iterations, with the first 5,000 serving as a burn-in period. We keep every fifth post-burn-in sample, resulting in 1,000 posterior samples. Other hyperparameter settings are explored in the Chapter 3 Supplementary Materials.

To evaluate performance we fit an oracle conditional logistic regression by creating a design matrix consisting of the true interactions and/or functional forms of the moderators, each interacting with the exposure. For each simulation run we compute the average bias (3.18), root mean squared error (RMSE) (3.19), and average 95% posterior credible interval coverage (3.20) of the individual exposure effects.

$$\widehat{\text{Bias}}_\beta = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{\beta}(\boldsymbol{z}_i) - \beta(\boldsymbol{z}_i)\right] \tag{3.18}$$

$$\widehat{\text{RMSE}}_\beta = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\hat{\beta}(\boldsymbol{z}_i) - \beta(\boldsymbol{z}_i)\right]^2} \tag{3.19}$$

$$\widehat{\text{Coverage}}_\beta = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\beta(\boldsymbol{z}_i) \in \left[\hat{\beta}(\boldsymbol{z}_i)_{0.025}, \hat{\beta}(\boldsymbol{z}_i)_{0.975}\right]\right] \tag{3.20}$$

In (3.18), (3.19), and (3.20), $\hat{\beta}(\boldsymbol{z}_i)$, $\hat{\beta}(\boldsymbol{z}_i)_{0.025}$, $\hat{\beta}(\boldsymbol{z}_i)_{0.975}$ are the posterior mean, 2.5th percentile, and 97.5th percentile of the individual exposure effect, respectively, and $\mathbb{I}$ is an indicator function which takes value 1 when the true individual exposure effect is within the credible interval. Results are summarized over 200 simulations for each setting and are presented in Tables 3.1 and 3.2.

3.4.1 CART Simulation

Table 3.1: CART Simulation Results - BART Predictions

Type	T[a]	Bias[b]	RMSE[b]	Coverage[b]	Width[b]
Oracle		0.002 (0.001)	0.036 (0.001)	0.940 (0.017)	0.144 (0.000)
CL-BART	1	0.000 (0.001)	0.067 (0.001)	0.819 (0.027)	0.187 (0.003)
CL-BART	5	0.002 (0.001)	0.056 (0.001)	0.933 (0.018)	0.211 (0.002)
CL-BART	10	0.002 (0.001)	0.058 (0.001)	0.952 (0.015)	0.235 (0.002)
CL-BART	25	0.002 (0.001)	0.063 (0.001)	0.960 (0.014)	0.266 (0.001)
CL-BART	50	0.002 (0.001)	0.069 (0.001)	0.958 (0.014)	0.286 (0.001)

[a] T: Number of trees.

[b] Monte Carlo mean and standard errors across 200 simulations reported.

The oracle shows overall unbiasedness and near 95% coverage, confirming the validity of the case-crossover design setup (Table 3.1). CL-BART also has negligible bias, but generally has greater RMSE and wider intervals. The latter is to be expected since CL-BART estimates individual (not average) effects. RMSE is lowest for the 5 and 10 tree settings, and the average coverage generally increases as the number of trees is increased. Bias and coverage of the confounders is on par with the oracle (see Tables 3.6 and 3.7 in the Chapter 3 Supplementary Materials).

The WAIC is lower for the 5, 10, and 25 tree settings for the default tree regularization priors, suggesting the potential for using WAIC to select hyperparameters (see Figure 3.7 in

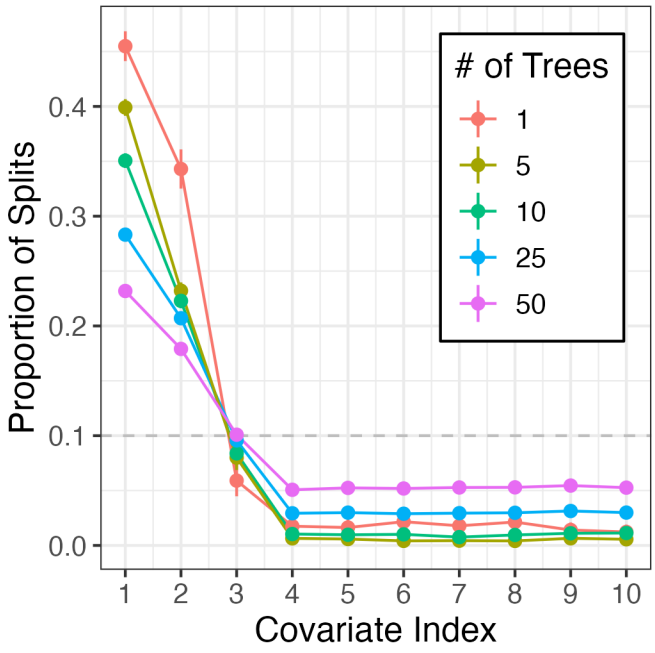

Figure 3.2: **CART Simulation Variable Importance**: Observed split proportions across 200 simulations (Monte Carlo mean and 95% uncertainty interval presented).

the Chapter 3 Supplementary Materials).

Across all simulations, the important covariates (Z_1', Z_2', Z_3') are typically split on with greater frequency than the remaining seven covariates (Figure 3.2). While these values are not perfect indicators of variable importance, this trend suggests the Dirichlet hyperprior is at least somewhat effective at selecting important covariates.

3.4.2 Friedman Simulation

For the Friedman scenario, the oracle achieves low bias and near 95% average coverage. CL-BART is unbiased even in small ensembles (Table 3.2). As more trees are added, RMSE and average coverage improve, but interval widths increase. Once again, this is likely due to CL-BART making predictions on the individual level. Estimates of the confounders exhibit low bias and good coverage, and the WAIC for this scenario suggests that larger ensembles perform better, but the improvements diminish as the number of trees approaches 100 (see Figure 3.10 in the Chapter 3 Supplementary Materials).

Table 3.2: Friedman Simulation Results - BART Predictions

Type	T[a]	Bias[b]	RMSE[b]	Coverage[b]	Width[b]
Oracle		-0.001 (0.001)	0.040 (0.001)	0.949 (0.016)	0.160 (0.000)
CL-BART	5	-0.001 (0.001)	0.165 (0.001)	0.801 (0.028)	0.431 (0.002)
CL-BART	10	-0.001 (0.001)	0.144 (0.001)	0.914 (0.020)	0.502 (0.002)
CL-BART	25	-0.001 (0.001)	0.130 (0.001)	0.967 (0.013)	0.568 (0.002)
CL-BART	50	-0.001 (0.001)	0.127 (0.001)	0.978 (0.010)	0.596 (0.003)
CL-BART	100	-0.001 (0.001)	0.126 (0.001)	0.980 (0.010)	0.600 (0.003)

[a] T: Number of trees.

[b] Monte Carlo mean and standard errors across 200 simulations reported.

We see that the important covariates $(Z_1, Z_2, Z_3, Z_4, Z_5)$ are all split on with greater frequencies, on average, than the remaining covariates (Figure 3.3). The Dirichlet hyperprior is particularly effective in this setting since there are many available splitting points for all covariates. Also, CL-BART does well to capture the true marginal partial dependence for each covariate (see Figure 3.11 in the Chapter 3 Supplementary Materials).

3.5 Application: Alzheimer's Disease and Heat Waves in California

3.5.1 Descriptive Statistics

There were 633,639 ED visits with an AD diagnosis reported during the study period. Patient sex was not reported for 62 cases, race was not reported for 7,662 cases, and ethnicity was not reported for 8,930 cases. Further, only 72,413 cases contained a heat wave per our definition within their referent window (most occurring in the summer months), and thus are the only cases which are informative for estimating heat wave effects. Dropping these cases and implementing the time-stratified case-crossover design resulted in a total of 71,020 cases

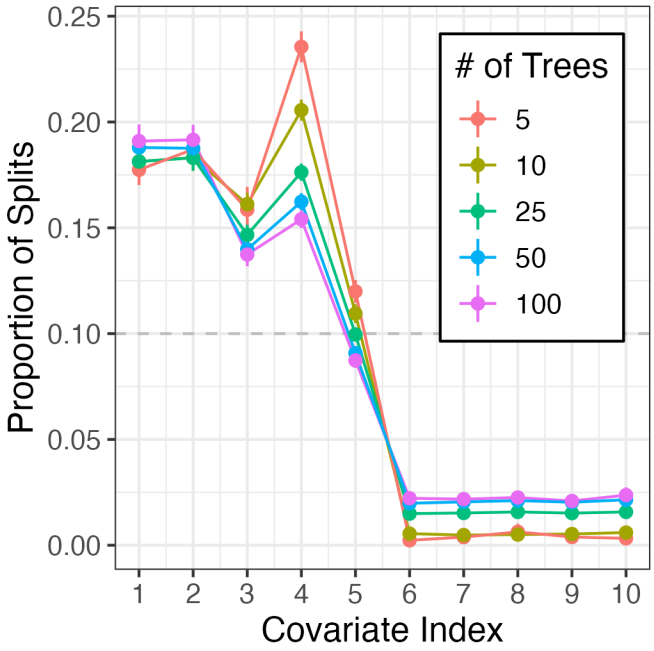

Figure 3.3: **Friedman Simulation Variable Importance**: Observed split proportions across 200 simulations (Monte Carlo mean and 95% uncertainty interval presented).

(319,336 observations).

The sample is primarily Non-Hispanic White (64.8%) and female (63.7%). The median age is 84 years (IQR: 79, 89). The median number of comorbid conditions is 2 (IQR: 1, 3), and hypertension is the most prevalent condition (65.2%) (Table 3.3). Over half of the sample has multiple conditions (56.1%), with the most common pairings being hypertension and hyperlipidemia (25.0%), hypertension and CKD (20.7%), hypertension and diabetes (19.6%), and hypertension and CHF (12.4%).

3.5.2 Model Considerations

Previous studies have found that associations between heat waves and health outcomes may differ by race and ethnicity [76, 60], so in addition to the overall analysis, we also conduct a stratified analysis with the following mutually exclusive subgroups: Hispanic, Non-Hispanic

Table 3.3: Descriptive Statistics for Emergency Department Visits Among Alzheimer's Disease Patients, CA 2005-2015

Characteristic	Overall
N[a]	71,020
Race/Ethnicity[b]	
Hispanic	11,959 (16.8%)
Non-Hispanic White	46,019 (64.8%)
Non-Hispanic Black	5,635 (7.9%)
Non-Hispanic Asian and Pacific Islander	5,521 (7.8%)
Non-Hispanic Other	1,886 (2.7%)
Sex[b]	
Male	25,762 (36.3%)
Female	45,258 (63.7%)
Age, yrs[c]	84 (79, 89)
Number of Comorbid Conditions[c]	2 (1, 3)
Congestive Heart Failure (CHF)[b]	11,494 (16.2%)
Chronic Kidney Disease (CKD)[b]	17,937 (25.3%)
Chronic Obstructive Pulmonary Disease (COPD)[b]	8,483 (11.9%)
Depression (DEP)[b]	9,005 (12.7%)
Diabetes (DIAB)[b]	17,654 (24.9%)
Hypertension (HT)[b]	46,281 (65.2%)
Hyperlipidemia (HLD)[b]	21,575 (30.4%)

[a] N; [b] N (%); [c] Median (IQR).

White, Non-Hispanic Black, Non-Hispanic Asian and Pacific Islander, and Non-Hispanic "other". The overall analysis includes these subgroups as potential effect moderators via one-hot encoding, while the stratified analysis is effectively forcing a split on race first. The overall analysis has the added benefit of having a larger sample size, but it may also mask heterogeneity within smaller subgroups, so we present both for comparison. In all analyses, we include sex and age alongside the comorbid conditions as potential moderators, with age being the only continuous moderator. The intuition behind including age is to allow it to serve as a proxy for other conditions that are not among those collected. The distribution of sex and age is similar across subgroups, but the prevalence of the comorbid conditions varies (e.g., hypertension and diabetes are less prevalent among the Non-Hispanic White subgroup) (see Table 3.12 in the Chapter 3 Supplementary Materials).

On the confounder side, both the daily average temperature and daily average dew-point temperature are modeled using natural cubic splines with four degrees of freedom. Federal holidays are included as a single indicator variable.

We fit a CL-BART model within each subgroup using the following hyperparameter settings: $T = 25$, $k = 1$, $\alpha_{\mathcal{T}} = 0.95$, and $\beta_{\mathcal{T}} = 2$. The WAIC was generally similar across different settings, so we only present the results for these particular values. For the overall model, we use the same settings except with $T = 100$, which had the lowest WAIC. We ran all models for 10,000 iterations, setting aside the first 5,000 as burn-in and only keeping every fifth sample, resulting in a total of 1,000 posterior samples. When fitting the CL-BART model, we monitor trace plots for σ_μ, $\bar{\beta}$, and the average number of nodes to ensure adequate mixing and convergence in the final model fits. Examples of these plots are included in Figure 3.15 in the Chapter 3 Supplementary Materials).

Table 3.4: Homogeneous vs. Average Heterogeneous Estimate for Heat Wave Effect

	CLR		CL-BART	
Subgroup	$\exp\left(\hat{\beta}\right)^a$ (95% CrI)	WAIC	$\exp\left(\bar{\beta}\right)^b$ (95% CrI)	WAIC
Overall	1.02 (0.99, 1.05)	216,788	1.01 (0.99, 1.05)	212,623
Hispanic	0.98 (0.91, 1.06)	36,170	0.99 (0.92, 1.05)	35,724
Non-Hispanic API	0.99 (0.89, 1.09)	16,512	0.99 (0.90, 1.08)	16,513
Non-Hispanic Black	1.09 (0.97, 1.21)	16,874	1.07 (0.96, 1.21)	16,868
Non-Hispanic Other	0.90 (0.73, 1.08)	5,658	0.92 (0.76, 1.09)	5,657
Non-Hispanic White	1.03 (0.99, 1.07)	137,904	1.01 (0.98, 1.05)	137,894

API: Asian and Pacific Islander.

CLR: Conditional Logistic Regression. CrI: Posterior Credible Interval.

[a] Estimated odds ratio from CLR with no effect moderators.

[b] Average exposure effect from CL-BART model.

3.5.3 Results

Both overall and within each subgroup, estimates of the average exposure effect $\bar{\beta}$ are similar to what one would obtain had they ignored effect heterogeneity entirely and simply fit a conditional logistic regression model as specified in (3.4) (Table 3.4). The WAIC is similar or better for the CL-BART model in all subgroups (Table 3.4), suggesting that the overall fit of the models are improved by considering effect heterogeneity, but the additional complexity introduced by using BART may limit the ability of the model to generalize to new data. Density plots of the posterior mean individual exposure effects illustrate the varying degrees of heterogeneity captured by CL-BART in each subgroup (see Figure 3.13 in the Chapter 3 Supplementary Materials).

To explore the heterogeneity estimated by CL-BART, we begin by visualizing the proportions of splits attributable to each moderator in Figure 3.4A. Unsurprisingly, age is split on with greater frequency than any of the binary moderators since it has more available splitting values. We also note that in some cases, certain binary covariates are split on more often

than others. Notably, CKD is more prominent for the Hispanic subgroup, and hypertension status is more prominent for Non-Hispanic Black subgroup. We have omitted the proportions for one-hot encoded race variables in the overall model, but together these accounted for 28% of splits, pointing toward the importance of race/ethnicity in the analysis.

Additionally, we present the marginal contributions as defined in (3.15) for each binary covariate in Figure 3.4B. These estimates are ratios of ORs, and thus represent the multiplicative effect associated with the given moderator on the underlying OR estimate for the association between ED visits and heat waves. In this way, they are similar to the interaction coefficient in a traditional regression model. For example, the presence of CKD among the Hispanic subgroup appears to be associated with a harmful modification of the exposure effect. Similarly, the presence of hypertension among the Non-Hispanic Black subgroup is associated with a protective modification of the exposure effect. While the harmful effect of CKD is most pronounced among the Hispanic subgroup, the estimated OR is greater than 1 across all subgroups, and the posterior credible interval is greater than 1 in the overall analysis. Other covariates have mixed effects on the heat wave effect across groups, but these are the most notable.

To examine interaction effects estimated via CL-BART, we fit CART models using the `rpart` R package [103] to obtain lower-dimensional summaries of the posterior mean individual exposure effects (see Figure 3.5 in the Chapter 3 Supplementary Materials). For these models, we drop the demographic moderators from the list of predictors to see how well the heterogeneity can be described by the comorbid conditions alone. We then compute $\bar{\beta}_{PD}$ for each leaf node represented in the summaries and plot the results in Figure 3.5. These plots are helpful in that they allow one to view the actual exposure effect, as opposed to just ratios of exposure effects in Figure 3.4B.

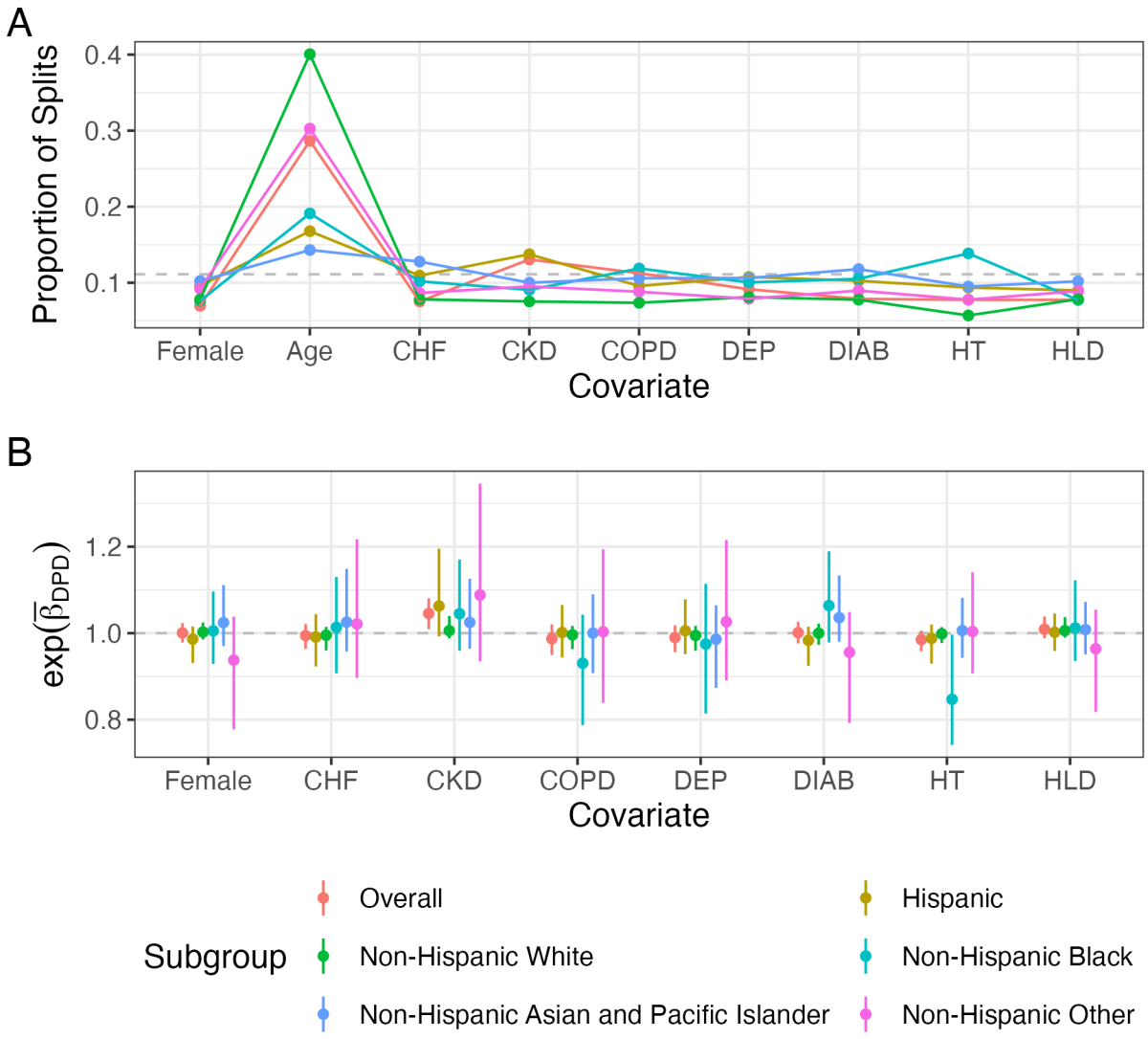

Figure 3.4: **Variable Importance and Marginal Partial Dependence for the Alzheimer's Disease Application**: Panel A displays the proportion of splits in the CL-BART model based on each covariate. Panel B displays the difference in partial average exposure effects for each binary covariate on the odds ratio scale (posterior means and 95% credible intervals presented). Numeric values corresponding to the estimates in panel B are provided in Tables 3.13 and 3.14 in the Chapter 3 Supplementary Materials. Abbreviations: DEP: depression, DIAB: diabetes, HT: hypertension, HLD: hyperlipidemia.

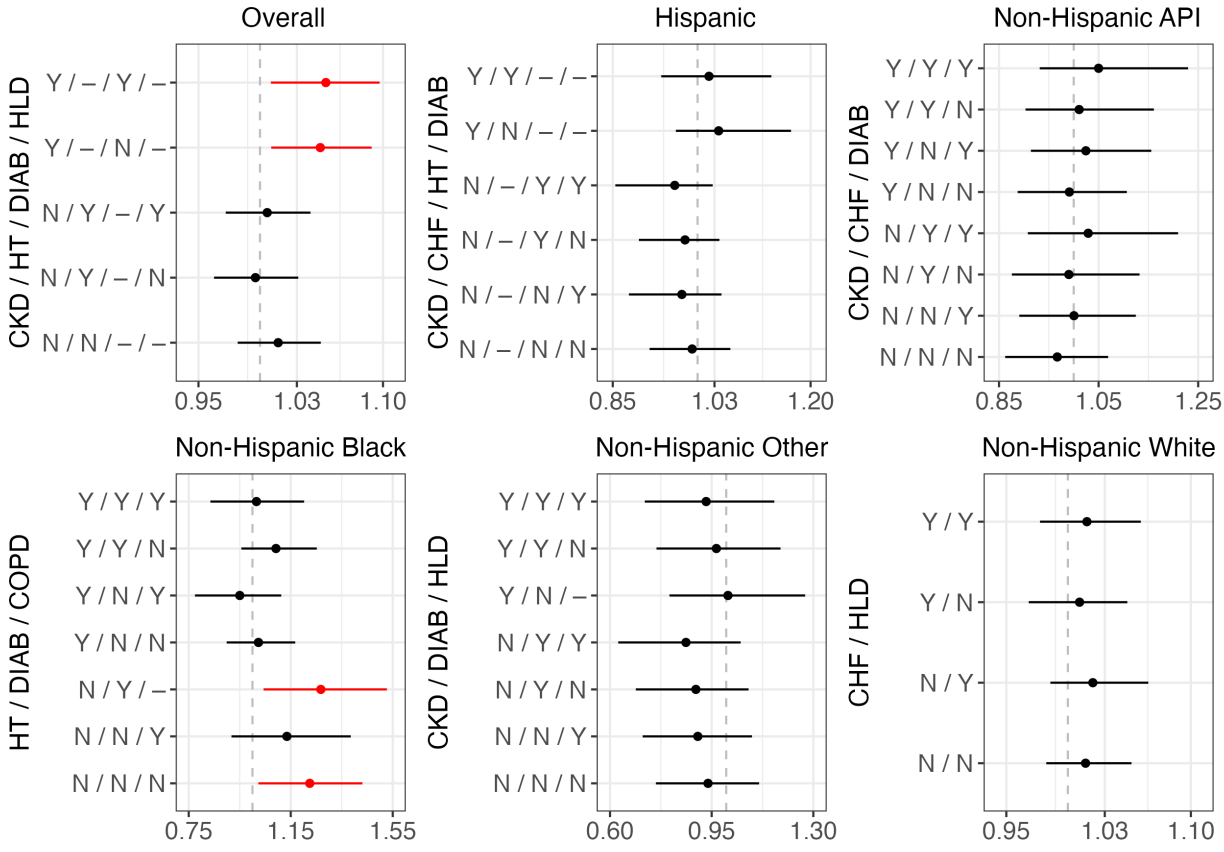

Figure 3.5: **CART-Informed Partial Average Heat Wave Effects for the Alzheimer's Disease Application**: Posterior mean and 95% credible intervals for the partial average heat wave effects within each leaf of the lower-dimensional CART summaries. Each condition is either present (Y), not present (N), or irrelevant (-). Results are presented on the odds ratio scale. Numeric values and CART diagrams are provided Table 3.15 and Figure 3.5 in the Chapter 3 Supplementary Materials. Abbreviations: DIAB: diabetes, HT: hypertension, HLD: hyperlipidemia.

In Figure 3.5, we observe that diabetes appears in 4 out of 5 subgroups, while CKD and CHF are the next most common moderators. Depression is the only condition that did not show up in any of the summaries. In the overall analysis, the estimated association between ED visits and heat waves is very strong for those with CKD, regardless of other important conditions. The subgroup with the most pronounced risks is Non-Hispanic Black. Among patients in this subgroup, the estimated association between ED visit and heat wave is greatest for those without hypertension, though an additional diagnosis of COPD may reduce this. Findings such as this illustrate the importance of considering interaction among moderators when modeling effect heterogeneity.

Unfortunately, the nice interpretations of the CART summaries come at a cost. The summary R^2 [116] for the CART summaries are relatively high for the Hispanic, Non-Hispanic Asian and Pacific Islander, and Non-Hispanic Black subgroups (0.89, 0.71, and 0.79, respectively), but suggest that much of the finer interactions estimated by CL-BART are not captured. The summary R^2 is very low for the Non-Hispanic Other (0.39) and Non-Hispanic White (0.09) subgroups. Age was split on with much greater frequency for these subgroups (Figure 3.4A), so in addition to struggling to summarize the CL-BART model fit, it is possible that age and/or sex are the drivers of effect heterogeneity in these subgroups. These findings should not dissuade one from studying effect heterogeneity, but they do illustrate the limitation of automating the analysis.

3.6 Discussion

CL-BART is a helpful tool for estimating heterogeneous effects in the case-crossover study design commonly used in environmental epidemiology. The primary benefit of CL-BART is

its ability to detect and estimate important high-order interactions and functional forms of potential effect moderators without requiring prespecification. Interpreting the heterogeneous effects can be challenging, but the proposed strategies revolving around variable importance, partial dependence, and lower-dimensional summaries provide a good start.

In terms of the application, the most consistent finding across most subgroups was that CKD and/or diabetes may modify the response to heat waves among people with AD. Specifically, having CKD was generally associated with an increased risk of ED visit during heat waves, which aligns with previous work that has established an association between kidney-related illness and extreme heat [56, 75, 47]. Another key finding was the protective-leaning effect of hypertension, particularly among the Non-Hispanic Black subgroup. This finding aligns well with previous studies of extreme heat both in California [94] and New York [68], and might be attributed to blood vessel dilation in hot weather, decreasing the risk of hypertension-associated morbidity [8]. We suspect that a mixture of biological and behavioral changes experienced by and medication(s) taken by those living with these comorbid conditions may be a contributing factor. Finally, while we used co-diagnosis codes at the ED visit to define comorbid conditions, other studies may consider other sources to ascertain pre-existing chronic conditions (e.g., medication use or medical history), to reduce classification error.

We acknowledge room for future improvements to CL-BART. The computation time is the largest limitation at this point in time. The bottleneck is the repeated evaluation of the conditional logistic regression likelihood required for Fisher scoring when determining the proposal distribution G of interim leaf node parameters and the adaptive rejection sampling of the final leaf node parameters. Exploring additional tree proposals such as those described in Pratola [90] and Deshpande [29] may improve mixing of the posterior chains, and thus

indirectly reduce computation time by lowering the number of MCMC iterations required to reach the stationary distribution. Average runtime for the simulation studies and application analyses are presented in the Chapter 3 Supplementary Materials.

Additionally, we have shown that the exposure may either be binary (application) or continuous (simulations). In both cases, a linear exposure-response relationship is assumed. There are many scenarios where this assumption may be violated. For example, we struggled to achieve model convergence in an exploratory analysis of the ED visit data using continuous daily average temperature as the exposure (results not shown). We suspect a nonlinear exposure-response relationship is at least partly responsible. Extending the CL-BART model to allow for polynomial, splines, and other flexible functions of the exposure could be desirable. For continuous exposure, it may also be interesting to consider short-term lagged effects via a distributed lag nonlinear model [38, 45]. Each basis function would require its own forest to be managed, substantially increasing the computational burden of an already burdensome algorithm.

In conclusion, CL-BART serves as a robust alternative to the typical strategies for estimating heterogeneous effects in the case-crossover design, using RJMCMC to integrate the flexibility of BART with traditional conditional logistic regression. This framework offers researchers a powerful tool to disentangle and model heterogeneous effects, whether it be in the context of treatment outcomes, environmental exposures, or any other one-to-many matched case-control study.

3.7 Supplementary Materials

3.7.1 CL-BART Algorithm Details

The Proposal Distribution G

In the reversible jump algorithm, the choice of the proposal distributions G_{GROW}, G_{PRUNE}, and G_{CHANGE} is crucial to ensuring good mixing. Linero [71] suggests a Normal proposal distribution based on the Laplace approximation. For any node η that is involved in a tree update proposal, we may sample $\mu_\eta \sim \text{Normal}(m_\eta, v_\eta^2)$, where

$$m_\eta = \arg\max_\mu \sum_{i:\boldsymbol{z}_i \mapsto \eta} \log p\left(\boldsymbol{y}_i \mid \lambda_i^{(t)} + \mu\right) + \log \pi_\mu(\mu)$$

and

$$v_\eta^{-2} = \sum_{i:\boldsymbol{z}_i \mapsto \eta} \mathcal{I}\left(\lambda_i^{(t)} + m_\eta\right) - \frac{d^2}{d\mu^2} \log \pi_\mu(\mu) \mid_{\mu = m_\eta}$$

Both m_η and v_η^{-2} may obtained using a Fisher scoring algorithm. Details for the gradient and Fisher information computation required in this step are provided in Section 3.7.1. For a starting point, we use the parameter from the leaf node that was split in either a GROW or CHANGE move, or a weighted average of the leaf node parameters that were deleted in a PRUNE move. We have found that this strategy typically works well, but in certain cases these starting values may be inadequate, causing the Fisher scoring algorithm to not converge. This usually occurs when there exists an extreme imbalance in the covariate space or the true predictions for leaf nodes are very different from that of their parent. In this rare scenario, we estimate m_η by first maximizing the likelihood using frequentist conditional logistic regression, followed by an application of the optimization technique introduced in

Brent [16] to incorporate prior information during the maximization.

Conditional Logistic Regression Derivations

Recall that in the case-crossover design, the conditional data likelihood is given by (3.4). Now suppose $\boldsymbol{\gamma}$ is known and we are interested only in the BART parameter β. Write the contribution to the conditional data likelihood for individual i as:

$$L_i^c(\beta) = p\left(\boldsymbol{y}_i \mid \beta\right) = \frac{\exp\left\{\boldsymbol{w}_{ij_i}^T\boldsymbol{\gamma} + \beta x_{ij_i}\right\}}{\sum_{j\in\mathcal{W}_i}\exp\left\{\boldsymbol{w}_{ij}^T\boldsymbol{\gamma} + \beta x_{ij}\right\}}. \tag{3.21}$$

The log-likelihood is given by:

$$l(\beta) = \boldsymbol{w}_{ij_i}^T\boldsymbol{\gamma} + \beta x_{ij_i} - \log\sum_{j\in\mathcal{W}_i}\exp\left\{\boldsymbol{w}_{ij}^T\boldsymbol{\gamma} + \beta x_{ij}\right\}. \tag{3.22}$$

The score with respect to β is given by:

$$U(\beta) = z_{ij_i} - \frac{\sum_{j\in\mathcal{W}_i} x_{ij}\exp\left\{\boldsymbol{w}_{ij}^T\boldsymbol{\gamma} + \beta x_{ij}\right\}}{\sum_{j\in\mathcal{W}_i}\exp\left\{\boldsymbol{w}_{ij}^T\boldsymbol{\gamma} + \beta x_{ij}\right\}} = \sum_{j\in\mathcal{W}_i} y_{ij}x_{ij} - \sum_{j\in\mathcal{W}_i} x_{ij}p_{ij}^c, \tag{3.23}$$

where p_{ij}^c is as defined in (3.3).

Let $\theta_{ij} = \boldsymbol{w}_{ij}^T\boldsymbol{\gamma} + \beta x_{ij}$. Then the Fisher information for β is given by:

$$\begin{aligned}\mathcal{I}(\beta) &= \frac{\left(\sum_{j\in\mathcal{W}_i}\exp\left(\theta_{ij}\right)\right)\left(\sum_{j\in\mathcal{W}_i} x_{ij}^2\exp\left(\theta_{ij}\right)\right) - \left(\sum_{j\in\mathcal{W}_i} x_{ij}\exp\left(\theta_{ij}\right)\right)^2}{\left(\sum_{j\in\mathcal{W}_i}\exp\left(\theta_{ij}\right)\right)^2} \\ &= \sum_{j\in\mathcal{W}_i} x_{ij}^2 p_{ij}^c - \left(\sum_{j\in\mathcal{W}_i} x_{ij}p_{ij}^c\right)^2.\end{aligned} \tag{3.24}$$

Note that in this scenario the Fisher information is equal to the observed information. These forms for $\mathcal{I}(\beta)$ and $U(\beta)$ are used for the Fisher scoring algorithm mentioned in Section 3.7.1.

Notes on Posterior Contraction

Linero [71] provides a theoretical base for the posterior contraction of the RJMCMC BART methodology for several models classes under certain conditions on the function of interest $\beta(\cdot)$ and the BART prior. One such class is those models belonging to an exponential family. Here, we show that CL-BART fits into this class of models through its connection to the multinomial likelihood. In showing this, we suggest that CL-BART has similar contraction properties outlined in the source reference.

Recall that the conditional logistic regression likelihood for a stratum (individual) in a case-crossover analysis corresponds to the joint probability of observing the case at a single time point. We can rewrite the contribution of single individual to the total conditional likelihood provided in (3.4) as follows:

$$\begin{aligned} L_i^c(\beta) &= \prod_{j\in\mathcal{W}_i}\left(\frac{\exp\left\{\boldsymbol{w}_{ij}^T\boldsymbol{\gamma}+\beta x_{ij}\right\}}{\sum_{j\in\mathcal{W}_i}\exp\left\{\boldsymbol{w}_{ij}^T\boldsymbol{\gamma}+\beta x_{ij}\right\}}\right)^{y_{ij}} \\ &= \exp\left(\sum_{j\in\mathcal{W}_i}\left[\left(y_{ij}\boldsymbol{w}_{ij}^T\boldsymbol{\gamma}+\beta x_{ij}\right)-y_{ij}\log\sum_{j\in\mathcal{W}_i}\exp\left(\boldsymbol{w}_{ij}^T\boldsymbol{\gamma}+\beta x_{ij}\right)\right]\right) \\ &= \exp\left(\sum_{j\in\mathcal{W}_i}y_{ij}(\boldsymbol{w}_{ij}^T\boldsymbol{\gamma}+\beta x_{ij})-\log\sum_{j\in\mathcal{W}_i}\exp\left(\boldsymbol{w}_{ij}^T\boldsymbol{\gamma}+\beta x_{ij}\right)\right) \\ &= \exp\left(\beta\sum_{j\in\mathcal{W}_i}y_{ij}x_{ij}-\log\sum_{j\in\mathcal{W}_i}\exp\left(\boldsymbol{w}_{ij}^T\boldsymbol{\gamma}+\beta x_{ij}\right)+\sum_{j\in\mathcal{W}_i}y_{ij}\boldsymbol{w}_{ij}^T\boldsymbol{\gamma}\right). \end{aligned} \tag{3.25}$$

Equation (3.25) has the form of an 1-dimensional exponential family with sufficient statistic $\sum_{j\in\mathcal{W}_i}y_{ij}x_{ij}$, natural parameter β, and $b(\beta)=\log\sum_{j\in\mathcal{W}_i}\exp\left(\boldsymbol{w}_{ij}^T\boldsymbol{\gamma}+\beta x_{ij}\right)$. We can verify some properties of the exponential form of the conditional data likelihood. The conditional

expectation is given by:

$$\mathbb{E}\left[\sum_{j\in\mathcal{W}_i} Y_{ij}x_{ij} \mid \sum_{j\in\mathcal{W}_i} Y_{ij} = 1\right] = b'(\beta) = \frac{\sum_{j\in\mathcal{W}_i} x_{ij}\exp\left\{\boldsymbol{w}_{ij}^T\boldsymbol{\gamma} + \beta x_{ij}\right\}}{\sum_{j\in\mathcal{W}_i}\exp\left\{\boldsymbol{w}_{ij}^T\boldsymbol{\gamma} + \beta x_{ij}\right\}} = \sum_{j\in\mathcal{W}_i} x_{ij}p_{ij}^c,$$

and the conditional variance is given by:

$$\begin{aligned}
\mathrm{Var}^{\,c}\left[\sum_{j\in\mathcal{W}_i} Y_{ij}x_{ij}\right] &= \mathrm{Var}\left[\sum_{j\in\mathcal{W}_i} Y_{ij}x_{ij} \mid \sum_{j\in\mathcal{W}_i} Y_{ij} = 1\right] \\
&= b''(\beta) \\
&= \frac{\left(\sum_{j\in\mathcal{W}_i}\exp\left(\theta_{ij}\right)\right)\left(\sum_{j\in\mathcal{W}_i} x_{ij}^2\exp\left(\theta_{ij}\right)\right) - \left(\sum_{j\in\mathcal{W}_i} x_{ij}\exp\left(\theta_{ij}\right)\right)^2}{\left(\sum_{j\in\mathcal{W}_i}\exp\left(\theta_{ij}\right)\right)^2} \\
&= \sum_{j\in\mathcal{W}_i} x_{ij}^2 p_{ij}^c - \left(\sum_{j\in\mathcal{W}_i} x_{ij}p_{ij}^c\right)^2,
\end{aligned}$$

where $\theta_{ij} = \boldsymbol{w}_{ij}^T\boldsymbol{\gamma} + \beta x_{ij}$ and p_{ij}^c is as defined in (3.3). Note the conditional expectation and variance match the derivations in Section 3.7.1.

Suitable bounds for the Kullback-Leibler divergence from $L_i^c(\beta)$ to $L_i^c(\beta+\Delta)$, such as those derived in the supplemental material of Linero [71], may be used to satisfy the model condition outlined therein.

3.7.2 Additional Simulation Materials

Table 3.5: Extended CART Simulation Results - BART Predictions

Type	T[a]	k[b]	$(\alpha_{\mathcal{T}}, \beta_{\mathcal{T}})$[c]	Bias[d]	RMSE[d]	Coverage[d]	Width[d]
Oracle				0.002 (0.001)	0.036 (0.001)	0.940 (0.017)	0.144 (0.000)
CL-BART	1	0.1	(0.5, 3)	-0.001 (0.001)	0.070 (0.002)	0.766 (0.030)	0.170 (0.002)
CL-BART	1	0.1	(0.95, 2)	-0.001 (0.001)	0.069 (0.001)	0.813 (0.028)	0.189 (0.003)
CL-BART	1	0.5	(0.5, 3)	0.000 (0.001)	0.069 (0.002)	0.771 (0.030)	0.168 (0.002)
CL-BART	1	0.5	(0.95, 2)	0.000 (0.001)	0.069 (0.001)	0.807 (0.028)	0.189 (0.003)
CL-BART	1	1.0	(0.5, 3)	0.000 (0.001)	0.068 (0.001)	0.762 (0.030)	0.164 (0.002)
CL-BART	1	1.0	(0.95, 2)	0.000 (0.001)	0.067 (0.001)	0.819 (0.027)	0.187 (0.003)
CL-BART	5	0.1	(0.5, 3)	0.002 (0.001)	0.058 (0.001)	0.862 (0.024)	0.174 (0.002)
CL-BART	5	0.1	(0.95, 2)	0.001 (0.001)	0.055 (0.001)	0.938 (0.017)	0.214 (0.002)
CL-BART	5	0.5	(0.5, 3)	0.002 (0.001)	0.059 (0.001)	0.856 (0.025)	0.174 (0.002)
CL-BART	5	0.5	(0.95, 2)	0.001 (0.001)	0.056 (0.001)	0.937 (0.017)	0.213 (0.002)
CL-BART	5	1.0	(0.5, 3)	0.002 (0.001)	0.059 (0.001)	0.856 (0.025)	0.173 (0.002)
CL-BART	5	1.0	(0.95, 2)	0.002 (0.001)	0.056 (0.001)	0.933 (0.018)	0.211 (0.002)
CL-BART	10	0.1	(0.5, 3)	0.002 (0.001)	0.059 (0.001)	0.882 (0.023)	0.185 (0.002)
CL-BART	10	0.1	(0.95, 2)	0.001 (0.001)	0.058 (0.001)	0.950 (0.015)	0.235 (0.002)
CL-BART	10	0.5	(0.5, 3)	0.002 (0.001)	0.059 (0.001)	0.883 (0.023)	0.184 (0.002)
CL-BART	10	0.5	(0.95, 2)	0.002 (0.001)	0.058 (0.001)	0.952 (0.015)	0.236 (0.002)
CL-BART	10	1.0	(0.5, 3)	0.002 (0.001)	0.059 (0.001)	0.879 (0.023)	0.183 (0.002)
CL-BART	10	1.0	(0.95, 2)	0.002 (0.001)	0.058 (0.001)	0.952 (0.015)	0.235 (0.002)
CL-BART	25	0.1	(0.5, 3)	0.002 (0.001)	0.060 (0.001)	0.900 (0.021)	0.199 (0.002)
CL-BART	25	0.1	(0.95, 2)	0.002 (0.001)	0.063 (0.001)	0.955 (0.015)	0.261 (0.001)
CL-BART	25	0.5	(0.5, 3)	0.002 (0.001)	0.059 (0.001)	0.902 (0.021)	0.200 (0.002)
CL-BART	25	0.5	(0.95, 2)	0.002 (0.001)	0.064 (0.001)	0.959 (0.014)	0.266 (0.001)
CL-BART	25	1.0	(0.5, 3)	0.002 (0.001)	0.059 (0.001)	0.907 (0.021)	0.199 (0.002)
CL-BART	25	1.0	(0.95, 2)	0.002 (0.001)	0.063 (0.001)	0.960 (0.014)	0.266 (0.001)
CL-BART	50	0.1	(0.5, 3)	0.002 (0.001)	0.061 (0.001)	0.908 (0.020)	0.208 (0.001)
CL-BART	50	0.1	(0.95, 2)	0.002 (0.001)	0.069 (0.001)	0.953 (0.015)	0.279 (0.001)
CL-BART	50	0.5	(0.5, 3)	0.002 (0.001)	0.061 (0.001)	0.912 (0.020)	0.209 (0.001)
CL-BART	50	0.5	(0.95, 2)	0.002 (0.001)	0.069 (0.001)	0.957 (0.014)	0.284 (0.001)
CL-BART	50	1.0	(0.5, 3)	0.002 (0.001)	0.060 (0.001)	0.916 (0.020)	0.210 (0.001)
CL-BART	50	1.0	(0.95, 2)	0.002 (0.001)	0.069 (0.001)	0.958 (0.014)	0.286 (0.001)

[a] T: Number of trees.

[b] k: Numerator of scale parameter for half-Cauchy hyper-prior.

[c] $(\alpha_{\mathcal{T}}, \beta_{\mathcal{T}})$: Hyperparameters for tree depth prior.

[d] Monte Carlo mean and standard errors across 200 simulations reported.

Table 3.6: Extended CART Simulation Results - Confounder Estimates (Bias)

				Bias × 1,000[d]				
Type	T[a]	k[b]	$(\alpha_{\mathcal{T}}, \beta_{\mathcal{T}})$[c]	γ_1	γ_2	γ_3	γ_4	γ_5
Oracle				6.54 (4.49)	-11.05 (4.77)	1.59 (5.10)	-3.90 (4.65)	-0.46 (4.69)
CL-BART	1	0.1	(0.5, 3)	5.60 (4.51)	-11.30 (4.78)	1.56 (5.13)	-3.85 (4.66)	-0.06 (4.65)
CL-BART	1	0.1	(0.95, 2)	5.68 (4.49)	-11.56 (4.80)	2.08 (5.10)	-4.17 (4.66)	0.16 (4.66)
CL-BART	1	0.5	(0.5, 3)	5.96 (4.50)	-11.90 (4.79)	1.55 (5.09)	-4.39 (4.67)	0.31 (4.70)
CL-BART	1	0.5	(0.95, 2)	5.53 (4.48)	-11.76 (4.75)	1.56 (5.10)	-3.80 (4.66)	0.22 (4.69)
CL-BART	1	1.0	(0.5, 3)	6.56 (4.48)	-11.50 (4.77)	1.46 (5.11)	-4.22 (4.65)	-0.06 (4.64)
CL-BART	1	1.0	(0.95, 2)	6.06 (4.51)	-11.43 (4.80)	1.30 (5.09)	-3.48 (4.66)	0.07 (4.63)
CL-BART	5	0.1	(0.5, 3)	5.63 (4.49)	-10.93 (4.79)	1.64 (5.10)	-4.34 (4.65)	-0.38 (4.71)
CL-BART	5	0.1	(0.95, 2)	5.90 (4.50)	-11.43 (4.76)	1.77 (5.11)	-4.02 (4.65)	0.20 (4.69)
CL-BART	5	0.5	(0.5, 3)	6.17 (4.48)	-11.10 (4.81)	1.59 (5.11)	-3.98 (4.65)	0.02 (4.70)
CL-BART	5	0.5	(0.95, 2)	5.43 (4.50)	-11.48 (4.76)	1.42 (5.15)	-3.74 (4.65)	0.76 (4.69)
CL-BART	5	1.0	(0.5, 3)	6.24 (4.50)	-11.31 (4.77)	1.55 (5.10)	-3.71 (4.67)	0.30 (4.69)
CL-BART	5	1.0	(0.95, 2)	5.86 (4.50)	-11.22 (4.79)	1.81 (5.11)	-3.64 (4.67)	0.02 (4.68)
CL-BART	10	0.1	(0.5, 3)	6.11 (4.46)	-11.37 (4.78)	0.80 (5.11)	-4.06 (4.69)	0.25 (4.67)
CL-BART	10	0.1	(0.95, 2)	5.95 (4.51)	-11.78 (4.80)	1.62 (5.13)	-4.00 (4.63)	0.69 (4.71)
CL-BART	10	0.5	(0.5, 3)	6.28 (4.49)	-11.31 (4.80)	1.78 (5.11)	-4.04 (4.65)	0.53 (4.71)
CL-BART	10	0.5	(0.95, 2)	5.44 (4.49)	-11.57 (4.79)	0.75 (5.13)	-4.04 (4.66)	0.74 (4.71)
CL-BART	10	1.0	(0.5, 3)	6.16 (4.49)	-11.39 (4.79)	1.35 (5.14)	-4.21 (4.64)	0.21 (4.67)
CL-BART	10	1.0	(0.95, 2)	5.54 (4.49)	-11.01 (4.82)	1.86 (5.08)	-3.93 (4.69)	0.96 (4.69)
CL-BART	25	0.1	(0.5, 3)	5.78 (4.51)	-11.56 (4.79)	1.64 (5.11)	-3.41 (4.64)	0.21 (4.71)
CL-BART	25	0.1	(0.95, 2)	5.80 (4.50)	-10.96 (4.81)	1.66 (5.13)	-3.94 (4.68)	0.84 (4.72)
CL-BART	25	0.5	(0.5, 3)	6.25 (4.51)	-11.64 (4.78)	1.59 (5.13)	-3.80 (4.68)	0.23 (4.71)
CL-BART	25	0.5	(0.95, 2)	5.19 (4.53)	-11.45 (4.78)	1.54 (5.10)	-3.87 (4.66)	0.44 (4.73)
CL-BART	25	1.0	(0.5, 3)	5.71 (4.49)	-11.49 (4.80)	1.72 (5.14)	-3.69 (4.67)	0.34 (4.70)
CL-BART	25	1.0	(0.95, 2)	5.65 (4.51)	-11.34 (4.79)	1.97 (5.13)	-3.98 (4.63)	0.86 (4.72)
CL-BART	50	0.1	(0.5, 3)	5.58 (4.50)	-11.22 (4.80)	1.60 (5.14)	-3.79 (4.66)	0.21 (4.71)
CL-BART	50	0.1	(0.95, 2)	5.64 (4.56)	-11.51 (4.78)	1.70 (5.11)	-3.64 (4.68)	1.06 (4.68)
CL-BART	50	0.5	(0.5, 3)	5.86 (4.49)	-11.21 (4.80)	1.01 (5.09)	-3.96 (4.68)	-0.27 (4.74)
CL-BART	50	0.5	(0.95, 2)	5.97 (4.53)	-11.80 (4.77)	1.47 (5.11)	-3.61 (4.68)	1.16 (4.68)
CL-BART	50	1.0	(0.5, 3)	5.87 (4.55)	-11.56 (4.82)	1.30 (5.11)	-3.64 (4.64)	0.40 (4.69)
CL-BART	50	1.0	(0.95, 2)	5.33 (4.50)	-11.11 (4.78)	1.50 (5.13)	-3.90 (4.67)	0.21 (4.74)

[a] T: Number of trees.

[b] k: Numerator of scale parameter for half-Cauchy hyper-prior.

[c] $(\alpha_{\mathcal{T}}, \beta_{\mathcal{T}})$: Hyperparameters for tree depth prior.

[d] Monte Carlo mean and standard errors across 200 simulations reported.

Table 3.7: Extended CART Simulation Results - Confounder Estimates (Coverage)

				Coverage[d]				
Type	T[a]	k[b]	$(\alpha_{\mathcal{T}}, \beta_{\mathcal{T}})$[c]	γ_1	γ_2	γ_3	γ_4	γ_5
Oracle				0.95	0.93	0.92	0.95	0.95
CL-BART	1	0.1	(0.5, 3)	0.94	0.93	0.92	0.95	0.94
CL-BART	1	0.1	(0.95, 2)	0.96	0.93	0.92	0.93	0.95
CL-BART	1	0.5	(0.5, 3)	0.94	0.92	0.91	0.94	0.93
CL-BART	1	0.5	(0.95, 2)	0.95	0.94	0.92	0.94	0.95
CL-BART	1	1.0	(0.5, 3)	0.96	0.93	0.91	0.94	0.94
CL-BART	1	1.0	(0.95, 2)	0.93	0.92	0.92	0.94	0.94
CL-BART	5	0.1	(0.5, 3)	0.95	0.93	0.92	0.94	0.94
CL-BART	5	0.1	(0.95, 2)	0.95	0.94	0.92	0.95	0.96
CL-BART	5	0.5	(0.5, 3)	0.95	0.92	0.92	0.95	0.94
CL-BART	5	0.5	(0.95, 2)	0.94	0.93	0.92	0.94	0.94
CL-BART	5	1.0	(0.5, 3)	0.94	0.94	0.92	0.95	0.92
CL-BART	5	1.0	(0.95, 2)	0.94	0.94	0.92	0.94	0.93
CL-BART	10	0.1	(0.5, 3)	0.95	0.92	0.90	0.94	0.95
CL-BART	10	0.1	(0.95, 2)	0.95	0.92	0.90	0.95	0.94
CL-BART	10	0.5	(0.5, 3)	0.93	0.93	0.92	0.93	0.95
CL-BART	10	0.5	(0.95, 2)	0.94	0.93	0.92	0.94	0.94
CL-BART	10	1.0	(0.5, 3)	0.95	0.92	0.91	0.94	0.95
CL-BART	10	1.0	(0.95, 2)	0.95	0.93	0.91	0.94	0.95
CL-BART	25	0.1	(0.5, 3)	0.95	0.94	0.92	0.94	0.95
CL-BART	25	0.1	(0.95, 2)	0.95	0.94	0.91	0.93	0.94
CL-BART	25	0.5	(0.5, 3)	0.94	0.93	0.92	0.93	0.95
CL-BART	25	0.5	(0.95, 2)	0.95	0.92	0.92	0.94	0.95
CL-BART	25	1.0	(0.5, 3)	0.95	0.94	0.92	0.95	0.94
CL-BART	25	1.0	(0.95, 2)	0.94	0.93	0.90	0.94	0.94
CL-BART	50	0.1	(0.5, 3)	0.94	0.92	0.90	0.94	0.95
CL-BART	50	0.1	(0.95, 2)	0.94	0.93	0.92	0.94	0.96
CL-BART	50	0.5	(0.5, 3)	0.95	0.93	0.90	0.94	0.94
CL-BART	50	0.5	(0.95, 2)	0.95	0.93	0.91	0.93	0.94
CL-BART	50	1.0	(0.5, 3)	0.93	0.94	0.92	0.95	0.94
CL-BART	50	1.0	(0.95, 2)	0.95	0.93	0.92	0.94	0.93

[a] T: Number of trees.

[b] k: Numerator of scale parameter for half-Cauchy hyper-prior.

[c] $(\alpha_{\mathcal{T}}, \beta_{\mathcal{T}})$: Hyperparameters for tree depth prior.

[d] 95% credible interval coverage across 200 simulations.

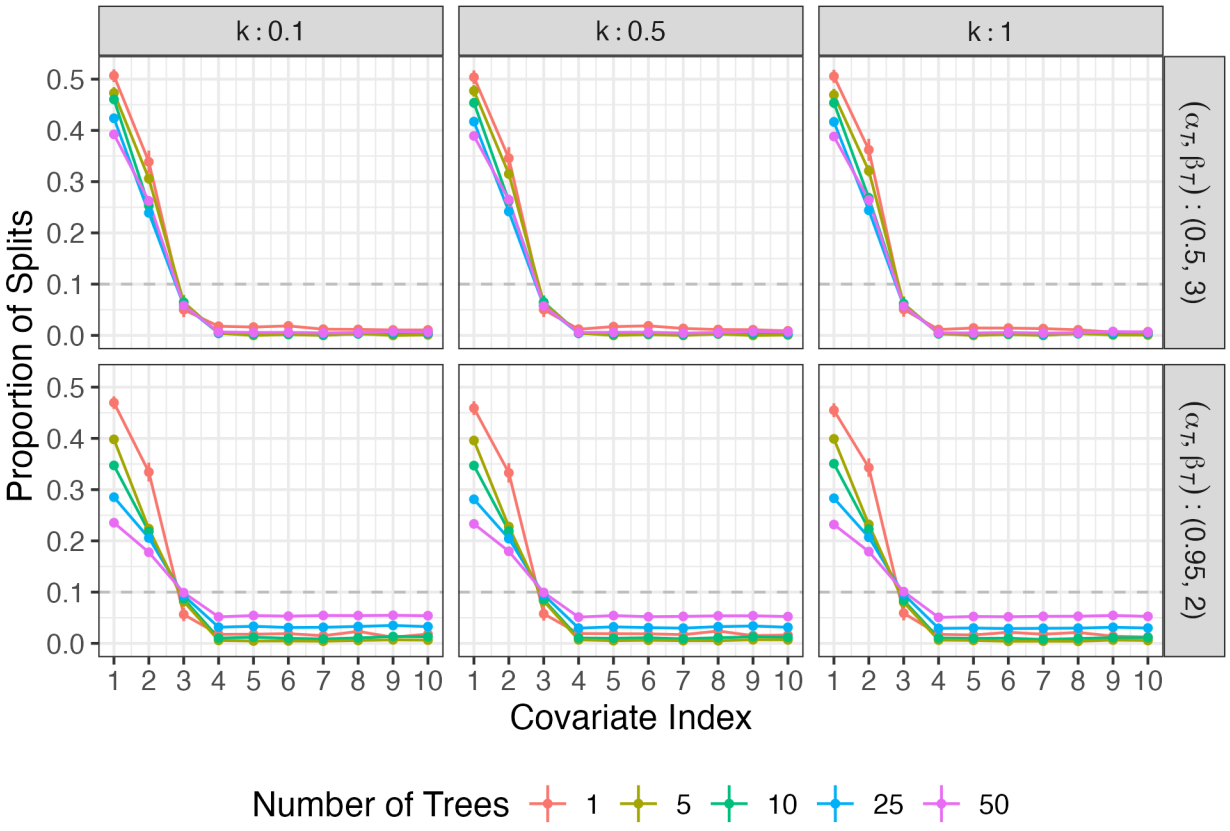

Figure 3.6: **Extended CART Simulation Variable Importance**: Plot of observed split proportions across 200 simulations (Monte Carlo mean and 95% uncertainty interval presented).

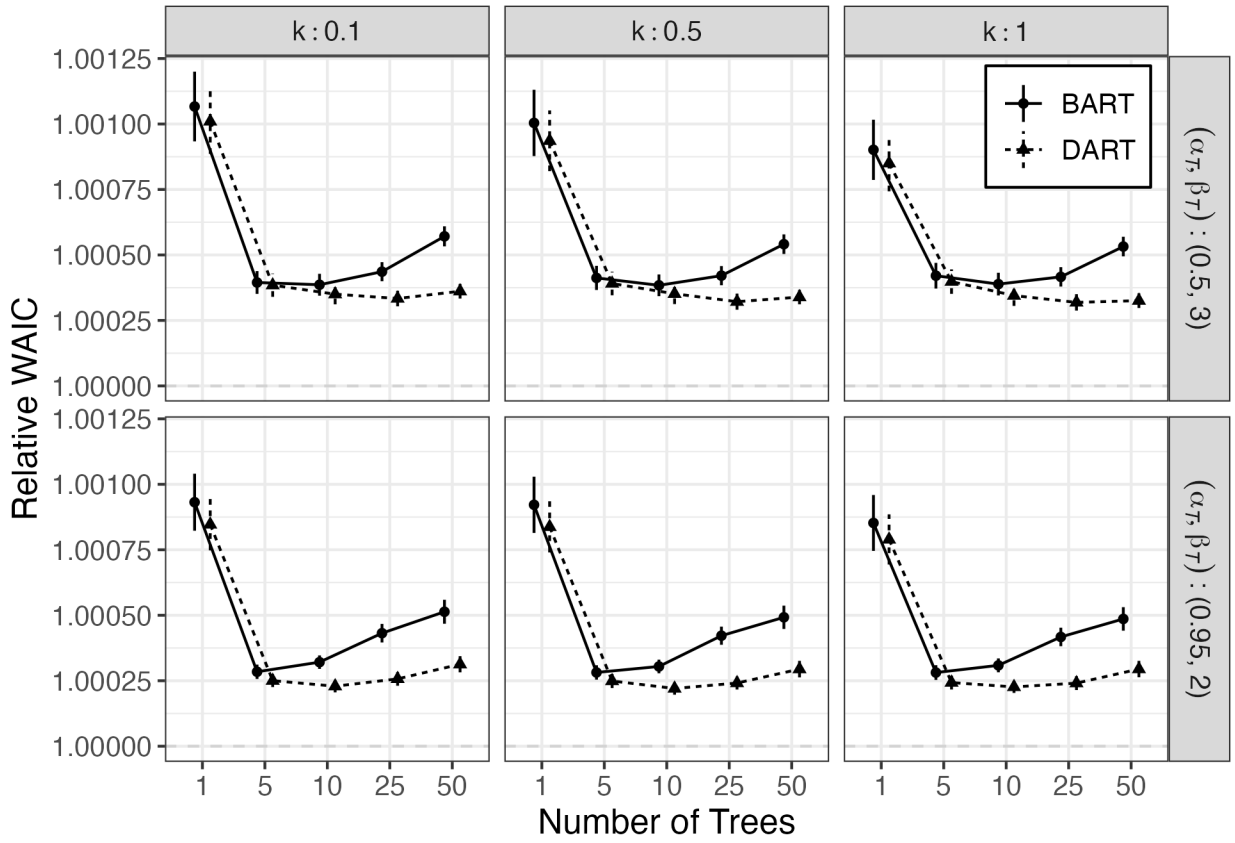

Figure 3.7: **WAIC for CART Simulation**: Relative WAIC for each simulation setting across 200 simulations (Monte Carlo mean and 95% uncertainty intervals presented). BART included for reference to demonstrate the effect of the Dirichlet prior on covariate selection.

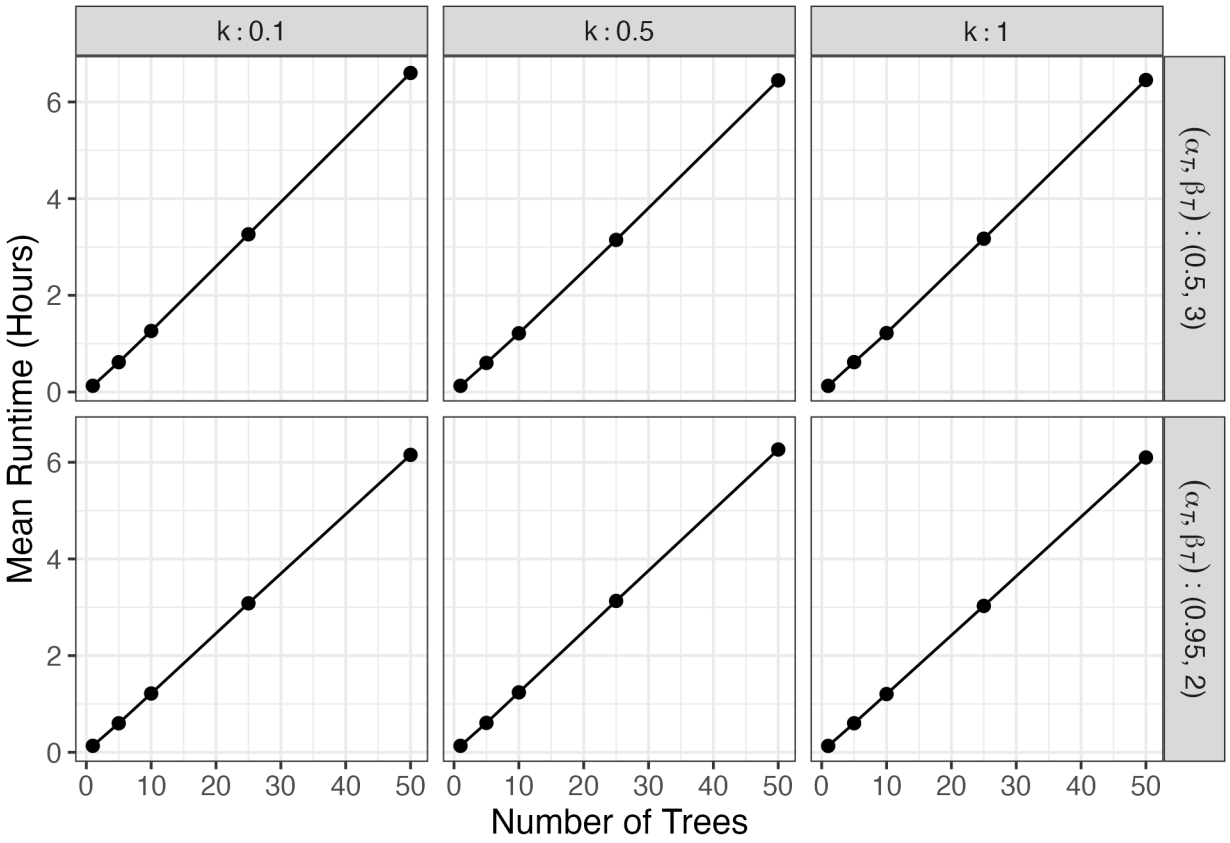

Figure 3.8: **CART Simulation Runtime**: Average time taken to run one CL-BART chain across 200 simulations. All models were run using 1 CPU on the high performance computing cluster at the Rollins School of Public Health, Emory University.

Table 3.8: Extended Friedman Simulation Results - BART Predictions

Type	T[a]	k[b]	$(\alpha_{\mathcal{T}}, \beta_{\mathcal{T}})$[c]	Bias[d]	RMSE[d]	Coverage[d]	Width[d]
Oracle				-0.001 (0.001)	0.040 (0.001)	0.949 (0.016)	0.160 (0.000)
CL-BART	5	0.1	(0.5, 3)	-0.001 (0.001)	0.169 (0.001)	0.740 (0.031)	0.387 (0.003)
CL-BART	5	0.1	(0.95, 2)	-0.001 (0.001)	0.164 (0.001)	0.804 (0.028)	0.430 (0.003)
CL-BART	5	0.5	(0.5, 3)	-0.001 (0.001)	0.170 (0.001)	0.740 (0.031)	0.390 (0.003)
CL-BART	5	0.5	(0.95, 2)	-0.001 (0.001)	0.164 (0.001)	0.807 (0.028)	0.434 (0.003)
CL-BART	5	1.0	(0.5, 3)	-0.001 (0.001)	0.171 (0.001)	0.737 (0.031)	0.388 (0.003)
CL-BART	5	1.0	(0.95, 2)	-0.001 (0.001)	0.165 (0.001)	0.801 (0.028)	0.431 (0.002)
CL-BART	10	0.1	(0.5, 3)	-0.001 (0.001)	0.151 (0.001)	0.846 (0.026)	0.435 (0.002)
CL-BART	10	0.1	(0.95, 2)	-0.001 (0.001)	0.142 (0.001)	0.914 (0.020)	0.498 (0.002)
CL-BART	10	0.5	(0.5, 3)	-0.001 (0.001)	0.151 (0.001)	0.845 (0.026)	0.437 (0.002)
CL-BART	10	0.5	(0.95, 2)	-0.001 (0.001)	0.143 (0.001)	0.915 (0.020)	0.502 (0.002)
CL-BART	10	1.0	(0.5, 3)	-0.001 (0.001)	0.152 (0.001)	0.842 (0.026)	0.436 (0.002)
CL-BART	10	1.0	(0.95, 2)	-0.001 (0.001)	0.144 (0.001)	0.914 (0.020)	0.502 (0.002)
CL-BART	25	0.1	(0.5, 3)	-0.001 (0.001)	0.138 (0.001)	0.913 (0.020)	0.477 (0.002)
CL-BART	25	0.1	(0.95, 2)	-0.001 (0.001)	0.129 (0.001)	0.967 (0.013)	0.560 (0.002)
CL-BART	25	0.5	(0.5, 3)	-0.001 (0.001)	0.138 (0.001)	0.915 (0.020)	0.479 (0.002)
CL-BART	25	0.5	(0.95, 2)	-0.001 (0.001)	0.130 (0.001)	0.967 (0.013)	0.564 (0.002)
CL-BART	25	1.0	(0.5, 3)	-0.001 (0.001)	0.138 (0.001)	0.914 (0.020)	0.480 (0.002)
CL-BART	25	1.0	(0.95, 2)	-0.001 (0.001)	0.130 (0.001)	0.967 (0.013)	0.568 (0.002)
CL-BART	50	0.1	(0.5, 3)	-0.001 (0.001)	0.132 (0.001)	0.942 (0.017)	0.503 (0.002)
CL-BART	50	0.1	(0.95, 2)	-0.001 (0.001)	0.126 (0.001)	0.977 (0.011)	0.583 (0.003)
CL-BART	50	0.5	(0.5, 3)	-0.001 (0.001)	0.132 (0.001)	0.943 (0.016)	0.506 (0.002)
CL-BART	50	0.5	(0.95, 2)	-0.001 (0.001)	0.126 (0.001)	0.978 (0.010)	0.590 (0.003)
CL-BART	50	1.0	(0.5, 3)	-0.001 (0.001)	0.133 (0.001)	0.943 (0.016)	0.509 (0.002)
CL-BART	50	1.0	(0.95, 2)	-0.001 (0.001)	0.127 (0.001)	0.978 (0.010)	0.596 (0.003)
CL-BART	100	0.1	(0.5, 3)	-0.001 (0.001)	0.128 (0.001)	0.955 (0.015)	0.520 (0.002)
CL-BART	100	0.1	(0.95, 2)	-0.001 (0.001)	0.126 (0.001)	0.979 (0.010)	0.591 (0.003)
CL-BART	100	0.5	(0.5, 3)	-0.001 (0.001)	0.129 (0.001)	0.956 (0.014)	0.523 (0.002)
CL-BART	100	0.5	(0.95, 2)	-0.001 (0.001)	0.126 (0.001)	0.980 (0.010)	0.596 (0.003)
CL-BART	100	1.0	(0.5, 3)	-0.001 (0.001)	0.129 (0.001)	0.958 (0.014)	0.529 (0.002)
CL-BART	100	1.0	(0.95, 2)	-0.001 (0.001)	0.126 (0.001)	0.980 (0.010)	0.600 (0.003)

[a] T: Number of trees.

[b] k: Numerator of scale parameter for half-Cauchy hyper-prior.

[c] $(\alpha_{\mathcal{T}}, \beta_{\mathcal{T}})$: Hyperparameters for tree depth prior.

[d] Monte Carlo mean and standard errors across 200 simulations reported.

Table 3.9: Extended Friedman Simulation Results - Confounder Estimates (Bias)

				Bias × 1,000[d]				
Type	T[a]	k[b]	$(\alpha_{\mathcal{T}}, \beta_{\mathcal{T}})$[c]	γ_1	γ_2	γ_3	γ_4	γ_5
Oracle				2.28 (4.36)	-0.76 (4.37)	-4.09 (4.70)	0.58 (4.57)	4.06 (4.58)
CL-BART	5	0.1	(0.5, 3)	-0.80 (4.37)	-1.52 (4.38)	-4.99 (4.73)	1.60 (4.57)	7.05 (4.63)
CL-BART	5	0.1	(0.95, 2)	-0.60 (4.35)	-2.09 (4.36)	-3.54 (4.72)	1.59 (4.61)	7.94 (4.60)
CL-BART	5	0.5	(0.5, 3)	-1.02 (4.37)	-1.24 (4.38)	-4.46 (4.72)	1.28 (4.59)	7.07 (4.57)
CL-BART	5	0.5	(0.95, 2)	-0.84 (4.34)	-1.05 (4.40)	-4.34 (4.70)	1.66 (4.59)	8.28 (4.58)
CL-BART	5	1.0	(0.5, 3)	-0.94 (4.32)	-1.18 (4.41)	-4.18 (4.71)	1.29 (4.58)	7.10 (4.58)
CL-BART	5	1.0	(0.95, 2)	-1.07 (4.34)	-1.51 (4.38)	-3.99 (4.70)	1.60 (4.56)	7.83 (4.59)
CL-BART	10	0.1	(0.5, 3)	-0.22 (4.38)	-1.02 (4.39)	-4.29 (4.73)	0.97 (4.55)	7.36 (4.61)
CL-BART	10	0.1	(0.95, 2)	-1.04 (4.39)	-1.91 (4.36)	-4.32 (4.77)	1.45 (4.65)	7.76 (4.63)
CL-BART	10	0.5	(0.5, 3)	-0.10 (4.39)	-1.33 (4.39)	-4.51 (4.71)	1.30 (4.58)	7.27 (4.58)
CL-BART	10	0.5	(0.95, 2)	-1.14 (4.40)	-1.66 (4.42)	-3.49 (4.72)	1.21 (4.60)	8.18 (4.64)
CL-BART	10	1.0	(0.5, 3)	-0.80 (4.37)	-2.03 (4.42)	-4.25 (4.75)	0.99 (4.60)	7.55 (4.58)
CL-BART	10	1.0	(0.95, 2)	-0.48 (4.36)	-1.40 (4.41)	-4.42 (4.73)	1.53 (4.58)	8.01 (4.60)
CL-BART	25	0.1	(0.5, 3)	-0.60 (4.38)	-1.47 (4.39)	-3.96 (4.71)	1.32 (4.60)	7.63 (4.61)
CL-BART	25	0.1	(0.95, 2)	-1.45 (4.39)	-1.53 (4.43)	-4.77 (4.71)	1.15 (4.64)	8.07 (4.61)
CL-BART	25	0.5	(0.5, 3)	-0.49 (4.38)	-1.26 (4.42)	-3.86 (4.71)	1.56 (4.60)	7.54 (4.59)
CL-BART	25	0.5	(0.95, 2)	-1.58 (4.36)	-2.28 (4.41)	-3.82 (4.72)	1.86 (4.62)	8.86 (4.61)
CL-BART	25	1.0	(0.5, 3)	-0.76 (4.41)	-1.95 (4.36)	-4.58 (4.72)	1.59 (4.58)	7.23 (4.62)
CL-BART	25	1.0	(0.95, 2)	-1.59 (4.39)	-1.94 (4.41)	-3.88 (4.74)	1.81 (4.59)	8.62 (4.60)
CL-BART	50	0.1	(0.5, 3)	-0.23 (4.41)	-1.84 (4.36)	-3.92 (4.74)	1.69 (4.60)	7.60 (4.59)
CL-BART	50	0.1	(0.95, 2)	-1.72 (4.40)	-1.93 (4.37)	-3.87 (4.76)	1.98 (4.62)	8.86 (4.62)
CL-BART	50	0.5	(0.5, 3)	-0.42 (4.40)	-1.96 (4.42)	-4.27 (4.75)	1.36 (4.63)	7.88 (4.67)
CL-BART	50	0.5	(0.95, 2)	-1.77 (4.37)	-1.34 (4.43)	-4.08 (4.70)	1.80 (4.63)	9.19 (4.64)
CL-BART	50	1.0	(0.5, 3)	-1.09 (4.41)	-2.00 (4.41)	-3.96 (4.70)	1.51 (4.60)	7.98 (4.65)
CL-BART	50	1.0	(0.95, 2)	-1.91 (4.37)	-1.72 (4.41)	-4.13 (4.75)	1.79 (4.64)	8.58 (4.66)
CL-BART	100	0.1	(0.5, 3)	-0.88 (4.38)	-2.29 (4.40)	-4.21 (4.72)	0.85 (4.60)	7.15 (4.60)
CL-BART	100	0.1	(0.95, 2)	-1.58 (4.39)	-1.89 (4.41)	-3.95 (4.75)	1.34 (4.61)	8.48 (4.66)
CL-BART	100	0.5	(0.5, 3)	-0.56 (4.40)	-1.71 (4.39)	-4.21 (4.73)	1.48 (4.62)	7.76 (4.59)
CL-BART	100	0.5	(0.95, 2)	-1.26 (4.37)	-1.50 (4.42)	-4.26 (4.73)	1.55 (4.61)	8.66 (4.62)
CL-BART	100	1.0	(0.5, 3)	-1.11 (4.39)	-1.81 (4.40)	-3.87 (4.69)	1.20 (4.64)	8.17 (4.64)
CL-BART	100	1.0	(0.95, 2)	-2.09 (4.41)	-1.98 (4.42)	-4.07 (4.73)	2.08 (4.62)	9.33 (4.62)

[a] T: Number of trees.

[b] k: Numerator of scale parameter for half-Cauchy hyper-prior.

[c] $(\alpha_{\mathcal{T}}, \beta_{\mathcal{T}})$: Hyperparameters for tree depth prior.

[d] Monte Carlo mean and standard errors across 200 simulations reported.

Table 3.10: Extended Friedman Simulation Results - Confounder Estimates (Coverage)

				Coverage[d]				
Type	T[a]	k[b]	$(\alpha_{\mathcal{T}}, \beta_{\mathcal{T}})$[c]	γ_1	γ_2	γ_3	γ_4	γ_5
Oracle				0.98	0.96	0.92	0.95	0.96
CL-BART	5	0.1	(0.5, 3)	0.97	0.96	0.92	0.93	0.94
CL-BART	5	0.1	(0.95, 2)	0.98	0.96	0.93	0.93	0.96
CL-BART	5	0.5	(0.5, 3)	0.97	0.97	0.93	0.93	0.95
CL-BART	5	0.5	(0.95, 2)	0.97	0.96	0.94	0.95	0.94
CL-BART	5	1.0	(0.5, 3)	0.98	0.96	0.93	0.93	0.94
CL-BART	5	1.0	(0.95, 2)	0.98	0.95	0.92	0.94	0.96
CL-BART	10	0.1	(0.5, 3)	0.98	0.96	0.92	0.95	0.95
CL-BART	10	0.1	(0.95, 2)	0.97	0.96	0.92	0.93	0.95
CL-BART	10	0.5	(0.5, 3)	0.97	0.96	0.92	0.93	0.95
CL-BART	10	0.5	(0.95, 2)	0.97	0.95	0.92	0.93	0.94
CL-BART	10	1.0	(0.5, 3)	0.97	0.96	0.94	0.93	0.95
CL-BART	10	1.0	(0.95, 2)	0.97	0.95	0.92	0.93	0.95
CL-BART	25	0.1	(0.5, 3)	0.98	0.97	0.92	0.93	0.96
CL-BART	25	0.1	(0.95, 2)	0.97	0.96	0.92	0.94	0.94
CL-BART	25	0.5	(0.5, 3)	0.97	0.97	0.92	0.93	0.95
CL-BART	25	0.5	(0.95, 2)	0.97	0.96	0.93	0.94	0.95
CL-BART	25	1.0	(0.5, 3)	0.97	0.96	0.92	0.93	0.95
CL-BART	25	1.0	(0.95, 2)	0.98	0.96	0.92	0.92	0.95
CL-BART	50	0.1	(0.5, 3)	0.96	0.96	0.92	0.94	0.95
CL-BART	50	0.1	(0.95, 2)	0.97	0.96	0.93	0.94	0.95
CL-BART	50	0.5	(0.5, 3)	0.97	0.96	0.92	0.94	0.95
CL-BART	50	0.5	(0.95, 2)	0.97	0.96	0.92	0.93	0.95
CL-BART	50	1.0	(0.5, 3)	0.97	0.95	0.93	0.94	0.94
CL-BART	50	1.0	(0.95, 2)	0.97	0.96	0.92	0.92	0.96
CL-BART	100	0.1	(0.5, 3)	0.97	0.96	0.93	0.94	0.95
CL-BART	100	0.1	(0.95, 2)	0.97	0.96	0.92	0.95	0.96
CL-BART	100	0.5	(0.5, 3)	0.96	0.95	0.92	0.93	0.95
CL-BART	100	0.5	(0.95, 2)	0.97	0.95	0.94	0.93	0.95
CL-BART	100	1.0	(0.5, 3)	0.97	0.96	0.92	0.94	0.94
CL-BART	100	1.0	(0.95, 2)	0.97	0.96	0.93	0.93	0.96

[a] T: Number of trees.

[b] k: Numerator of scale parameter for half-Cauchy hyper-prior.

[c] $(\alpha_{\mathcal{T}}, \beta_{\mathcal{T}})$: Hyperparameters for tree depth prior.

[d] 95% credible interval coverage across 200 simulations.

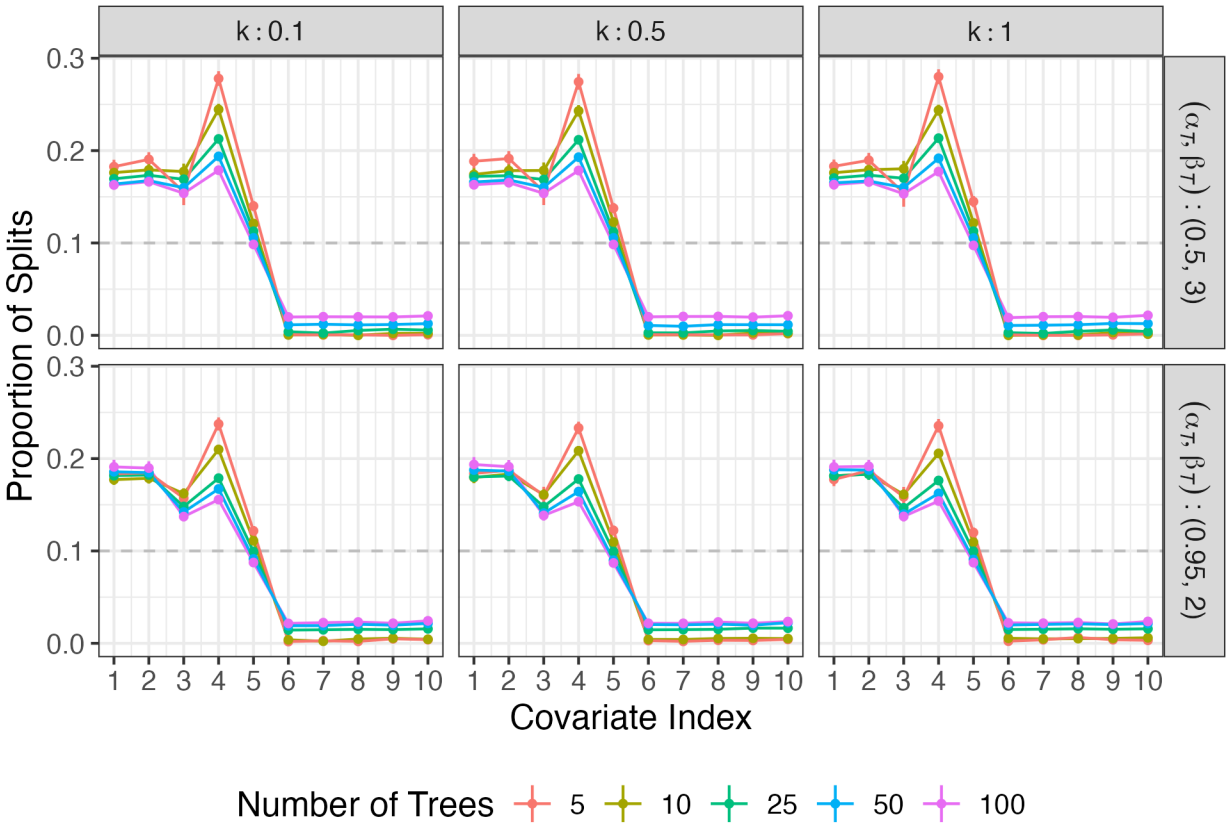

Figure 3.9: **Extended Friedman Simulation Variable Importance**: Plot of observed split proportions across 200 simulations (Monte Carlo mean and 95% uncertainty interval presented).

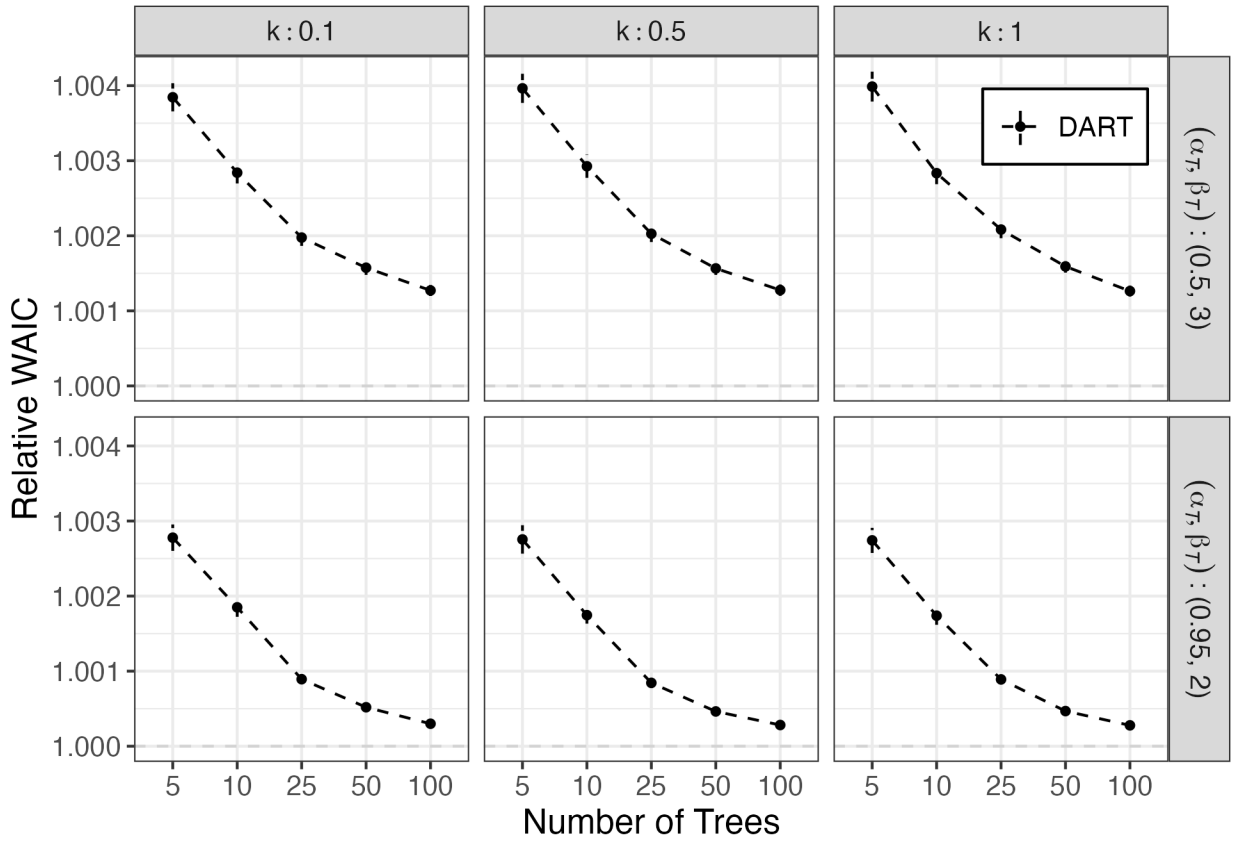

Figure 3.10: **WAIC for Friedman Simulation**: Relative WAIC for each simulation setting across 200 simulations (Monte Carlo mean and 95% uncertainty intervals presented).

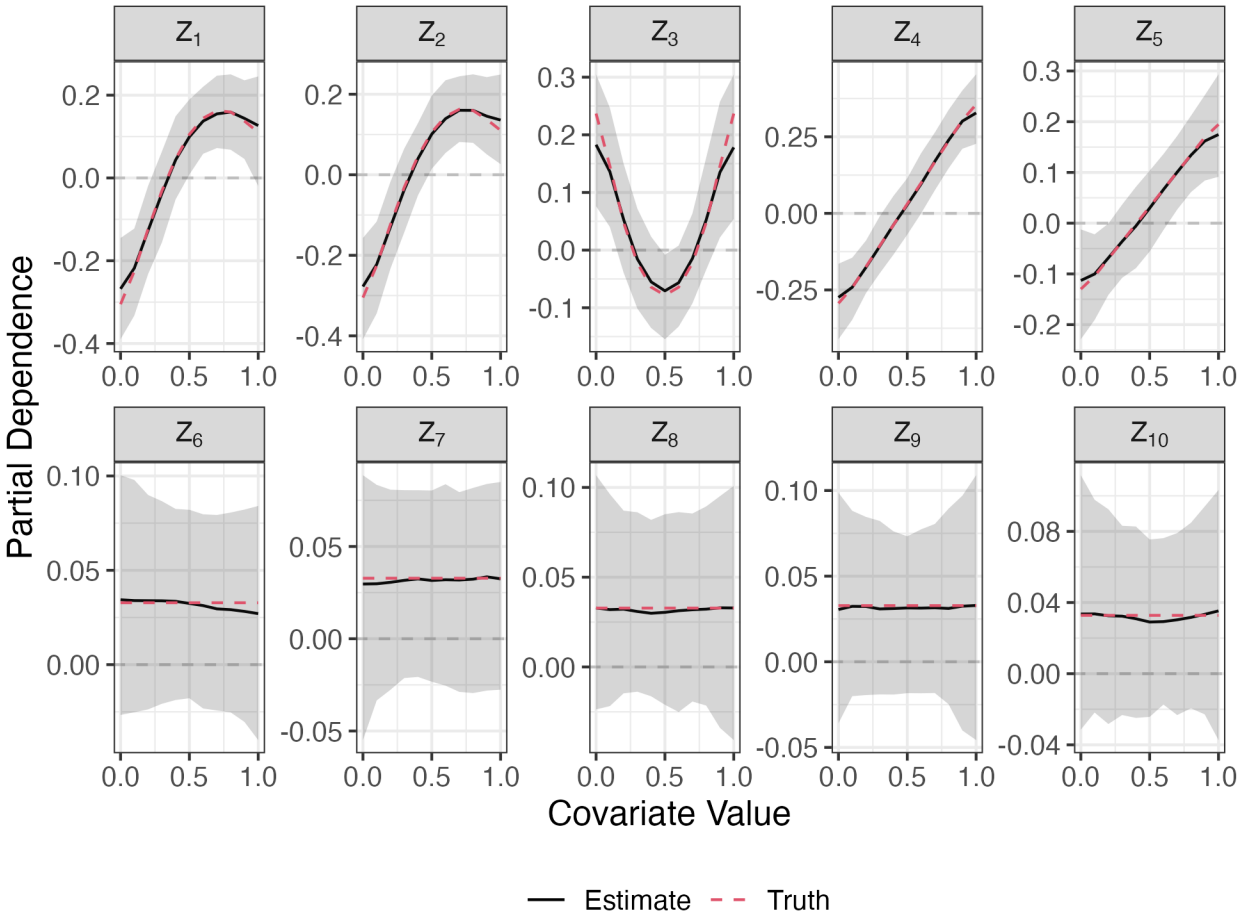

Figure 3.11: **Partial Dependence Plots for Friedman Simulation**: Posterior mean estimates of the partial dependence function for each covariate across 200 simulations with $(T, k, \alpha_{\mathcal{T}}, \beta_{\mathcal{T}}) = (100, 1, 0.95, 2)$ (simulation mean and 95% quantile intervals presented).

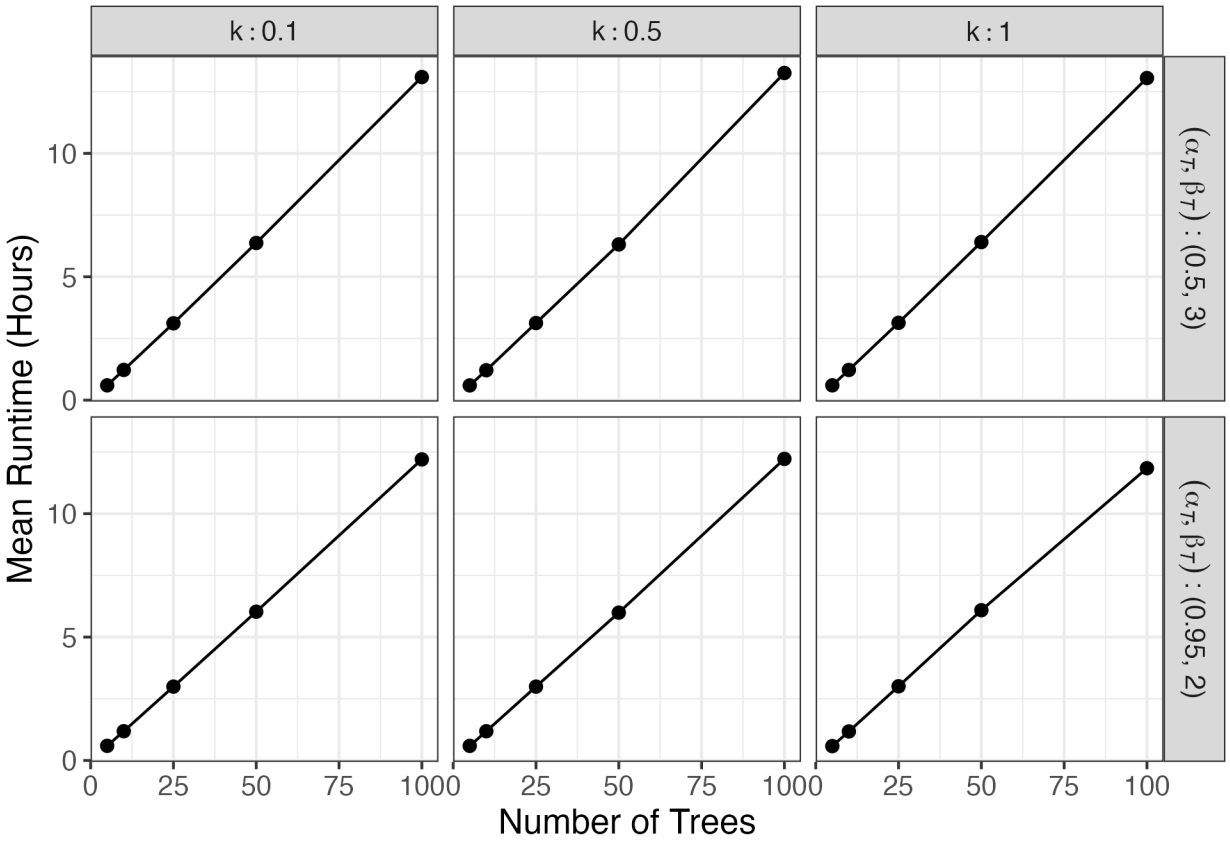

Figure 3.12: **Friedman Simulation Runtime**: Average time taken to run one CL-BART chain across 200 simulations. All models were run using 1 CPU on the high performance computing cluster at the Rollins School of Public Health, Emory University.

3.7.3 Additional Application Materials

List of Comorbid Condition Diagnosis Codes

We used both primary and secondary International Classification of Diseases (ICD) diagnosis codes to identify AD ED visits and comorbid conditions including chronic kidney disease, chronic obstructive pulmonary disease, congestive heart failure, depression, diabetes, hypertension, and hyperlipidemia. The exact codes are provided in Table 3.11. ICD-9 codes are used for records observed prior to October 1, 2015.

Table 3.11: List of International Classification of Diseases (ICD) Codes used to identify Alzheimer's Disease Emergency Department Visits and Comorbid Conditions

Condition	ICD-9 Codes	ICD-10 Codes
Alzheimer's Disease	331.0	G30.0, G30.1, G30.8, G30.9
Chronic Kidney Disease	016.00, 016.01, 016.02, 016.03, 016.04, 016.05, 016.06, 095.4, 189.0, 189.9, 223.0, 236.91, 249.40, 249.41, 250.40, 250.41, 250.42, 250.43, 271.4, 274.10, 283.11, 403.01, 403.11, 403.91, 404.02, 404.03, 404.12, 404.13, 404.92, 404.93, 440.1, 442.1, 572.4, 580.0, 580.4, 580.81, 580.89, 580.9, 581.0, 581.1, 581.2, 581.3, 581.81, 581.89, 581.9, 582.0, 582.1, 582.2, 582.4, 582.81, 582.89, 582.9, 583.0, 583.1, 583.2, 583.4, 583.6, 583.7, 583.81, 583.89, 583.9, 584.5, 584.6, 584.7, 584.8, 584.9, 585.1, 585.2, 585.3, 585.4, 585.5, 585.6, 585.9, 586, 587, 588.0, 588.1, 588.81, 588.89, 588.9, 591, 753.12, 753.13, 753.14, 753.15, 753.16, 753.17, 753.19, 753.20, 753.21, 753.22, 753.23, 753.29, 794.4	A18.11, A52.75, B52.0, C64.1, C64.2, C64.9, C68.9, D30.00, D30.01, D30.02, D41.00, D41.01, D41.02, D41.10, D41.11, D41.12, D41.20, D41.21, D41.22, D59.3, E08.21, E08.22, E08.29, E08.65, E09.21, E09.22, E09.29, E10.21, E10.22, E10.29, E10.65, E11.21, E11.22, E11.29, E11.65, E13.21, E13.22, E13.29, E74.8, I12.0, I12.9, I13.0, I13.10, I13.11, I13.2, I70.1, I72.2, K76.7, M10.30, M10.311, M10.312, M10.319, M10.321, M10.322, M10.329, M10.331, M10.332, M10.339, M10.341, M10.342, M10.349, M10.351, M10.352, M10.359, M10.361, M10.362, M10.369, M10.371, M10.372, M10.379, M10.38, M10.39, M32.14, M32.15, M35.04, N00.0, N00.1, N00.2, N00.3, N00.4, N00.5, N00.6, N00.7, N00.8, N00.9, N00.A, N01.0, N01.1, N01.2, N01.3, N01.4, N01.5, N01.6, N01.7, N01.8, N01.9, N01.A, N02.0, N02.1, N02.2, N02.3, N02.4, N02.5, N02.6, N02.7, N02.8, N02.9, N02.A, N03.0, N03.1, N03.2, N03.3, N03.4, N03.5, N03.6, N03.7, N03.8, N03.9, N03.A, N04.0, N04.1, N04.2, N04.3, N04.4, N04.5, N04.6, N04.7, N04.8, N04.9, N04.A, N05.0, N05.1, N05.2, N05.3, N05.4, N05.5, N05.6, N05.7, N05.8, N05.9, N05.A, N06.0, N06.1, N06.2, N06.3, N06.4, N06.5, N06.6, N06.7, N06.8, N06.9, N06.A, N07.0, N07.1, N07.2, N07.3, N07.4, N07.5, N07.6, N07.7, N07.8, N07.9, N07.A, N08, N13.1, N13.2, N13.30, N13.39, N14.0, N14.1, N14.2, N14.3, N14.4, N15.0, N15.8, N15.9, N16, N17.0, N17.1, N17.2, N17.8, N17.9, N18.1, N18.2, N18.3, N18.30, N18.31, N18.32, N18.4, N18.5, N18.6, N18.9, N19, N25.0, N25.1, N25.81, N25.89, N25.9, N26.1, N26.9, Q61.02, Q61.11, Q61.19, Q61.2, Q61.3, Q61.4, Q61.5, Q61.8, Q62.0, Q62.2, Q62.10, Q62.11, Q62.12, Q62.31, Q62.32, Q62.39, R94.4

(continued . . .)

Table 3.11: List of International Classification of Diseases (ICD) Codes used to identify Alzheimer's Disease Emergency Department Visits and Comorbid Conditions. *(continued)*

Condition	ICD-9 Codes	ICD-10 Codes
Chronic Obstructive Pulmonary Disease	491, 492, 496	J41, J42, J43, J44
Congestive Heart Failure	428	I42, I50, I51
Depression	296.20, 296.21, 296.22, 296.23, 296.24, 296.25, 296.26, 296.30, 296.31, 296.32, 296.33, 296.34, 296.35, 296.36, 296.51, 296.52, 296.53, 296.54, 296.55, 296.56, 296.60, 296.61, 296.62, 296.63, 296.64, 296.65, 296.66, 296.89, 298.0, 300.4, 309.1, 311	F31.30, F31.31, F31.32, F31.4, F31.5, F31.60, F31.61, F31.62, F31.63, F31.64, F31.75, F31.76, F31.77, F31.78, F31.81, F32.0, F32.1, F32.2, F32.3, F32.4, F32.5, F32.9, F33.0, F33.1, F33.2, F33.3, F33.40, F33.41, F33.42, F33.8, F33.9, F34.1, F43.21, F43.23
Diabetes	249, 250	E08, E09, E10, E11, E12, E13
Hyperlipidemia	272.0, 272.1, 272.2, 272.3, 272.4	E78.0, E78.00, E78.01, E78.1, E78.2, E78.3, E78.4, E78.41, E78.49, E78.5
Hypertension	401, 402, 403, 404, 405	I10, I11, I12, I13, I15

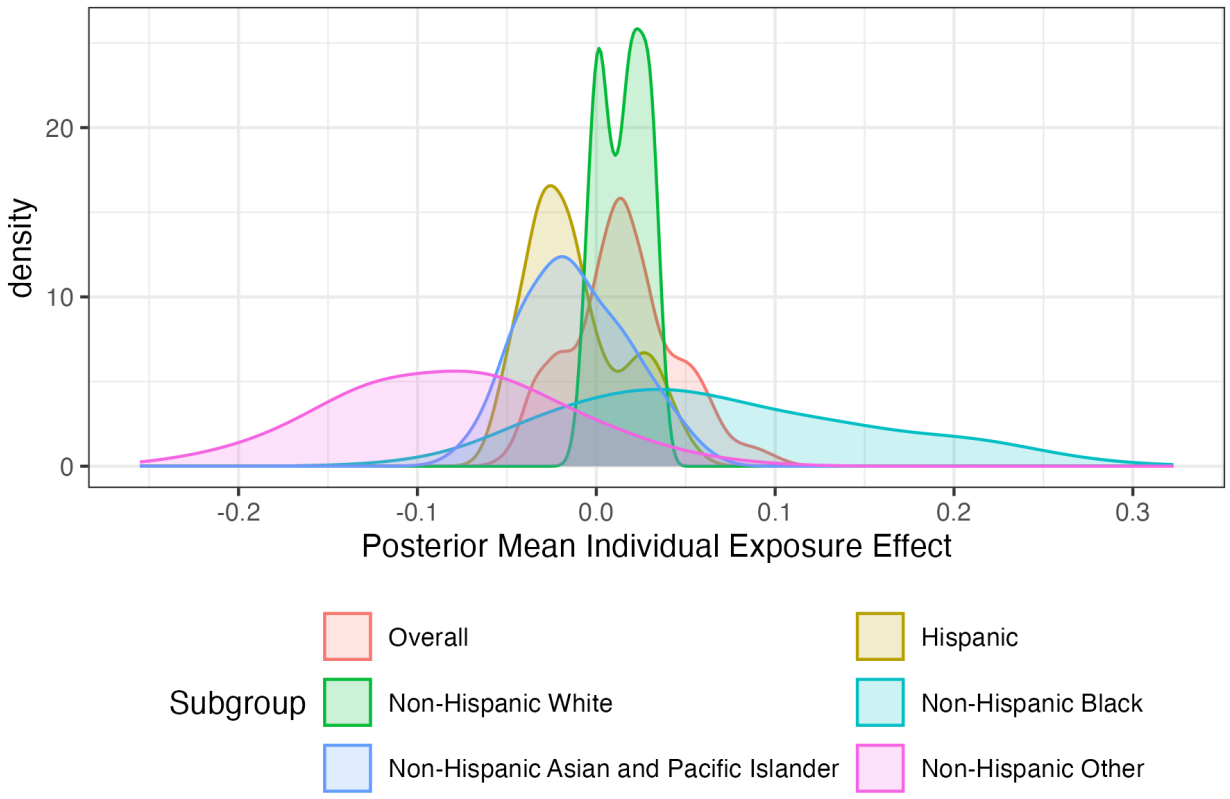

Figure 3.13: **Distribution of Individual Exposure Effects for the Alzheimer's Disease Application**: Density plots demonstrating the heterogeneity in point estimates of individual conditional exposure effects.

Table 3.12: Descriptive Statistics for Emergency Department Visits Among Alzheimer's Disease Patients by Subgroup, CA 2005-2015

Characteristic	HISP	NHW	NHB	NHAPI	NHO
n	11,959	46,019	5,635	5,521	1,886
Sex					
Male	4,356 (36.4%)	16,857 (36.6%)	1,839 (32.6%)	1,979 (35.8%)	731 (38.8%)
Female	7,603 (63.6%)	29,162 (63.4%)	3,796 (67.4%)	3,542 (64.2%)	1,155 (61.2%)
Age, yrs	83 (78, 88)	85 (80, 89)	82 (77, 88)	85 (80, 89)	84 (78, 88)
Number of Comorbid Conditions	2 (1, 3)	2 (1, 3)	2 (1, 3)	2 (1, 3)	2 (1, 3)
CHF	1,965 (16.4%)	7,287 (15.8%)	965 (17.1%)	960 (17.4%)	317 (16.8%)
CKD	3,042 (25.4%)	10,775 (23.4%)	1,925 (34.2%)	1,784 (32.3%)	411 (21.8%)
COPD	1,319 (11.0%)	5,655 (12.3%)	644 (11.4%)	670 (12.1%)	195 (10.3%)
Depression	1,519 (12.7%)	6,259 (13.6%)	432 (7.7%)	546 (9.9%)	249 (13.2%)
Diabetes	4,605 (38.5%)	8,687 (18.9%)	1,788 (31.7%)	2,066 (37.4%)	508 (26.9%)
Hypertension	8,212 (68.7%)	28,500 (61.9%)	4,296 (76.2%)	4,085 (74.0%)	1,188 (63.0%)
Hyperlipidemia	3,610 (30.2%)	13,731 (29.8%)	1,766 (31.3%)	1,936 (35.1%)	532 (28.2%)

n: sample size.

Median (IQR) reported for age. number of conditions.

HISP: Hispanic, NHW: Non-Hispanic White, NHB: Non-Hispanic Black, NHAPI: Non-Hispanic Asian and Pacific Islander, NHO: Non-Hispanic Other.

CHF: Congestive Heart Failure, CKD: Chronic Kidney Disease, COPD: Chronic Obstructive Pulmonary Disease.

Table 3.13: Marginal Partial Dependence Estimates for the Alzheimer's Disease Application (Overall)

Covariate	$\exp\left(\bar{\beta}_{DPD}\right)^a$
Hispanic	0.99 (0.95, 1.02)
Non-Hispanic White	0.98 (0.95, 1.01)
Non-Hispanic Black	1.01 (0.97, 1.07)
Non-Hispanic Asian and Pacific Islander	1.02 (0.99, 1.07)
Non-Hispanic Other	0.99 (0.94, 1.05)
Female	1.00 (0.98, 1.02)
CHF	0.99 (0.96, 1.02)
CKD	1.05 (1.01, 1.08)
COPD	0.99 (0.95, 1.02)
Depression	0.99 (0.96, 1.02)
Diabetes	1.00 (0.98, 1.03)
Hypertension	0.99 (0.96, 1.01)
Hyperlipidemia	1.01 (0.99, 1.04)

[a] $\bar{\beta}_{DPD}$: Difference in marginal partial average exposure effects.

Posterior mean and 95% credible interval presented.

Race/ethnicity covariates are mutually exclusive and one-hot encoded.

Table 3.14: Marginal Partial Dependence Estimates for the Alzheimer's Disease Application (Stratified Analysis)

Subgroup	Covariate	$\exp\left(\bar{\beta}_{DPD}\right)^a$
Hispanic	Female	0.99 (0.93, 1.02)
	CHF	0.99 (0.92, 1.04)
	CKD	1.06 (0.99, 1.20)
	COPD	1.00 (0.94, 1.07)
	Depression	1.01 (0.95, 1.08)
	Diabetes	0.98 (0.92, 1.02)
	Hypertension	0.99 (0.93, 1.02)
	Hyperlipidemia	1.00 (0.96, 1.05)
Non-Hispanic White	Female	1.00 (0.99, 1.02)
	CHF	1.00 (0.96, 1.01)
	CKD	1.01 (0.99, 1.04)
	COPD	1.00 (0.96, 1.02)
	Depression	0.99 (0.96, 1.02)
	Diabetes	1.00 (0.97, 1.02)
	Hypertension	1.00 (0.98, 1.01)
	Hyperlipidemia	1.01 (0.99, 1.04)
Non-Hispanic Black	Female	1.01 (0.93, 1.10)
	CHF	1.01 (0.91, 1.13)
	CKD	1.05 (0.96, 1.17)
	COPD	0.93 (0.79, 1.04)
	Depression	0.97 (0.81, 1.11)
	Diabetes	1.06 (0.98, 1.19)
	Hypertension	0.85 (0.74, 1.00)
	Hyperlipidemia	1.01 (0.94, 1.12)
Non-Hispanic Asian and Pacific Islander	Female	1.02 (0.97, 1.11)
	CHF	1.03 (0.96, 1.15)
	CKD	1.02 (0.96, 1.13)
	COPD	1.00 (0.91, 1.09)
	Depression	0.99 (0.87, 1.06)
	Diabetes	1.04 (0.98, 1.13)
	Hypertension	1.01 (0.94, 1.08)
	Hyperlipidemia	1.01 (0.95, 1.07)
Non-Hispanic Other	Female	0.94 (0.78, 1.04)
	CHF	1.02 (0.90, 1.22)
	CKD	1.09 (0.93, 1.35)
	COPD	1.00 (0.84, 1.19)
	Depression	1.03 (0.89, 1.22)
	Diabetes	0.96 (0.79, 1.05)
	Hypertension	1.00 (0.91, 1.14)
	Hyperlipidemia	0.96 (0.82, 1.05)

[a] $\bar{\beta}_{DPD}$: Difference in marginal partial average exposure effects.
Posterior mean and 95% credible interval presented.

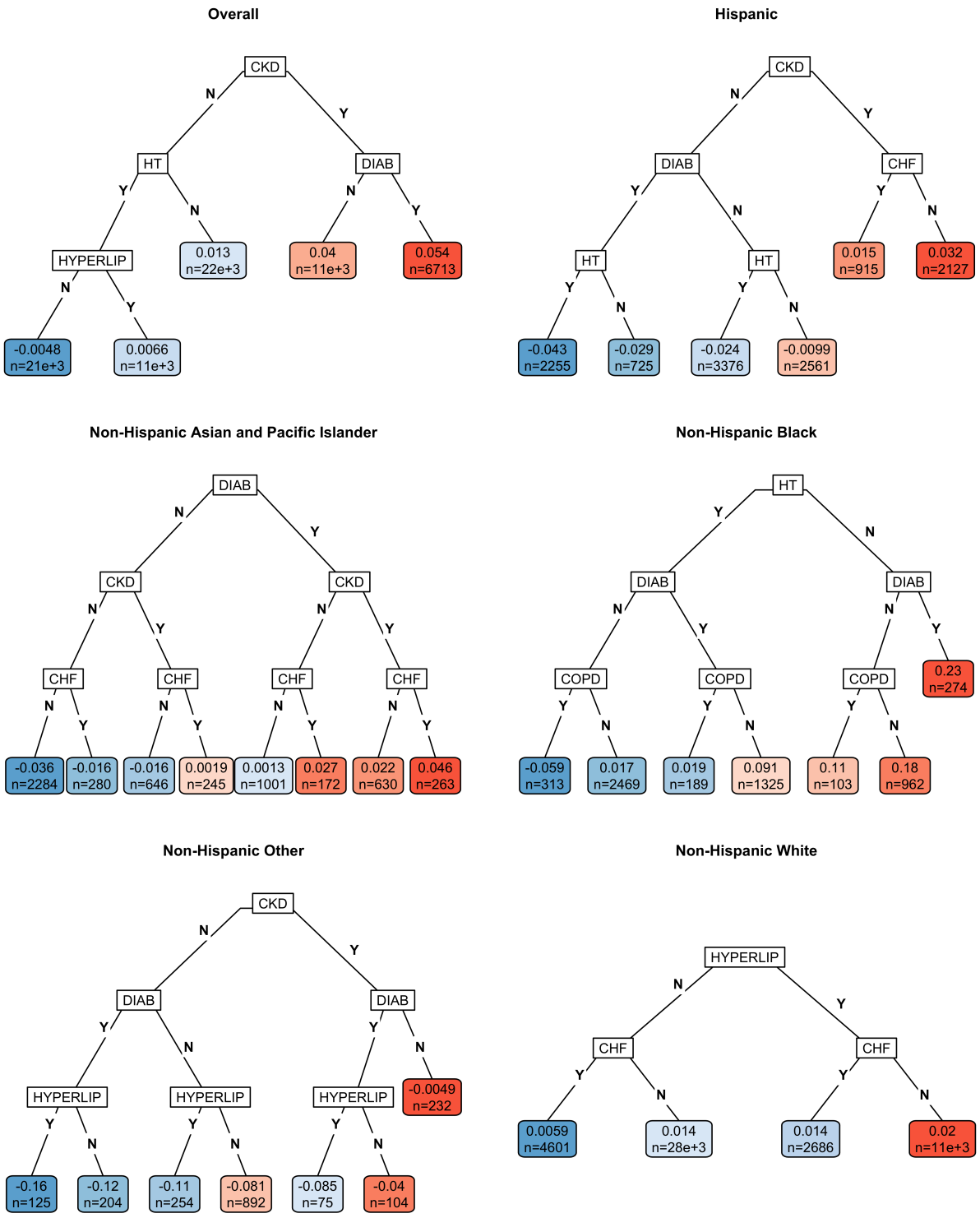

Figure 3.14: **Lower-dimensional CART summaries for the Alzheimer's Disease Application**: CART diagrams of lower-dimensional summaries for CL-BART model predictions (log odds ratio scale).

Table 3.15: Lower Dimensional CART Summary Partial Dependence for the Alzheimer's Disease Application

Subgroup	CHF	CKD	COPD	DEP	DIAB	HT	HLD	$\exp\left(\bar{\beta}_{PD}\right)^{a}$	$\Pr\left(\bar{\beta}_{PD} > 0\right)$	$R^{2\ b}$
Overall		-				-		1.01 (0.98, 1.05)	0.80	0.41
		+			-			1.05 (1.01, 1.09)	1.00	0.41
		+			+			1.05 (1.01, 1.10)	0.99	0.41
		-				+	-	1.00 (0.96, 1.03)	0.41	0.41
		-				+	+	1.01 (0.97, 1.04)	0.62	0.41
HISP	+	+						1.02 (0.94, 1.13)	0.65	0.89
	-	+						1.04 (0.96, 1.17)	0.74	0.89
		-			+	+		0.96 (0.85, 1.03)	0.20	0.89
		-			+	-		0.97 (0.88, 1.04)	0.27	0.89
		-			-	+		0.98 (0.90, 1.04)	0.28	0.89
		-			-	-		0.99 (0.92, 1.06)	0.40	0.89
NHW	+						-	1.01 (0.97, 1.05)	0.69	0.09
	-						-	1.01 (0.98, 1.05)	0.79	0.09
	+						+	1.02 (0.98, 1.06)	0.75	0.09
	-						+	1.02 (0.99, 1.07)	0.84	0.09
NHB					+	-		1.27 (1.04, 1.53)	0.99	0.79
			+		-	+		0.95 (0.77, 1.11)	0.29	0.79
			-		-	+		1.02 (0.90, 1.17)	0.63	0.79
			+		+	+		1.02 (0.84, 1.20)	0.56	0.79
			-		+	+		1.09 (0.96, 1.25)	0.88	0.79
			+		-	-		1.13 (0.92, 1.39)	0.89	0.79
			-		-	-		1.22 (1.02, 1.43)	0.99	0.79
NHAPI	-	-			-			0.97 (0.86, 1.07)	0.25	0.71
	+	-			-			0.99 (0.88, 1.13)	0.40	0.71
	-	+			-			0.99 (0.89, 1.11)	0.41	0.71
	+	+			-			1.01 (0.90, 1.16)	0.53	0.71
	-	-			+			1.00 (0.89, 1.12)	0.48	0.71
	+	-			+			1.03 (0.91, 1.21)	0.65	0.71
	-	+			+			1.02 (0.91, 1.16)	0.62	0.71
	+	+			+			1.05 (0.93, 1.23)	0.74	0.71
NHO		+			-			1.01 (0.80, 1.27)	0.46	0.39
		-			+		+	0.86 (0.63, 1.05)	0.10	0.39
		-			+		-	0.90 (0.69, 1.08)	0.16	0.39
		-			-		+	0.90 (0.71, 1.09)	0.16	0.39
		-			-		-	0.94 (0.76, 1.11)	0.24	0.39
		+			+		+	0.93 (0.72, 1.17)	0.25	0.39
		+			+		-	0.97 (0.76, 1.19)	0.34	0.39

[a] $\bar{\beta}_{PD}$: Partial average exposure effect. Posterior mean and 95% credible interval presented.

[b] Summary R^2.

HISP: Hispanic, NHW: Non-Hispanic White, NHB: Non-Hispanic Black, NHAPI: Non-Hispanic Asian and Pacific Islander, NHO: Non-Hispanic Other.

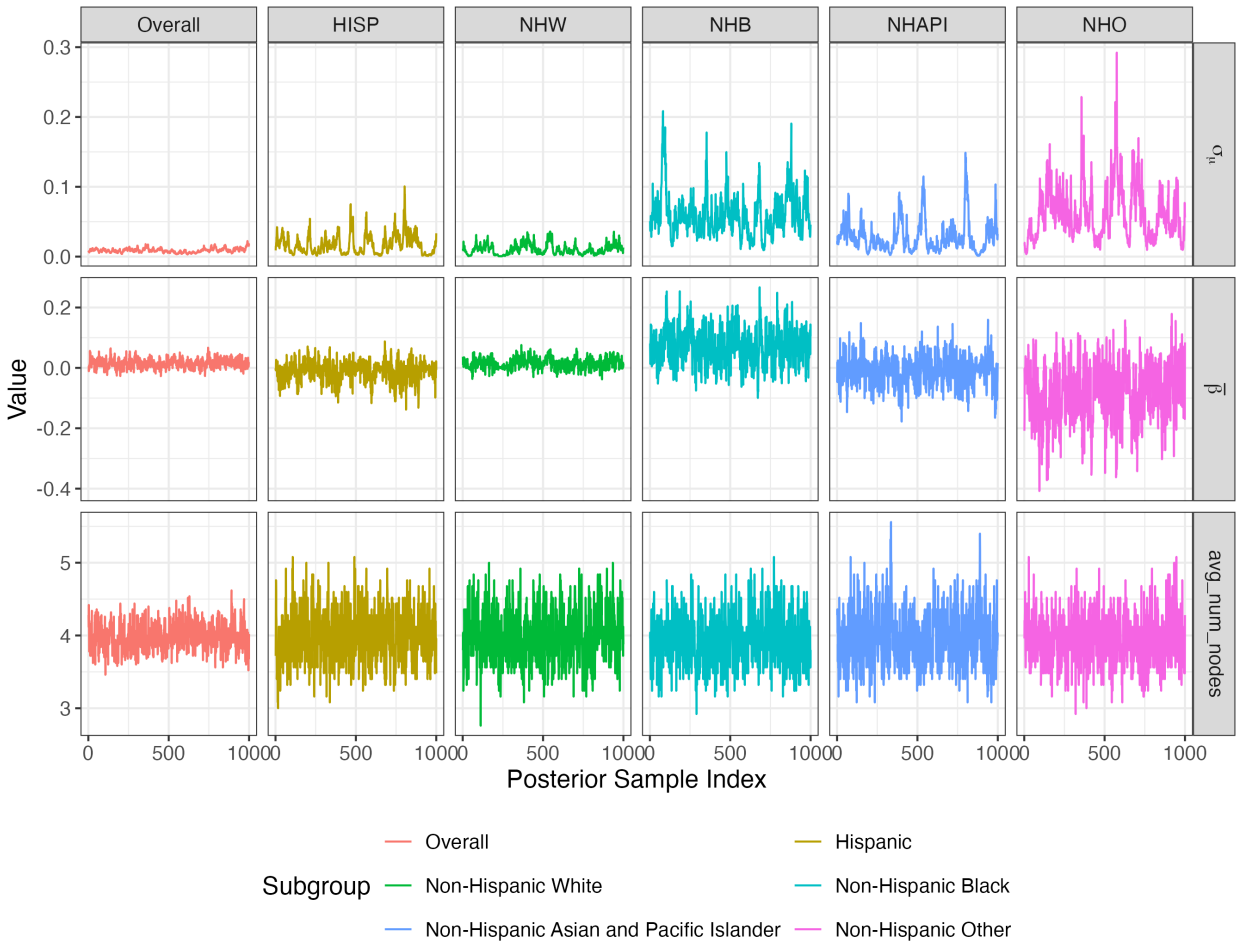

Figure 3.15: **Trace Plots for Selected Parameters**: Trace plots for each subgroup. Parameters from top to bottom include: σ_μ (leaf node prior standard deviation), $\bar{\beta}$ (average exposure effect, on the log scale), and the average number of nodes across all trees.

Table 3.16: Alzheimer's Disease Application CL-BART Model Runtimes

Subgroup	Sample Size	Runtime[a]
Overall	71,020	150.77
Hispanic	11,959	6.28
Non-Hispanic White	46,019	34.48
Non-Hispanic Black	5,635	3.77
Non-Hispanic Asian and Pacific Islander	5,521	3.11
Non-Hispanic Other	1,886	1.08

[a] Time taken to run one chain of a 25-tree CL-BART model (subgroups) or 100-tree CL-BART model (overall), in hours. All models were run using 1 CPU on the high performance computing cluster at the Rollins School of Public Health, Emory University.

Chapter 4

Modeling Joint Health Effects of Environmental Exposure Mixtures using BART

4.1 Introduction

Asthma affected an estimated 25 million (7.7%) individuals in the United States in the year 2021 per the National Health Interview Survey [21]. In that same year, the Healthcare Cost and Utilization Project estimated a total of 5.8 million asthma-related emergency department (ED) visits, of which 1.4 million required hospitalization and 930,000 listed asthma as the primary diagnosis [1]. In this work we are interested in studying the marginal and joint associations between elevated concentrations of multiple airborne chemical pollutants on asthma-related ED visit rates in Atlanta, Georgia.

Previous studies of Atlanta and other U.S. cities have found harmful associations between

asthma-related ED visits and environmental pollutants such as fine particulate matter of equal to or less than 2.5µm in diameter ($PM_{2.5}$), nitrogen dioxide (NO_2), ozone (O_3), and carbon monoxide (CO), among others [98, 55, 115, 2, 99, 86, 11]. The majority of these studies analyze exposures individually due to the challenges associated with modeling mixtures of correlated exposures. However, studies do often stratify analyses to study effect modification. For example, O'Lenick et al. [86] reported that neighborhood-level socioeconomic status may affect the association between between air pollution and pediatric asthma morbidity. Additionally, some studies have also reported modification of the effect of ozone on mortality by temperature [112].

Modeling of environmental mixtures is often framed in one of two ways: targeting a restricted class of research questions using easily interpretable parametric models, or estimating the true exposure-response surface with fancier (but less interpretable) models based on Gaussian processes, regression tree ensembles, etc. The former includes summary index approaches such as quantile g-computation [58] and weighted quantile sum [19], while the latter encompasses tools like Bayesian kernel machine regression (BKMR) [13], and more recently treed distributed lag mixture models (TDLMM) [80] and multiple exposure distributed lag models [5]. BKMR is most useful for estimating smooth exposure-response functions containing interactions and nonlinearities, while the latter two focus primarily on interaction and lagged effects over discrete time intervals. In our review, existing methods typically address at most two of 1) nonlinearity, 2) interaction, or 3) lagged effects (see Wilson et al. [114] for an approach which seeks to address all three). We propose using soft Bayesian additive regression trees (BART) [73] as an alternative to the BKMR approach for estimating interactions and nonlinearities. While soft BART is computationally slower than traditional BART, the tree-based approach is more feasible than BKMR when working with datasets

with a large number of observations, as in the motivating Atlanta dataset.

Following the fit of a mixture model, summarization of the estimated exposure-risk function with respect to one or two exposures often involves fixing the other exposures within the mixture to some quantile, or perhaps makes use of partial dependence statistics [36]. These approaches tend to extrapolate to implausible mixture levels in the estimation process, particularly in the context of correlated exposures. We propose using accumulated local effects [6] as an alternative approach that avoids this issue and has other benefits as well.

The main contribution of this work is to leverage a modeling approach based on Bayesian regression tree ensembles and subsequent summarization strategy for evaluating the effects of multi-pollutant mixtures on asthma morbidity in the city of Atlanta. We introduce the data for this application in Section 4.2 and outline the methodology in Section 4.3. We then demonstrate the utility of the proposed approach though a simulation study (Section 4.4) before finally presenting our main findings from the application (Section 4.5) and discussing areas of future work (Section 4.6).

4.2 Data

4.2.1 Health Data

Patient-level billing records for ED visits to hospitals in the metropolitan Atlanta area from 2011-2018 were obtained from the Georgia Hospital Association. These data included admission date, billing address, International Classification of Disease (ICD) version 9 or 10 discharge diagnosis codes, and various patient characteristics. We restricted the ED visit data to only include visits containing an asthma diagnosis (ICD-9 code 493; ICD-10 code J45) and occurring during Atlanta's warm season (April-October). These visits were then

aggregated by ZIP code and date for the analysis. These data have previously been used for various asthma and air pollution association studies [98, 99, 86, 62, 11].

4.2.2 Air Pollution Data

Air pollution concentration data are collected daily and include fine particulate matter with diameter 2.5µm and smaller ($PM_{2.5}$, 24-hr average, $\mu m/m^3$), ozone (O_3, 8-hr max, ppm), nitrogen dioxide (NO_2, 1-hr max, ppb), and carbon monoxide (CO, 1-hr max, ppb). The estimates are derived from the data fusion model described by Senthilkumar et al. [93], which utilized simulations from the Community Multiscale Air Quality Model and monitoring data from the Environmental Protection Agency's Air Quality System database. The data product is available at a 12km gridded spatial resolution and is linked to each ZIP code based on area-weighted averaging.

4.2.3 Other Data

Maximum daily temperature is obtained from Daymet [104]. The 1km gridded product was spatially averaged within each ZIP code, and linked to the ED visit data by both date and ZIP code. Annual ZIP code-level estimates of total population and the percent of the population below the poverty level are obtained from the 5-year American Community Survey for years 2011-2018.

4.3 Methods

4.3.1 Soft BART

Bayesian additive regression trees (BART) is a nonparametric machine learning approach which approximates complicated functions using sums of shallow Bayesian decision trees [25]. In this sense, the approach is similar to *boosting* from the machine learning literature. In recent years BART has soared in popularity due to its performance on prediction, classification, and causal inference tasks [51]. Ultimately, BART is a tree-based approach that can only approximate smooth functions with rigid fits. Linero and Yang [73] propose a *soft* version of BART, which adapts to smooth functions better than traditional BART by swapping traditional decision trees for soft decision trees [122]. In a soft decision tree, the prediction for an observation is a weighted average of all of the leaf node parameters, where the weights are defined as the probability an observation is mapped to each leaf node as determined by, say, a logistic gating function [73]. When compared to the deterministic predictions obtained from a traditional decision tree, this has the effect of smoothing over the otherwise rigid decision rules that form the binary tree. Since we generally expect the exposure-risk surface to be smooth, we opt to use this version of BART in our implementation. For more details on soft BART, see Section 2.1.3.

4.3.2 Negative Binomial Regression with BART

The model we propose is:

$$Y_{ij} \mid \theta_{ij} \sim \text{Poisson}(\theta_{ij}) \tag{4.1}$$

$$\theta_{ij} \sim \text{Gamma}(\xi, \exp(\eta_{ij})) \tag{4.2}$$

$$\eta_{ij} = \log(\text{Pop}_{ij}) + \boldsymbol{w}_{ij}^T\boldsymbol{\gamma} + f(\boldsymbol{x}_{ij}) + \nu_i, \tag{4.3}$$

where Y_{ij} and θ_{ij} represent the actual and expected asthma-related ED counts in region i on day j, respectively, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. The log expected counts are offset by Pop_{ij}, the population of region i at time j, and the overdispersion in the counts is represented by ξ. Potential confounders such as federal holidays and socioeconomic factors are represented by the $P_w \times 1$ vector $\boldsymbol{w}_{ij}$, while all exposures are represented by $P_x \times 1$ vector $\boldsymbol{x}_{ij}$. To account for additional unexplained variation in the counts due to location we include a ZIP-code specific random intercept ν_i.

The confounders are modeled linearly (or parametrically using splines) with regression coefficients given by the $P_w \times 1$ vector $\boldsymbol{\gamma}$, following previous studies. The exposures are modeled using soft BART [73]. Specifically, in Equation (4.3) we use $f(\boldsymbol{x}_{ij}) = \sum_{t=1}^{T} \text{Tree}(\boldsymbol{x}_{ij}; \mathcal{T}_t, \mathcal{M}_t)$. The $\{\mathcal{T}_t, \mathcal{M}_t\}_{t=1}^{T}$ parameters correspond to the tree structures and scalar-valued leaf node parameters associated with the soft BART model, and "Tree" is the function which maps a set of exposures $\boldsymbol{x}_{ij}$ to its prediction from a single soft decision tree. Our approach is similar to that used in Mutiso et al. [83], with the main difference being that we substitute the BKMR, which uses Gaussian processes, for soft BART.

4.3.3 Model Estimation

To reiterate, the log predictor for the model we propose is $\eta_{ij} = \log(\text{Pop}_{ij}) + \boldsymbol{w}_{ij}^T\boldsymbol{\gamma} + \sum_{t=1}^{T} \text{Tree}(\boldsymbol{x}_{ij}; \mathcal{T}_t, \mathcal{M}_t) + \nu_i$. The parameters in this component include $\boldsymbol{\gamma}, \{\mathcal{T}_t, \mathcal{M}_t\}_{t=1}^{T}$, and $\boldsymbol{\nu}$. To estimate these parameters, along with the dispersion parameter ξ, we adopt a Markov-chain Monte Carlo (MCMC) algorithm whose details are described in this section.

Negative Binomial Representation

The hierarchical Poisson-gamma model provided in (4.1) and (4.2) can be shown to have a marginal negative binomial distribution with density given by (4.4)

$$p(y_{ij} \mid \xi, \eta_{ij}) = \frac{\Gamma(y_{ij} + \xi)}{\Gamma(y_{ij} + 1)\Gamma(\xi)}(1 - p_{ij})^{\xi} p_{ij}^{y_{ij}} \propto (1 - p_{ij})^{\xi} p_{ij}^{y_{ij}}, \tag{4.4}$$

where

$$p_{ij} = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}. \tag{4.5}$$

The mean (4.6) and variance (4.7) may be obtained in a straightforward fashion by applying the laws of total expectation and total variance to the hierarchical model.

$$\mathbb{E}[Y_{ij}] = \mathbb{E}[\mathbb{E}(Y_{ij} \mid \theta_{ij})] = \mathbb{E}[\theta_{ij}] = \xi \exp(\eta_{ij}). \tag{4.6}$$

$$\begin{aligned} \text{Var}[Y_{ij}] &= \text{Var}[\mathbb{E}(Y_{ij} \mid \theta_{ij})] + \mathbb{E}[\text{Var}(Y_{ij} \mid \theta_{ij})] \\ &= \text{Var}[\theta_{ij}] + \mathbb{E}[\theta_{ij}] \\ &= \xi \exp(2\eta_{ij}) + \xi \exp(\eta_{ij}) \\ &= \xi \exp(\eta_{ij})[1 + \exp(\eta_{ij})]. \end{aligned} \tag{4.7}$$

Since Var $[Y_{ij}] > \mathbb{E}[Y_{ij}]$, it is clear that ξ is exclusively modeling *overdispersion*. The marginal likelihood of all of the observed data can be written as in (4.8).

$$p(\boldsymbol{y} \mid \xi, \boldsymbol{\eta}) \propto \prod_{i=1}^{I}\prod_{j=1}^{J}(1-p_{ij})^{\xi}p_{ij}^{y_{ij}} = \prod_{i=1}^{I}\prod_{j=1}^{J}\frac{\{\exp(\eta_{ij})\}^{y_{ij}}}{\{1+\exp(\eta_{ij})\}^{y_{ij}+\xi}}. \tag{4.8}$$

Pólya-Gamma Data Augmentation

BART (and by extension, soft BART), has been extended to many different outcome types since the original model was proposed, however these extensions typically require conditional conjugacy between the outcome distribution and prior distribution on the individual leaf node parameters to facilitate the sampling of tree structures. Since the negative binomial likelihood in (4.8) does not itself admit a conditionally conjugate prior, we adopt the framework proposed by Pillow and Scott [88]. Specifically, we augment the outcome Y_{ij} with latent weights ω_{ij} sampled from a Pólya-gamma (PG) distribution [89].

If $\omega \sim \text{PG}(b, 0)$, then for any choice of a, we have the following result:

$$\frac{(e^{\eta})^{a}}{(1+e^{\eta})^{b}} = 2^{-b}e^{\kappa\eta}\int_{0}^{\infty} e^{-\omega\eta^{2}/2}p(\omega)d\omega, \tag{4.9}$$

where $\kappa = a - b/2$. This follows from the definition of PG random variables [89]. Substituting (4.8) into the LHS of this result gives (4.10)

$$p(\boldsymbol{y} \mid \xi, \boldsymbol{\eta}) \propto \prod_{i=1}^{I}\prod_{j=1}^{J}\exp(\kappa_{ij}\eta_{ij})\int_{0}^{\infty}\exp(-\omega_{ij}\eta_{ij}^{2}/2)p(\omega_{ij} \mid y_{ij}+\xi, 0)d\omega_{ij}, \tag{4.10}$$

where $\kappa_{ij} = \frac{y_{ij}-\xi}{2}$. If we condition on ω_{ij}, the expectation (integral) in (4.10) can be ignored, and after completing the square we may obtain the following form for the data augmented

data likelihood:

$$
\begin{aligned}
p(\boldsymbol{y} \mid \xi, \boldsymbol{\eta}, \boldsymbol{\omega}) &\propto \prod_{i=1}^{I} \prod_{j=1}^{J} \exp\left(\kappa_{ij}\eta_{ij}\right) \times \exp\left(-\omega_{ij}\eta_{ij}^2/2\right) \\
&\propto \prod_{i=1}^{I} \prod_{j=1}^{J} \exp\left\{-\frac{\omega_{ij}}{2}\left(\frac{y_{ij}-\xi}{2\omega_{ij}} - \eta_{ij}\right)^2\right\},
\end{aligned}
\tag{4.11}
$$

where $\boldsymbol{\omega} = (\omega_{11}, \ldots, \omega_{1J}, \ldots, \omega_{I1}, \ldots, \omega_{IJ})^T$. Thus, if we independently draw $\omega_{ij} \sim \text{PG}(y_{ij} + \xi, \eta_{ij})$, then the latent outcome $y_{ij}^* = \frac{y_{ij}-\xi}{2\omega_{ij}}$ is normally distributed with mean η_{ij} and variance $1/\omega_{ij}$. It follows that the vector of latent outcomes $\boldsymbol{y}^* \sim \text{MVN}(\boldsymbol{\eta}, \boldsymbol{\Omega}^{-1})$, where $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\omega})$. This leads to a convenient Gibbs sampler based on the Bayesian backfitting approach of Hastie and Tibshirani [49] for updating $\boldsymbol{\gamma}$, $\{\mathcal{M}_t\}_{t=1}^T$, and $\boldsymbol{\nu}$, when multivariate normal prior distributions are chosen. The full forms of these conjugate updates are provided in the Chapter 4 Supplemental Materials.

Updating BART Parameters

In BART, new tree structures are sampled from their marginal distribution and updated using a Metropolis-Hastings (M-H) step by first integrating out the leaf node parameters as described in Section 2.1.3. The tree structures themselves are proposed from a so-called *branching process* prior, which modify the existing structures from the previous iteration [25, 90]. A detailed explanation of how this approach works with weights, such as those introduced by the Pólya-gamma data augmentation scheme, is outlined in Bleich and Kapelner [12]. The tree prior used for soft BART is more involved, optionally including hyperparameters and hyperpriors responsible for the degree of smoothness and/or sparsity, but the general concept is the same. One of the benefits of BART is that default priors tend to work well in a variety of circumstances. We use default priors for updating $\{\mathcal{T}_t, \mathcal{M}_t\}_{t=1}^T$

as detailed in Linero and Yang [73]. In our implementation, the sampling of BART-related parameters is facilitated by the drop-in C++ module available in the `SoftBart` R package `https://github.com/theodds/SoftBART`.

A Spatial CAR Prior for ZIP Code Random Intercepts

We assign a mean-zero Besag proper conditional autoregressive (CAR) prior [10] for the ZIP code-level random intercepts $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_I)^T$. In mathematical terms, $\boldsymbol{\nu} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_\nu)$ with $\boldsymbol{\Sigma}_\nu = \tau^2(\mathbf{D} - \rho\mathbf{A})^{-1}$, where $\mathbf{A}$ is the $I \times I$ first-order adjacency matrix, and $\mathbf{D}$ is the $I \times I$ diagonal matrix containing the number of neighbors for each region. This allows for a portion of the variability in the response unexplained by the predictors to be attributed to unmeasured spatial factors. We assign an inverse gamma and discrete uniform hyperprior to τ^2 and ρ, respectively.

Updating the Dispersion Parameter

Lastly, we update the dispersion parameter ξ using the conjugate sampling routine described by Zhou et al. [121]. This technique relies on expressing the $Y_{ij} \sim \text{NB}(\xi, p_{ij})$ marginal distribution as a compound Poisson distribution. The details for this step and the others described in this section are outlined in Section 4.7.1 of the Chapter 4 Supplementary Materials. A summary of a single MCMC iteration is provided in Algorithm 4.5.

4.3.4 Model Interpretation via Accumulated Local Effects

When using flexible methods which target the response surface directly, interpretation of the resulting fit requires just as much thought as the estimation itself. In the environmental mixtures setting, it is common to focus on the marginal effect of a single exposure on the

Algorithm 4.5 One MCMC Iteration of the Spatial Soft BART Negative Binomial Algorithm

Input: $\mathcal{D} = \{\boldsymbol{y}, \mathbf{W}, \mathbf{X}, \mathbf{Z}, \text{Population}, \mathbf{D}, \mathbf{A}\}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \tau^2, \rho, \{\mathcal{T}_t, \mathcal{M}_t\}_{t=1}^T, \xi, \alpha_\tau, \beta_\tau, \alpha_\xi, \beta_\xi$
for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ **do**
    Draw $\omega_{ij} \sim \text{PG}(y_{ij} + \xi, \eta_{ij})$.
    Compute $y_{ij}^* = \frac{y_{ij} - \xi}{2\omega_{ij}}$.
end for
Form $\boldsymbol{\Omega} = \text{diag}(\omega_{11}, \ldots, \omega_{1,J}, \ldots, \omega_{I1}, \ldots, \omega_{IJ})$.
Draw $\boldsymbol{\gamma} \sim \text{MVN}\left(\boldsymbol{\mu}_\gamma^*, \boldsymbol{\Sigma}_\gamma^*\right)$, where $\boldsymbol{\mu}_\gamma^*$ and $\boldsymbol{\Sigma}_\gamma^*$ are defined as in Section 4.7.1.
Draw $\boldsymbol{\nu} \sim \text{MVN}\left(\boldsymbol{\mu}_\nu^*, \boldsymbol{\Sigma}_\nu^*\right)$, where $\boldsymbol{\mu}_\nu^*$ and $\boldsymbol{\Sigma}_\nu^*$ are defined as in Section 4.7.1.
Draw $\tau^2 \sim \text{IG}\left(\alpha_\tau + I/2, \beta_\tau + \boldsymbol{\nu}^T(\mathbf{D} - \rho\mathbf{A})\boldsymbol{\nu})/2\right)$.
Draw ρ from its discrete posterior distribution described in Section 4.7.1.
for $t = 1, \ldots, T$ **do**
    Propose/update $\mathcal{T}_t, \mathcal{M}_t$, and any associated hyperparameters governing the degree of smoothness and/or sparsity as described in Linero and Yang [73].
end for
Draw $L_{ij} \sim \text{CRT}(\xi, y_{ij})$ for $i = 1, \ldots, I$, $j = 1, \ldots, J$.
Draw $\xi \sim \text{Gamma}\left(\alpha_\xi + \sum_{i=1}^I \sum_{j=1}^J L_{ij}, \beta_\xi - \sum_{i=1}^I \sum_{j=1}^J \ln(1 - p_{ij})\right)$.

outcome by evaluating the exposure-response function at several levels of the chosen exposure and plotting the result. This strategy can also be used for studying the joint effects of two exposures using, say, contour plots. Analyzing or visualizing joint effects of more than two continuous exposures is rather difficult, and thus is not as common. Mathematically, if we are we are interested in evaluating the effect of $\mathbf{X}_p$ on a fitted exposure-response function $\hat{f}$, we evaluate $\hat{f}\left(\boldsymbol{x}_p \mid \mathbf{X}_{-p}\right)$ for several values $\boldsymbol{x}_p$ in the observed range of $\mathbf{X}_p$.

In settings where there are more than two exposures, a decision must be made regarding the treatment of the exposures that aren't of interest, $\mathbf{X}_{-p}$, when evaluating the exposure-response function for 1-2 exposures of interest $\mathbf{X}_p$. As outlined in Section 2.2, there are a few options for this. A common choice is to set $\mathbf{X}_{-p}$ to some fixed values (e.g., their observed medians) while varying $\mathbf{X}_p$. In fact, one might set $\mathbf{X}_{-p}$ to multiple values (e.g., their medians and 95th percentiles), and plot $\hat{f}\left(\boldsymbol{x}_p \mid \mathbf{X}_{-p}\right)$ for each setting. This approach is not ideal since the exposure-response function ultimately depends on the selected values for $\mathbf{X}_{-p}$, of which

there are many choices for each exposure in the model - none of which are the perfect choice, and many of which are poor choices.

Another option is to calculate partial dependence (PD) functions [36]. PD functions target the average exposure effect across the marginal distribution of the observed data, setting $\mathbf{X}_{-p}$ to their observed values $\boldsymbol{x}_{ij,-p}, i = 1, \ldots, I, j = 1, \ldots, J$. In this manner, PD functions avoid having to make a choice regarding the values of $\mathbf{X}_{-p}$, and the resulting estimates naturally incorporate the variability in $\mathbf{X}_{-p}$ across specified values of $\mathbf{X}_p$. One of the major limitations of PD functions is that they are computationally burdensome, requiring evaluations of $\hat{f}$ for every observation in the study at every value considered for $\mathbf{X}_p$.

Both the fixed-value and PD approaches run the risk of extrapolating when evaluating $\hat{f}\left(\boldsymbol{x}_p \mid \mathbf{X}_{-p}\right)$, particularly if the exposures are correlated (which they often are). By this we mean that some of the evaluations of $\hat{f}\left(\boldsymbol{x}_p \mid \mathbf{X}_{-p}\right)$ are made on implausible exposure profiles. This is a general issue for assessing covariate effects in black-box supervised learning models. One approach that has been proposed to combat this issue is accumulated local effects (ALE, [6]). The estimands for the partial effect of a single exposure X_p at some level x_p for each of the three approaches are provided in (4.12), (4.13), and (4.14).

$$f_{p,Fixed}(x_p) \equiv \mathbb{E}\left[\hat{f}\left(x_p, \boldsymbol{x}_{-p}\right)\right] \tag{4.12}$$

$$f_{p,PD}(x_p) \equiv \mathbb{E}\left[\ \hat{f}\left(x_p, \mathbf{X}_{-p}\right)\right] \tag{4.13}$$

$$f_{p,ALE}(x_p) \equiv \int_{x_{min,p}}^{x_p} \mathbb{E}\left[\frac{\partial \hat{f}}{\partial X_p}\left(X_p, \mathbf{X}_{-p}\right) \mid X_p = x_p'\right] dx_p' \tag{4.14}$$

The quantities in (4.12) and (4.13) are estimated using (4.15) and (4.16), respectively.

$$\hat{f}_{p,Fixed}(x_p) = f\left(x_p, \boldsymbol{x}_{-p}\right). \tag{4.15}$$

$$\hat{f}_{p,PD}(x_p) = \frac{1}{IJ}\sum_{i,j}\hat{f}\left(x_p, \boldsymbol{x}_{ij,-p}\right). \tag{4.16}$$

Estimation of the uncentered ALE main effect of X_p is a bit more involved. We first rewrite the estimand using the limit definition of a derivative as in (4.17).

$$f_{p,ALE}(x_p) \equiv \lim_{K\to\infty}\sum_{k=1}^{k_p^K(x_p)}\mathbb{E}\left[\hat{f}\left(x_{k,p}^K, \mathbf{X}_{-p}\right) - \hat{f}\left(x_{k-1,p}^K, \mathbf{X}_{-p}\right) \mid X_p \in \left(x_{k-1,p}^K, x_{k,p}^K\right]\right], \tag{4.17}$$

where K is the number of intervals in the range of X_p over which local effects are estimated. The k^{th} interval is defined as $\left(x_{k-1,p}^K, x_{k,p}^K\right]$, and $k_p^K(x_p)$ is the interval associated with the value of interest x_p. The estimator of the uncentered ALE main effect is given by (4.18)

$$\hat{f}_{p,ALE}(x_p) = \sum_{k=1}^{k_p^K(x_p)}\frac{1}{N_k}\sum_{i,j:x_{ij,p}\in\left(x_{k-1,p}^K, x_{k,p}^K\right]}\left[\hat{f}\left(x_{k,p}^K, \boldsymbol{x}_{ij,-p}\right) - \hat{f}\left(x_{k-1,p}^K, \boldsymbol{x}_{ij,-p}\right)\right], \tag{4.18}$$

where N_k is the number of observations having $X_p \in \left(x_{k-1,p}^K, x_{k,p}^K\right]$.

Usually we wish to estimate any of these functions for M values in the range of X_p. For estimation purposes, (4.15) requires M evaluations of $\hat{f}$ at $(x_p, \boldsymbol{x}_{-p})$ (one at each of the M values of interest), while (4.16) requires M evaluations of $\hat{f}$ at $(x_p, \boldsymbol{x}_{ij,-p})$ for all i, j (MIJ total evaluations). In contrast, the estimator in (4.18) requires two evaluations of $\hat{f}$ for all i, j – one at $\left(x_{k,p}^K, \boldsymbol{x}_{ij,-p}\right)$ and one at $\left(x_{k-1,p}^K, \boldsymbol{x}_{ij,-p}\right)$ – such that $x_{ij,p} \in \left(x_{k-1,p}^K, x_{k,p}^K\right]$, for a total of $2IJ$ total evaluations. The latter is a result of approximating $\frac{\partial \hat{f}}{\partial X_p}$ with small finite differences. An illustration for the ALE computation has been included in Section 2.2.3. For

more information regarding ALE computation, we refer the reader to Apley and Zhu [6]. When using a Bayesian approach, each of (4.15), (4.16), and (4.18) would be evaluated for each sample from the posterior distribution. This allows one to obtain pointwise posterior means and uncertainty estimates at each of the M values of interest.

The primary benefits of using ALE over the other approaches are threefold: 1) effects of correlated exposures are isolated by targeting the partial derivative of $\hat{f}$, 2) estimates are only informed by predictions on plausible exposure profiles since averaging is done using the conditional distribution (i.e., no extrapolation), and 3) the computation is relatively fast compared to PD functions since only two predictions are needed for each observation, regardless of the number of levels of X_p being considered. Since it is common to consider 40 or more values of X_p, the last point is a significant advantage to using ALE.

4.4 Simulation Study

To evaluate the proposed approach, specifically to different BART specifications, we conduct a brief simulation study. We use the populations and locations of the 128 ZIP codes from the first year (2011) of the application. We set confounder effects $\boldsymbol{\gamma} = (-2, -1, 1, 2)^T$ and the true exposure-risk surface function to $f(\mathbf{X}) = -10 + \frac{f_0(\mathbf{X})}{5}$, where $f_0(\mathbf{X}) = 10\sin(X_1X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5$ is the benchmark function proposed in Friedman [35]. While this surface only depends on five exposures, we generate five additional noise exposures (ten total exposures) to assess the performance in settings where not all exposures are important. Spatial random effects are sampled from a proper CAR prior with $\rho = 0.9$ and $\tau^2 = 0.3$. For $i = 1, \ldots, 128$ regions and $j = 1, \ldots, 300$ observations, we simulate outcomes using the following data generating process:

1. Generate $W_{ij,1}, W_{ij,2}, W_{ij,3}, W_{ij,4} \overset{i.i.d.}{\sim} \text{Uniform}(0,1)$.

2. Generate $\mathbf{X}_{ij} \sim \text{MVN}(\mathbf{0}_{10\times 1}, \mathbf{\Sigma}_{10\times 10})$, where $\mathbf{\Sigma}_{[1:5,1:5]}$ matches the observed correlation matrix of $PM_{2.5}$, NO_2, O_3, CO, and maximum temperature from the application. Only the first five exposures are used to generate the outcome. All values are scaled to $[0,1]$ using min-max normalization.

3. Sample $Y_{ij} \sim \text{NB}\left(\xi = 1, p_{ij} = \frac{e^{\eta_{ij}}}{1+e^{\eta_{ij}}}\right)$ where $\eta_{ij} = \log\left(\text{Pop}_{ij}\right) + \boldsymbol{w}_{ij}^T\boldsymbol{\gamma} + f(\boldsymbol{x}_{ij}) + \nu_i$

We consider ensembles of size $T = \{10, 25, 50, 100\}$, both hard and soft decision rules, and both the classic and sparse branching processes [70]. Each setting is repeated 200 times. The average bias, root mean squared error (RMSE), and 95% credible interval coverage for $f(\mathbf{X})$ are presented in Table 4.1, along with Monte Carlo standard error estimates.

Soft BART had excellent bias, coverage, and RMSE even when few trees were used. When many trees were used ($T = 100$), performance of traditional BART improved, but was still worse than soft BART in terms of coverage and RMSE. In general, increasing the number of trees beyond 25 did not appear to improve the performance of soft BART. Additionally, using the sparse branching process prior mostly resulted in improved coverage and reduced RMSE (Table 4.1). This suggests the sparsity-inducing Dirichlet prior was effective at identifying the important exposures and avoiding tree splitting rules based on the noise exposures.

We present simulation results for the marginal ALE plots for each exposure using the $T = 25$ soft, sparse trees setting in Figure 4.1. On average across simulations, the true functional forms of the five important exposures is recovered remarkably well. The null effects of the noise exposures are also accurately captured, primarily due to the ensembles avoiding splitting on these covariates entirely. Similar results for the pairwise ALE plots are included in Figures 4.9 and 4.10 in the Chapter 4 Supplementary Materials.

Table 4.1: Soft BART Simulation Results - BART Predictions.

T[a]	Soft[b]	Sparse[c]	Bias ×10 (MCSE)	Coverage (MCSE)	RMSE ×10 (MCSE)
10			0.11 (0.0138)	0.88 (0.0229)	1.17 (0.0050)
10		✓	0.11 (0.0142)	0.96 (0.0143)	1.06 (0.0043)
10	✓		0.02 (0.0139)	0.91 (0.0207)	0.41 (0.0065)
10	✓	✓	0.02 (0.0137)	0.92 (0.0188)	0.41 (0.0061)
25			0.05 (0.0141)	0.78 (0.0293)	0.98 (0.0037)
25		✓	0.05 (0.0138)	0.89 (0.0221)	0.89 (0.0039)
25	✓		0.01 (0.0139)	0.94 (0.0167)	0.39 (0.0056)
25	✓	✓	0.01 (0.0137)	0.95 (0.0155)	0.37 (0.0056)
50			0.02 (0.0136)	0.81 (0.0279)	0.89 (0.0035)
50		✓	0.02 (0.0139)	0.88 (0.0228)	0.82 (0.0037)
50	✓		0.01 (0.0137)	0.95 (0.0147)	0.42 (0.0050)
50	✓	✓	0.02 (0.0140)	0.96 (0.0140)	0.37 (0.0054)
100			0.01 (0.0140)	0.88 (0.0233)	0.85 (0.0035)
100		✓	0.01 (0.0137)	0.92 (0.0188)	0.77 (0.0037)
100	✓		0.02 (0.0138)	0.96 (0.0131)	0.46 (0.0047)
100	✓	✓	0.01 (0.0136)	0.96 (0.0130)	0.39 (0.0054)

CrI: Bayesian posterior credible interval.

MCSE: Monte Carlo Standard Error.

[a] T: Number of trees.

[b] Soft BART used [73].

[c] Sparse branching process used [70].

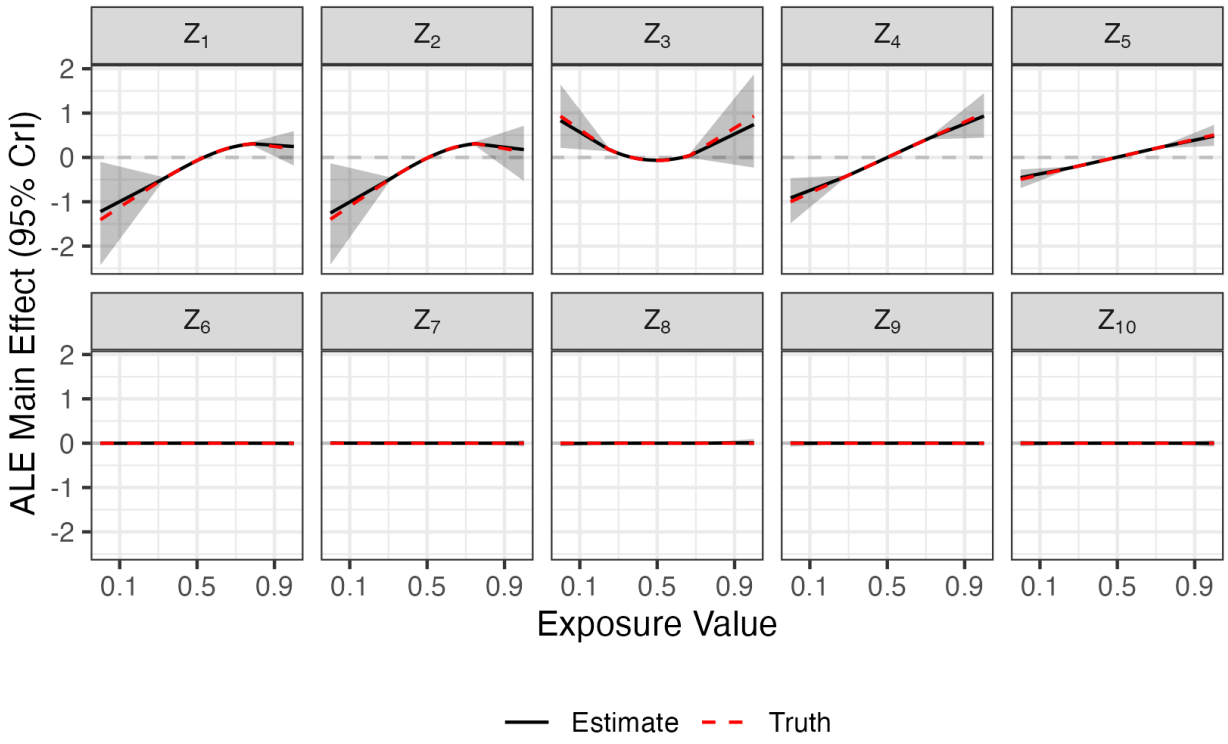

Figure 4.1: **ALE Main Effects for the Soft BART Simulation Study**: Depicted are the main effect ALEs from the simulation study with $T = 25$ soft, sparse trees setting. The solid black line represents the pointwise average ALE posterior mean, and the gray ribbon represents the average 95% posterior credible interval (CrI) bounds across 200 simulations. ALEs are calculated using $K = 40$ quantile intervals for each exposure.

In addition to the excellent performance on the recovery of f, estimates for $\boldsymbol{\gamma}$, ξ, τ^2, ρ, and $\boldsymbol{\nu}$ were also generally unbiased and exhibited reasonable 95% credible interval coverage (see Table 4.2 and Figures 4.6, 4.7, and 4.8 in the Chapter 4 Supplementary Materials).

4.5 Application: Asthma and Air Pollution in Atlanta, Georgia

4.5.1 Descriptive Statistics

We observed 478,311 asthma-related ED visits from 219,136 daily counts during the warm season in Atlanta from 2011-2018. These visits came from 128 ZIP codes from Clayton, DeKalb, Gwinnett, Fulton, and Cobb counties. The number of asthma-related ED visits was relatively stable year-over-year during this time frame, but in general more visits are observed at either end of the warm season (April and October), and occasionally coincide with federal holidays as well (see Figure 4.12 in the Chapter 4 Supplementary Materials). To account for potential confounding by these factors, we included an indicator variable representing federal holidays and a natural cubic spline on the day-of-year with 7 degrees of freedom per year (one per each warm season month). Given the previous findings of O'Lenick et al. [86] suggesting the importance of socioeconomic status in this same dataset, we also include the annual ZIP code-level percent below the poverty threshold as a time-varying linear confounder.

4.5.2 Model Considerations

We consider four chemical exposures $PM_{2.5}$, NO_2, O_3, and CO, as well as a meteorological exposure in maximum temperature. Each of these are recorded daily and included as 3-day

moving averages. We fit soft BART ensembles of 10, 25, 50, and 100 trees for each of the five primary exposures individually and as a mixture. We run each model for 5,000 burn-in iterations, and then draw 1,000 posterior samples using a thinning interval of 10 iterations (15,000 total MCMC iterations).

4.5.3 Results

For each model, we compute the Widely Applicable Information Criterion (WAIC) as an approximation to leave-one-out cross-validation [108, 40]. The results are plotted in Figure 4.2. For the single-exposure models, the WAIC is lowest for NO_2 and CO, suggesting that these two exposures are the most predictive of asthma-related ED visits when considered individually. As suspected, the mixture model containing all five exposures had a much lower WAIC than any of the single-exposure models. While increasing the ensemble size beyond 25 trees does not appear to improve the WAIC for any of the single-exposure models, larger ensembles may lead to some improved performance of the mixture model in terms of WAIC. However, when summarizing results of the mixture models with larger ensembles, we found the main findings to be generally similar to the fit with 25 trees. Due to this finding and for the sake of an even comparison, we will consider only the 25-tree single-exposure and mixture models in this section.

A popular approach for assessing the overall mixture effect is to evaluate the exposure-risk surface at a range of exposure values. For instance, one might plot the fitted exposure-risk function while simultaneously setting all exposures to specific quantiles (see Figure 4.3). Using this strategy, the overall mixture effect suggests a decreasing risk of asthma-related ED visits with increased exposure levels. One challenge with this approach is that it is difficult to assess the contribution of each individual exposure to the overall mixture effect. A larger

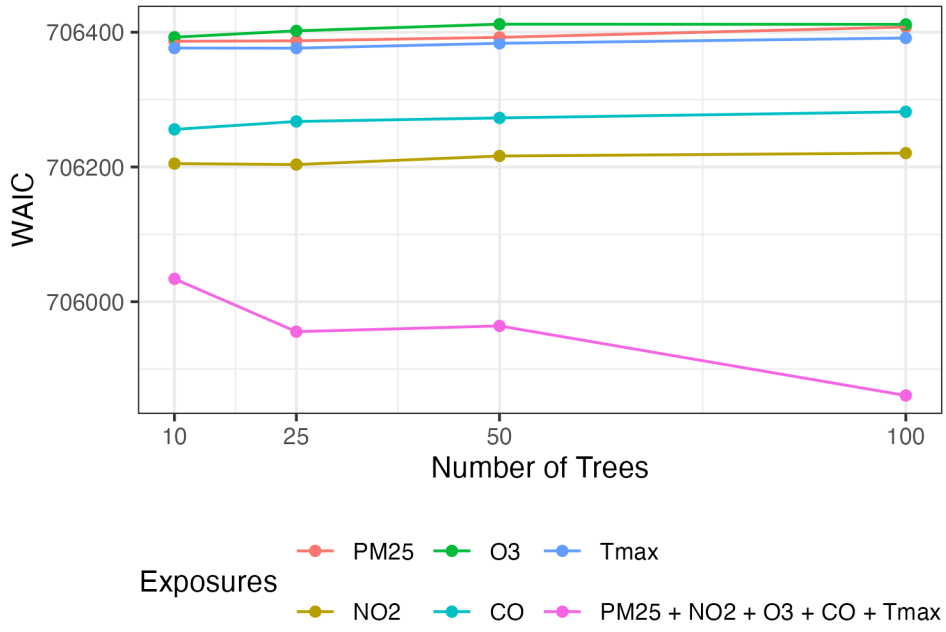

Figure 4.2: **WAIC for Single-Exposure and Mixture Models for the Asthma Application**: Results for all 25-tree Soft BART models fit to the asthma-related emergency department visit data.

issue is that these exposure profiles are not particularly realistic. The observed pairwise proportions of exposures across all ZIP code days belonging to the same decile range from just 10% (NO_2 and temperature, CO and temperature) to 23% (NO_2 and CO). Meanwhile, just 0.06% of all ZIP code days had all five exposures in the same decile. This observation underlines the need to evaluate the exposure-risk function in a more realistic manner.

Alternatively, we can avoid extrapolation in our assessment of the mixture effect by referencing the ALE. Estimates of each exposure's ALE shift slightly in the mixture model compared to their single-exposure models (Figure 4.4). Most notably, the largely null effect of O_3 shifts to harmful in the mixture model. Additionally, in the mixture model, $PM_{2.5}$ has a borderline harmful main effect, NO_2 has a strong negative association with ED visits, and CO and temperature have some upside-down "U-shaped" relationship with ED visits (Figure 4.4).

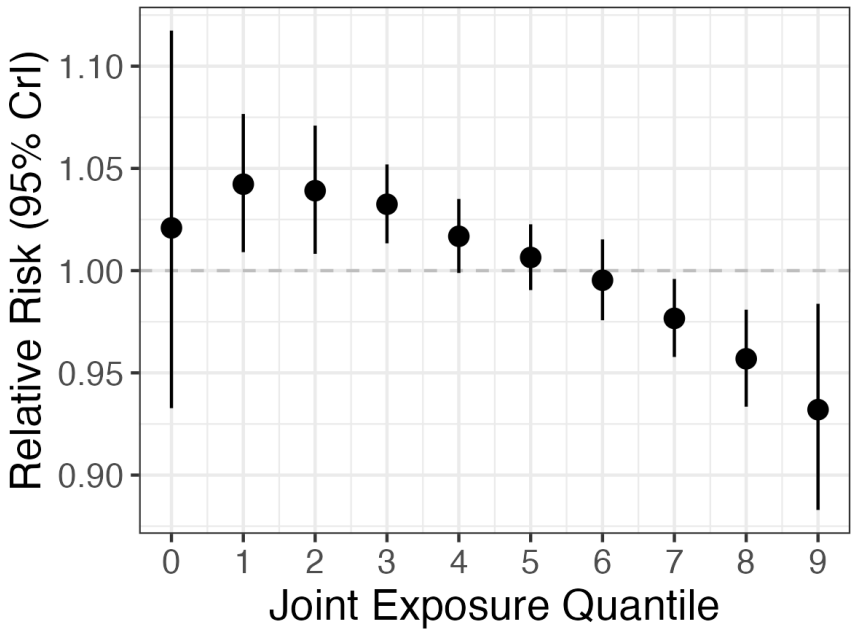

Figure 4.3: **Estimated Air Pollution Mixture Effect for the Asthma Application**: Posterior means and 95% posterior credible interval (CrI) for the estimated relative risk of an asthma-related emergency department visits when all exposures are simultaneously set to the same decile. Estimates are relative to the exposure profiles corresponding to the overall average risk.

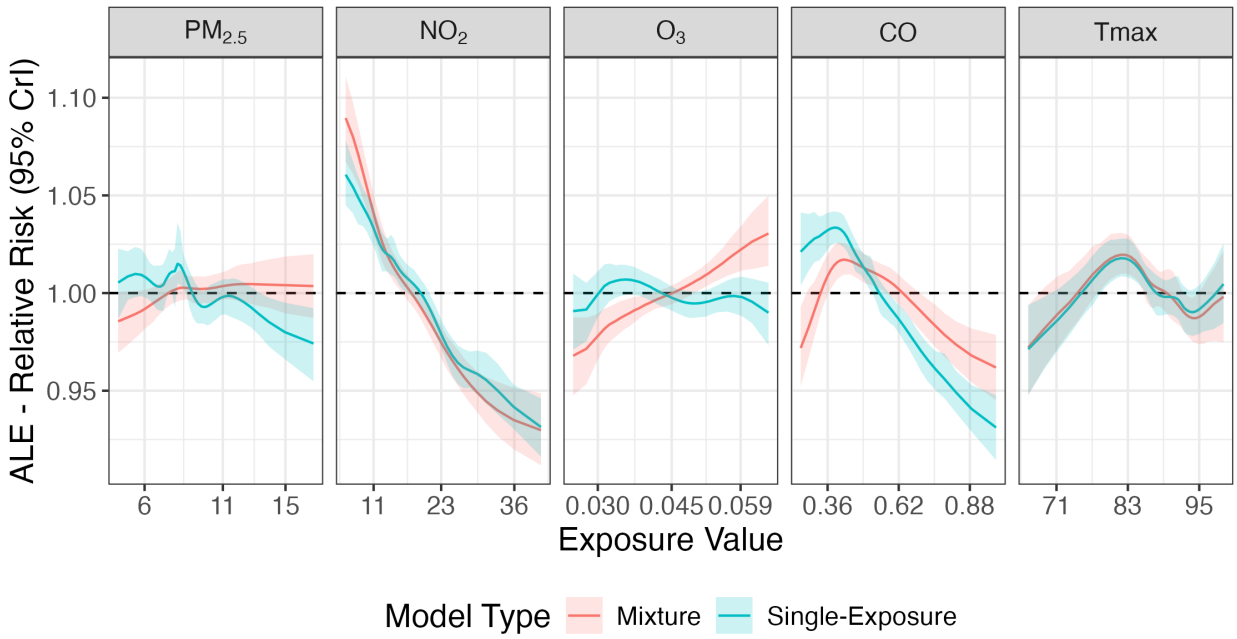

Figure 4.4: **Main Effect ALE for Single-Exposure and Mixture Models for the Asthma Application**: Estimated main effect ALE for single-exposure models (blue) and for each exposure in the mixture model (red). Plots are centered so that one on the y-axis represents an average risk level. ALEs are calculated using $K = 40$ quantile intervals for each exposure. Plots are trimmed so that only the central 95% of each exposure is displayed.

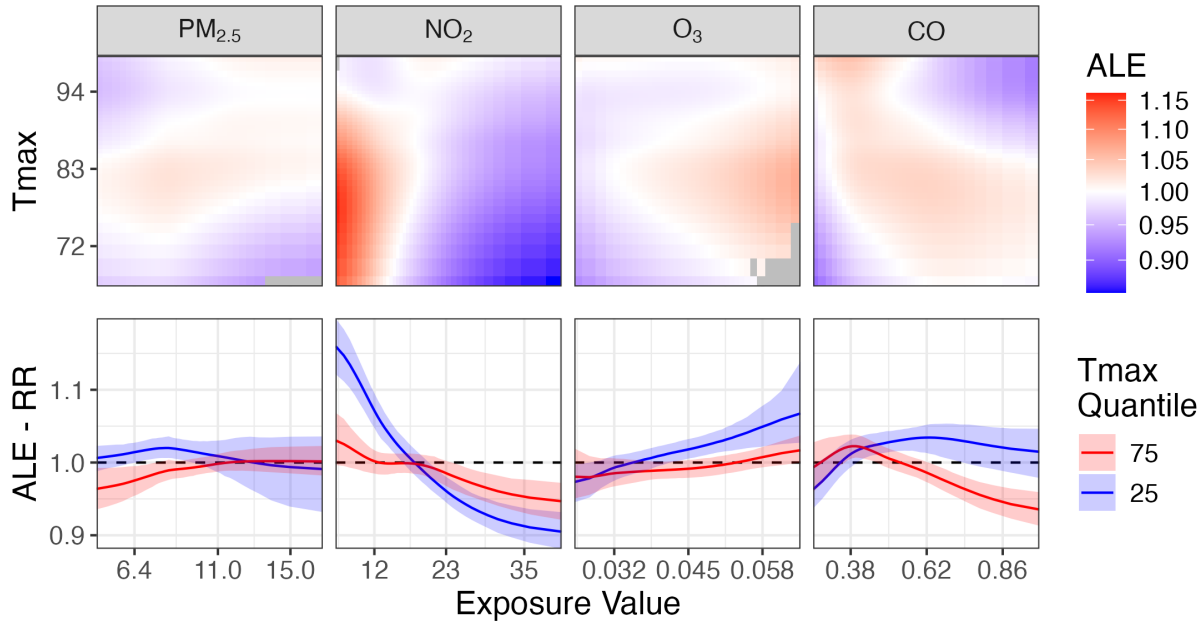

Figure 4.5: **Mixture Model Pairwise ALEs for the Asthma Application**: Depicted are the pairwise second-order ALE for each chemical exposure with maximum temperature for the mixture model with $T = 25$, with the corresponding main effect ALEs added on. The top row includes the posterior mean ALE for all observed pairwise combinations, while the bottom row plots horizontal slices at the first and third quartiles of maximum temperature along with 95% credible intervals. ALEs are calculated using $K = 40$ quantile intervals for each exposure. Plots are trimmed so that only the central 95% of each exposure is displayed.

In the mixture model, the exposures may interact with one another as well. Here we focus on the potential joint effects of each chemical exposure with temperature, but the resulting fit also showed some interaction among the chemical exposures (see Figure 4.15 in the Chapter 4 Supplementary Materials). In Figure 4.5, we note that the estimated ALE for each pollutant depends on temperature to some extent. For instance, the negative association between NO_2 and ED visits is more pronounced at lower temperatures. The positive association between O_3 and ED visits is more pronounced at lower temperatures, unless the O_3 concentration is very high. We also note that the estimated risk associated with $PM_{2.5}$ only appears to differ with temperature for lower $PM_{2.5}$ concentrations, while the reverse is true for CO.

4.6 Discussion

In summary, we show through a simulation study and real data application that BART (specifically soft BART) can be used for estimating complex mixture exposure-risk surfaces in the context of count responses, such as visits to an emergency department. Additionally, we have demonstrated the utility of ALE for analyzing marginal and joint effects of 1-2 exposures in the presence of other exposures. Plots such as those included in Figures 4.4 and 4.5 are straightforward to interpret individually since the ALE estimation process only averages over plausible exposure profiles supported by the data.

Our application findings regarding asthma-related emergency department visits are also interesting. The exposure concentration required to achieve above average risk may depend on the temperature. We observed interaction between ozone and temperature - specifically a stronger ozone effect at cooler warm-season temperatures. While the analysis framework we have proposed is not causal in nature, we hypothesize that modification of chemical exposure effects by temperature could be related to individual-level behavior - e.g., people may be less likely to experience the effects of air pollution on very hot days where they are more inclined to stay indoors. We also found a strong negative association between NO_2 and ED visits, which stands in contrast to some findings regarding NO_2 and respiratory outcomes. Due to the dense tree canopy and high traffic emissions in Atlanta, NO_2 and volatile organic compounds are a precursor to ozone, and higher NO_2 levels may be reflective of warmer days with lower ozone pollution.

One current limitation of our methodology is the ability to formally detect lagged effects. Since in our application we are focused on short-term effects, using 3-day moving averages for the daily exposures is sufficient. In general it is difficult to simultaneously estimate nonlinear,

interaction, and lagged effects in mixture modeling. Regardless of the specific model selected, we find it important to consider exposures as a mixture rather than individually both when it comes to estimating the exposure-response surface and interpreting the resulting fit.

Our modeling approach is also computationally demanding. Drawing Pólya-gamma latent variables for every observation in the data augmentation step and updating the soft BART ensemble takes time. The model shared in the application took approximately 20 hours to fit and summarize. While this is a long time, Gaussian process based modeling approaches (e.g. BKMR) would be infeasible given the sample size of 220,000. A BART ensemble using traditional "rigid" trees would also fit faster than soft BART, but may also require a greater number of trees to achieve comparable performance.

4.7 Supplementary Materials

4.7.1 Soft BART Negative Binomial Algorithm Details

The Bayesian model described in Section 4.3 is fit with a Markov chain Monte Carlo algorithm. Additional details for certain step in Algorithm 4.5 are described here.

Updating $\boldsymbol{\gamma}$

Let $r_{ij}^{\gamma} = y_{ij}^{*} - \log\left(\text{Pop}_{ij}\right) - \sum_{t=1}^{T} \text{Tree}(\boldsymbol{x}_{ij}; \mathcal{T}_t, \mathcal{M}_t) - \nu_i$, and define $\boldsymbol{r}^{\gamma} = (r_{11}^{\gamma}, \ldots, r_{IJ}^{\gamma})^T$. Then, assuming $\pi(\boldsymbol{\gamma}) = \text{MVN}(\boldsymbol{\gamma} \mid \mathbf{b}, \boldsymbol{\Sigma}_{\gamma})$, we derive the posterior distribution for $\boldsymbol{\gamma}$ as:

$$
\begin{aligned}
\pi(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{\eta}, \boldsymbol{\omega}, \xi) &\propto \pi(\boldsymbol{\gamma}) \times p(\boldsymbol{r}^{\gamma} \mid \boldsymbol{\gamma}, \boldsymbol{\Omega}) \\
&\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\gamma} - \mathbf{b}\right\}^T \boldsymbol{\Sigma}_{\gamma}^{-1} \left(\boldsymbol{\gamma} - \mathbf{b}\right)\right\} \times \exp\left\{-\frac{1}{2}\left(\mathbf{W}\boldsymbol{\gamma} - \boldsymbol{r}^{\gamma}\right)^T \boldsymbol{\Omega} \left(\mathbf{W}\boldsymbol{\gamma} - \boldsymbol{r}^{\gamma}\right)\right\} \\
&\propto \exp\left[-\frac{1}{2}\left\{(\boldsymbol{\gamma} - \mathbf{b})^T \boldsymbol{\Sigma}_{\gamma}^{-1} (\boldsymbol{\gamma} - \mathbf{b}) + \left(\mathbf{W}\boldsymbol{\gamma} - \boldsymbol{r}^{\gamma}\right)^T \boldsymbol{\Omega} \left(\mathbf{W}\boldsymbol{\gamma} - \boldsymbol{r}^{\gamma}\right)\right\}\right] \\
&\propto \text{Normal}\left(\boldsymbol{\mu}_{\gamma}^{*}, \boldsymbol{\Sigma}_{\gamma}^{*}\right)
\end{aligned}
\tag{4.19}
$$

where $\boldsymbol{\Sigma}_{\gamma}^{*} = \left(\boldsymbol{\Sigma}_{\gamma}^{-1} + \mathbf{W}^T \boldsymbol{\Omega} \mathbf{W}\right)^{-1}$ and $\boldsymbol{\mu}_{\gamma}^{*} = \boldsymbol{\Sigma}_{\gamma}^{*}\left(\boldsymbol{\Sigma}_{\gamma}^{-1}\mathbf{b} + \mathbf{W}^T \boldsymbol{\Omega} \boldsymbol{r}^{\gamma}\right)$.

Updating $\boldsymbol{\nu}$

Let $r_{ij}^{\nu} = y_{ij}^{*} - \log\left(\text{Pop}_{ij}\right) - \boldsymbol{w}_{ij}^T \boldsymbol{\gamma} - \sum_{t=1}^{T} \text{Tree}(\boldsymbol{x}_{ij}; \mathcal{T}_t, \mathcal{M}_t)$, and define $\boldsymbol{r}^{\nu} = (r_{11}^{\nu}, \ldots, r_{IJ}^{\nu})^T$. Recall that we assume a proper CAR prior for $\boldsymbol{\nu}$, i.e. $\boldsymbol{\nu} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{\nu})$ where $\boldsymbol{\Sigma}_{\nu} = \tau^2(\mathbf{D} - \rho\mathbf{A})^{-1}$. $\mathbf{A}$ is the $I \times I$ first-order adjacency matrix, and $\mathbf{D}$ is the $I \times I$ diagonal matrix containing the number of neighbors for each region. Let $\mathbf{X}_{\nu}$ be the $IJ \times I$ design matrix such that $\mathbf{X}_{\nu}\boldsymbol{\nu}$ is the $IJ \times 1$ vector of spatial random effects allocated to each observation in the

dataset. We derive the posterior distribution for $\boldsymbol{\nu}$ as:

$$\begin{aligned}\pi(\boldsymbol{\nu} \mid \boldsymbol{y}, \boldsymbol{\eta}, \boldsymbol{\omega}, \xi) &\propto \pi(\boldsymbol{\nu}) \times p(\boldsymbol{r}^{\nu} \mid \boldsymbol{\nu}, \boldsymbol{\Omega}) \\ &\propto \exp\left(-\frac{1}{2}\boldsymbol{\nu}^T \boldsymbol{\Sigma}_{\nu}^{-1} \boldsymbol{\nu}\right) \times \exp\left\{-\frac{1}{2}\left(\mathbf{X}_{\nu}\boldsymbol{\nu} - \boldsymbol{r}^{\nu}\right)^T \boldsymbol{\Omega}\left(\mathbf{X}_{\nu}\boldsymbol{\nu} - \boldsymbol{r}^{\nu}\right)\right\} \\ &\propto \exp\left[-\frac{1}{2}\left\{\boldsymbol{\nu}^T \boldsymbol{\Sigma}_{\nu}^{-1} \boldsymbol{\nu} + \left(\mathbf{X}_{\nu}\boldsymbol{\nu} - \boldsymbol{r}^{\nu}\right)^T \boldsymbol{\Omega}\left(\mathbf{X}_{\nu}\boldsymbol{\nu} - \boldsymbol{r}^{\nu}\right)\right\}\right] \\ &\propto \text{Normal}\left(\boldsymbol{\mu}_{\nu}^*, \boldsymbol{\Sigma}_{\nu}^*\right)\end{aligned} \tag{4.20}$$

where $\boldsymbol{\Sigma}_{\nu}^* = \left(\boldsymbol{\Sigma}_{\nu}^{-1} + \mathbf{X}_{\nu}^T \boldsymbol{\Omega} \mathbf{X}_{\nu}\right)^{-1}$ and $\boldsymbol{\mu}_{\nu}^* = \boldsymbol{\Sigma}_{\nu}^* \mathbf{X}_{\nu}^T \boldsymbol{\Omega} \boldsymbol{r}^{\nu}$.

Updating τ^2

Assuming $\pi(\tau^2) = \text{Inverse-Gamma}\left(\tau^2 \mid \alpha_{\tau}, \beta_{\tau}\right)$, the posterior distribution of τ^2 is derived as:

$$\begin{aligned}\pi(\tau^2 \mid \boldsymbol{\nu}) &\propto \pi(\tau^2) \times p(\boldsymbol{\nu} \mid \tau^2) \\ &\propto \frac{{\beta_{\tau}}^{\alpha_{\tau}}}{\Gamma(\alpha_{\tau})}(\tau^2)^{-\alpha_{\tau}-1} \exp\left(-\frac{\beta_{\tau}}{\tau^2}\right) \times \frac{1}{\sqrt{\det\left(\boldsymbol{\Sigma}_{\nu}\right)}} \times \exp\left(-\frac{1}{2}\boldsymbol{\nu}^T \boldsymbol{\Sigma}_{\nu}^{-1} \boldsymbol{\nu}\right)\end{aligned} \tag{4.21}$$

If we assume $\boldsymbol{\Sigma}_{\nu} = \tau^2(\mathbf{D} - \rho\mathbf{A})^{-1}$, as is the case when using a proper CAR prior for $\boldsymbol{\nu}$, then we have:

$$\begin{aligned}\pi(\tau^2 \mid \boldsymbol{\nu}) &\propto \left(\tau^2\right)^{-\left(\alpha_{\tau} + \frac{I}{2}\right)-1} \exp\left[-\frac{1}{\tau^2}\left\{\beta_{\tau} + \frac{\boldsymbol{\nu}^T(\mathbf{D} - \rho\mathbf{A})\boldsymbol{\nu}}{2}\right\}\right] \\ &\propto \text{Inverse-Gamma}\left(\alpha_{\tau} + \frac{I}{2}, \quad \beta_{\tau} + \frac{\boldsymbol{\nu}^T(\mathbf{D} - \rho\mathbf{A})\boldsymbol{\nu}}{2}\right)\end{aligned} \tag{4.22}$$

where the above uses the result that for constant c and square matrix $\mathbf{H}_{N \times N}$, we have $\det\left(c\mathbf{H}\right) = c^N \det\left(\mathbf{H}\right)$.

Updating ρ

For updating ρ, we assign a discrete uniform prior over 1,000 values between 0 and 1 (i.e., $\rho \sim \text{Unif}\left\{\frac{0}{999}, \ldots, \frac{999}{999}\right\}$). Since this prior assigns equal probability to all values in its support, the full conditional distribution for ρ is proportional to the density of the random effects (4.23).

$$\pi\left(\rho \mid \boldsymbol{\nu}, \tau^2\right) \propto \pi\left(\rho\right) \times p\left(\boldsymbol{\nu} \mid \rho, \tau^2\right) \propto p\left(\boldsymbol{\nu} \mid \rho, \tau^2\right) \tag{4.23}$$

Recall that we assume a proper CAR prior for $\boldsymbol{\nu}$. In other words, $\boldsymbol{\nu} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_\nu)$, where $\boldsymbol{\Sigma}_\nu = \tau^2(\mathbf{D} - \rho\mathbf{A})^{-1}$. Thus, the log density of $p(\boldsymbol{\nu} \mid \tau^2, \rho)$ is given by

$$\begin{aligned}
\log \pi\left(\boldsymbol{\nu} \mid \tau^2, \rho\right) &\propto \frac{1}{2}\log|\tau^2(\mathbf{D} - \rho\mathbf{A})^{-1}| - \frac{1}{2\tau^2}\boldsymbol{\nu}^T(\mathbf{D} - \rho\mathbf{A})\boldsymbol{\nu} \\
&\propto \frac{1}{2}\log|\mathbf{D} - \rho\mathbf{A}| + \frac{\rho}{2\tau^2}\boldsymbol{\nu}^T\mathbf{A}\boldsymbol{\nu} \\
&= \frac{1}{2}\log|\mathbf{D}(\mathbf{I} - \rho\mathbf{D}^{-1}\mathbf{A})| + \frac{\rho}{2\tau^2}\boldsymbol{\nu}^T\mathbf{A}\boldsymbol{\nu} \\
&\propto \frac{1}{2}\log|\mathbf{I} - \rho\mathbf{D}^{-1}\mathbf{A}| + \frac{\rho}{2\tau^2}\boldsymbol{\nu}^T\mathbf{A}\boldsymbol{\nu} \\
&= \frac{1}{2}\sum_{i=1}^{I}\log(1 - \rho\lambda_i) + \frac{\rho}{2\tau^2}\boldsymbol{\nu}^T\mathbf{A}\boldsymbol{\nu}
\end{aligned} \tag{4.24}$$

where λ_i is the i^{th} eigenvalue of $\mathbf{D}^{-1}\mathbf{A}$. The first term in the final expression may be calculated ahead of time for all candidate values of ρ so that only the second term needs to be updated during the MCMC. New values of ρ can be sampled from the discrete distribution with probabilities proportional to the computed values of (4.24) for all 1,000 prior values of ρ. Alternatively, an intrinsic CAR prior may be used, where ρ if fixed to 1 throughout the MCMC algorithm. In many applications the posterior distribution of ρ will be concentrated around 1, but using the approach detailed here will provide more flexibility when that is not the case.

Updating ξ

The final update in the MCMC algorithm is for the (over)dispersion parameter ξ. We perform this update using the conjugate sampling routine described in [121]. This technique relies on expressing the $Y_{ij} \sim \text{NB}(\xi, p_{ij})$ marginal distribution as a compound Poisson distribution (4.25).

$$\begin{aligned} y_{ij} &= \sum_{m=1}^{L_{ij}} l_m \\ L_{ij} &\sim \text{Poisson}(-\xi \ln(1-p_{ij})) \end{aligned} \tag{4.25}$$

Given y_{ij} and ξ, L_{ij} follows a Chinese Restaurant Table (CRT) distribution whose samples can be generated as $L_{ij} = \sum_{m=1}^{y_{ij}} l_m$ where $l_m \sim \text{Bernoulli}\left(\frac{\xi}{\xi+m-1}\right)$. If we assign a $\text{Gamma}(\alpha_\xi, \beta_\xi)$ prior for ξ, the full conditional distribution for ξ is derived as:

$$\begin{aligned} \pi(\xi \mid p) &\propto \pi(\xi) \prod_{i=1}^{I} \prod_{j=1}^{J} p(L_{ij} \mid \xi, p_{ij}) \\ &= \frac{\beta_\xi^{\alpha_\xi}}{\Gamma(\alpha_\xi)} \xi^{\alpha_\xi - 1} \exp\left(-\xi \beta_\xi\right) \times \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{\{-\xi \ln(1-p_{ij})\}^{L_{ij}} \exp\left\{\xi \ln(1-p_{ij})\right\}}{L_{ij}!} \\ &\propto \xi^{\alpha_\xi - 1} \exp\left(-\xi \beta_\xi\right) \times \xi^{\sum\sum L_{ij}} \exp\left\{\xi \sum_{i=1}^{I} \sum_{j=1}^{J} \ln(1-p_{ij})\right\} \\ &= \xi^{\alpha_\xi + \sum\sum L_{ij} - 1} \exp\left[-\xi \left\{\beta_\xi - \sum_{i=1}^{I} \sum_{j=1}^{J} \ln(1-p_{ij})\right\}\right] \\ &\propto \text{Gamma}\left(\alpha_\xi + \sum_{i=1}^{I} \sum_{j=1}^{J} L_{ij}, \beta_\xi - \sum_{i=1}^{I} \sum_{j=1}^{J} \ln(1-p_{ij})\right). \end{aligned} \tag{4.26}$$

See Zhou et al. [121] for further details regarding this procedure.

4.7.2 Implementation Details

We fit Soft BART models using the drop-in C++ module from the `SoftBart` R package. This module allows you to customize your own Gibbs samplers using BART. Here we have supplied an example of an R function one might use to fit the models described in this chapter.

```
# inputs:
#   - w: IJ x P_w matrix of confounders
#   - y: IJ x 1 vectors of counts
#   - x: IJ x P_x matrix of exposures
#   - offset: IJ x 1 vector of population offsets
#   - x_nu: spatial random effects design matrix
#   - A: spatial adjaceny matrix
#   - num_tree: number of trees for BART ensemble
run_mcmc <- function (w, y, x, offset, x_nu, A,
                      num_tree = 20, num_burn = 5000, num_save = 1000) {

  # Sample sizes and dimensions
  n   <- length(y); ns  <- nrow(A); p_w <- ncol(w); p_x <- ncol(x)
  D <- diag(rowSums(A))

  # Initialize parameters for MCMC sampler
  G     <- numeric(n)           # BART predictor
  gamma  <- numeric(p_w)        # Confounder regression coefficients
  xi    <- 1                    # Dispersion parameter
  nu    <- rnorm(ns)     # Spatial random effects
  tau2  <- 1 / rgamma(1, 0.1 + ns / 2, 0.1 + (1 / 2))
  rho   <- 0.9

  # Linear predictor
  fixeff <- as.numeric(w %*% gamma)
  raneff <- as.numeric(x_nu %*% nu)
  eta <- offset + fixeff + G + raneff

  # Pre-calculate discrete prior distribution for rho
  lambda    <- eigen(solve(D) %*% A, only.values = TRUE)$values
  rho_vals  <- q(seq(1e-4, 1-1e-4, length.out = 1000), 1, 1)
  rho_ll0   <- sapply(rho_vals, \(x) 0.5 * sum(log(1 - x * lambda)),
     simplify = TRUE)

  # Specify fixed effect prior distribution
  b <- rep(0, p_w); B <- diag(p_w) * 1e4; B_inv <- diag(1 / diag(B))

  # Create BART objects
  bart_hypers <- SoftBart::Hypers(X = x, Y = G, sigma_hat = 1,
                                  num_tree = num_tree)
  bart_opts <- SoftBart::Opts(update_sigma = FALSE, num_burn = num_burn,
     num_save = num_save)

  sampler <- SoftBart::MakeForest(hypers = bart_hypers, opts = bart_opts)
```

```
  # Run sampler
  for (k in seq_len(num_burn + num_save)) {

    # Step 1) Sample latent Polya-Gamma random variables
    omega <- jrpg::jrpg(y + xi, eta)[,1]

    # Convert to Gaussian form
    y_star <- (y - xi) / (2 * omega)  # y_star ~ N(eta, diag(1 / omega))

    # Update spatial weight matrices
    D_rho_A <- spam::as.spam(D - rho * A)

    # Step 2) Update spatial random effects
    r_nu <- y_star - offset - fixeff - G
    nu_Sigma <- spam::solve(spam::crossprod.spam(x_nu * sqrt(omega)) + (1
       / tau2) * (D_rho_A))
    nu_mu <- nu_Sigma %*% spam::crossprod.spam(x_nu, omega * r_nu)
    nu <- mvtnorm::rmvnorm(n = 1, mean = nu_mu, sigma = nu_Sigma)[1,]
    nu <- nu - mean(nu)
    raneff <- as.numeric(x_nu %*% nu)

    # Step 3) Update spatial random effects variance
    tau2 <- 1 / rgamma(1, 0.1 + ns / 2, 0.1 + (nu %*% (D_rho_A) %*% nu)/2)

    # Step 4) Update spatial random effects correlation
    rho_ll <- rho_ll0 + rho_vals / (2*tau2) * as.numeric(nu %*% A %*% nu)
    rho <- sample(rho_vals, size = 1, prob = exp(rho_ll - max(rho_ll)))

    # Step 5) Update fixed effects
    r_gamma <- y_star - offset - G - raneff
    gamma_Sigma <- solve(B_inv + crossprod(w * sqrt(omega)))
    gamma_mu <- gamma_Sigma %*% (B_inv %*% b + crossprod(w, omega * r_
       gamma))
    gamma <- mvtnorm::rmvnorm(n = 1, mean = gamma_mu, sigma = gamma_Sigma)
       [1,]
    fixeff <- as.numeric(w %*% gamma)

    # Step 6) Update BART
    r_G <- y_star - offset - fixeff - raneff
    G <- sampler$do_gibbs_weighted(x, r_G, omega, x, 1)[1,]

    # Update linear predictor
    eta <- offset + fixeff + G + raneff

    # Update dispersion parameter
    l <- sapply(1:n, function (i) sum(rbinom(y[i], 1, round(xi / (xi + 1:y
       [i] - 1), 6))))
    q <- pmin(0.9999, 1 / (1 + exp(eta))) # 1 - Pr(success)
    xi <- rgamma(1, 0.01 + sum(l), 0.01 - sum(log(q)))
  }
}
```

4.7.3 Additional Simulation Materials

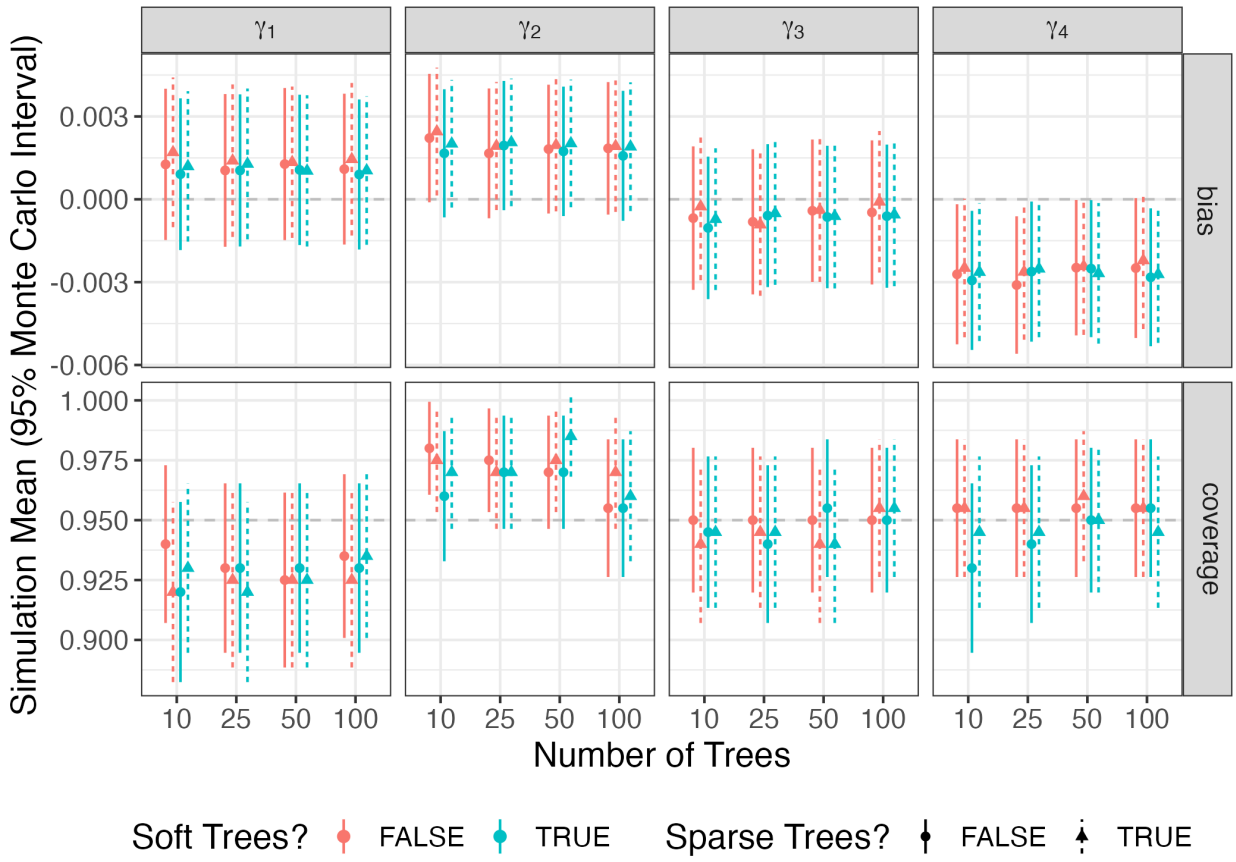

Figure 4.6: **Soft BART Simulation Results - Confounder Estimates**: Simulation average bias and 95% credible interval coverage for the four linearly modeled parameters. Estimates are plotted as the simulation mean +/- 1.96× the simulation Monte Carlo standard error.

Table 4.2: Soft BART Simulation Results - Global Parameter Estimates

			ρ		τ^2		ξ	
T[a]	Soft[b]	Sparse[c]	Bias (MCSE)	Coverage (MCSE)	Bias (MCSE)	Coverage (MCSE)	Bias (MCSE)	Coverage (MCSE)
10			-0.099 (0.008)	0.905 (0.021)	0.007 (0.003)	0.965 (0.013)	-0.020 (0.000)	0.205 (0.029)
10		✓	-0.098 (0.008)	0.910 (0.020)	0.007 (0.003)	0.970 (0.012)	-0.022 (0.000)	0.190 (0.028)
10	✓		-0.098 (0.008)	0.910 (0.020)	0.007 (0.003)	0.970 (0.012)	-0.004 (0.000)	0.930 (0.018)
10	✓	✓	-0.099 (0.008)	0.905 (0.021)	0.007 (0.003)	0.965 (0.013)	-0.004 (0.000)	0.960 (0.014)
25			-0.099 (0.008)	0.910 (0.020)	0.007 (0.003)	0.965 (0.013)	-0.009 (0.000)	0.775 (0.030)
25		✓	-0.098 (0.008)	0.905 (0.021)	0.007 (0.003)	0.975 (0.011)	-0.009 (0.000)	0.775 (0.030)
25	✓		-0.098 (0.008)	0.900 (0.021)	0.007 (0.003)	0.975 (0.011)	-0.003 (0.000)	0.945 (0.016)
25	✓	✓	-0.099 (0.008)	0.910 (0.020)	0.007 (0.003)	0.980 (0.010)	-0.003 (0.000)	0.955 (0.015)
50			-0.098 (0.008)	0.915 (0.020)	0.007 (0.003)	0.970 (0.012)	-0.005 (0.000)	0.930 (0.018)
50		✓	-0.098 (0.008)	0.910 (0.020)	0.006 (0.003)	0.970 (0.012)	-0.005 (0.000)	0.905 (0.021)
50	✓		-0.099 (0.008)	0.900 (0.021)	0.007 (0.003)	0.970 (0.012)	-0.003 (0.000)	0.950 (0.015)
50	✓	✓	-0.098 (0.008)	0.905 (0.021)	0.006 (0.003)	0.965 (0.013)	-0.003 (0.000)	0.950 (0.015)
100			-0.099 (0.008)	0.915 (0.020)	0.007 (0.003)	0.970 (0.012)	-0.003 (0.000)	0.960 (0.014)
100		✓	-0.099 (0.008)	0.895 (0.022)	0.006 (0.003)	0.975 (0.011)	-0.003 (0.000)	0.960 (0.014)
100	✓		-0.098 (0.008)	0.895 (0.022)	0.006 (0.003)	0.960 (0.014)	-0.003 (0.000)	0.975 (0.011)
100	✓	✓	-0.098 (0.008)	0.920 (0.019)	0.007 (0.003)	0.965 (0.013)	-0.003 (0.000)	0.970 (0.012)

MCSE: Monte Carlo Standard Error.

[a] T: Number of trees.

[b] Soft BART used [73].

[c] Sparse branching process used [70].

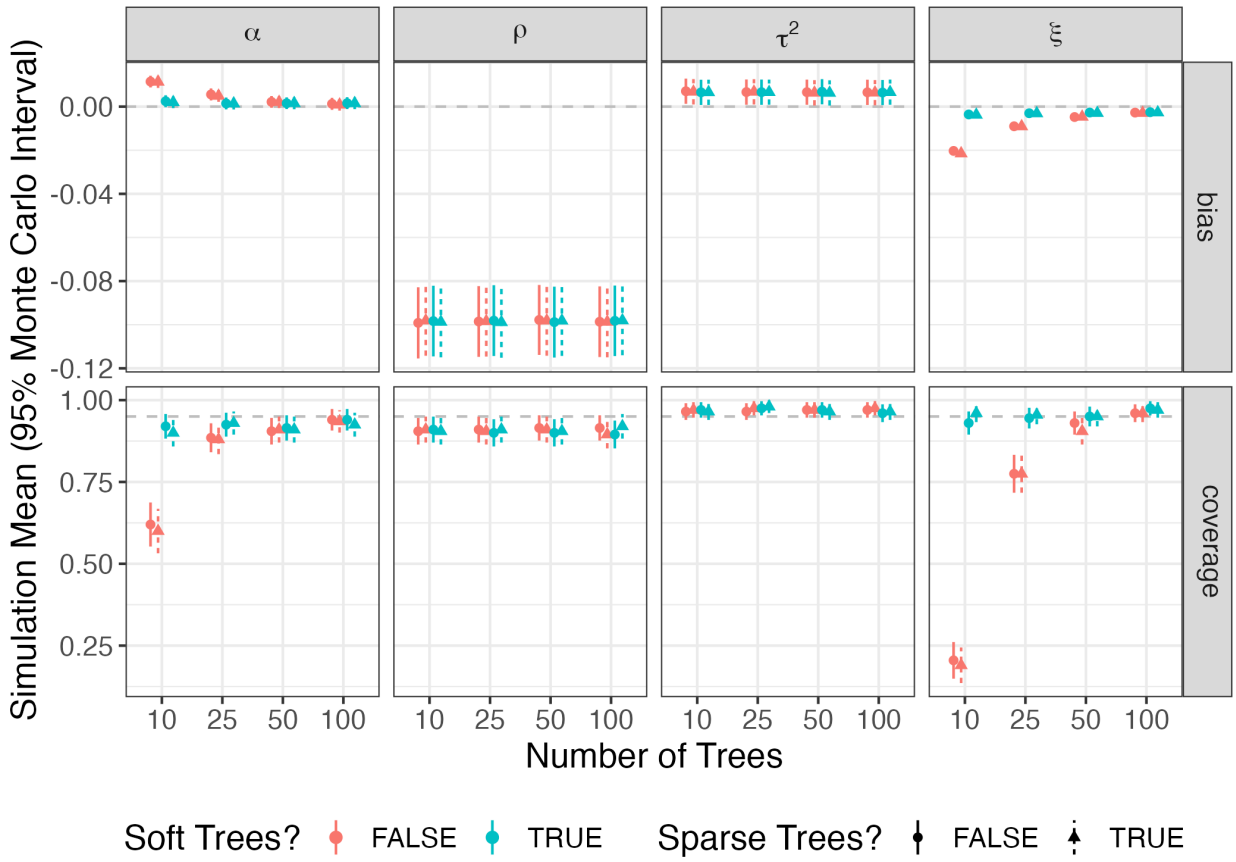

Figure 4.7: **Soft BART Simulation Results - Global Parameters Estimates**: Simulation average bias and 95% credible interval coverage for global parameters. Estimates are plotted as the simulation mean +/- 1.96× the simulation Monte Carlo standard error.

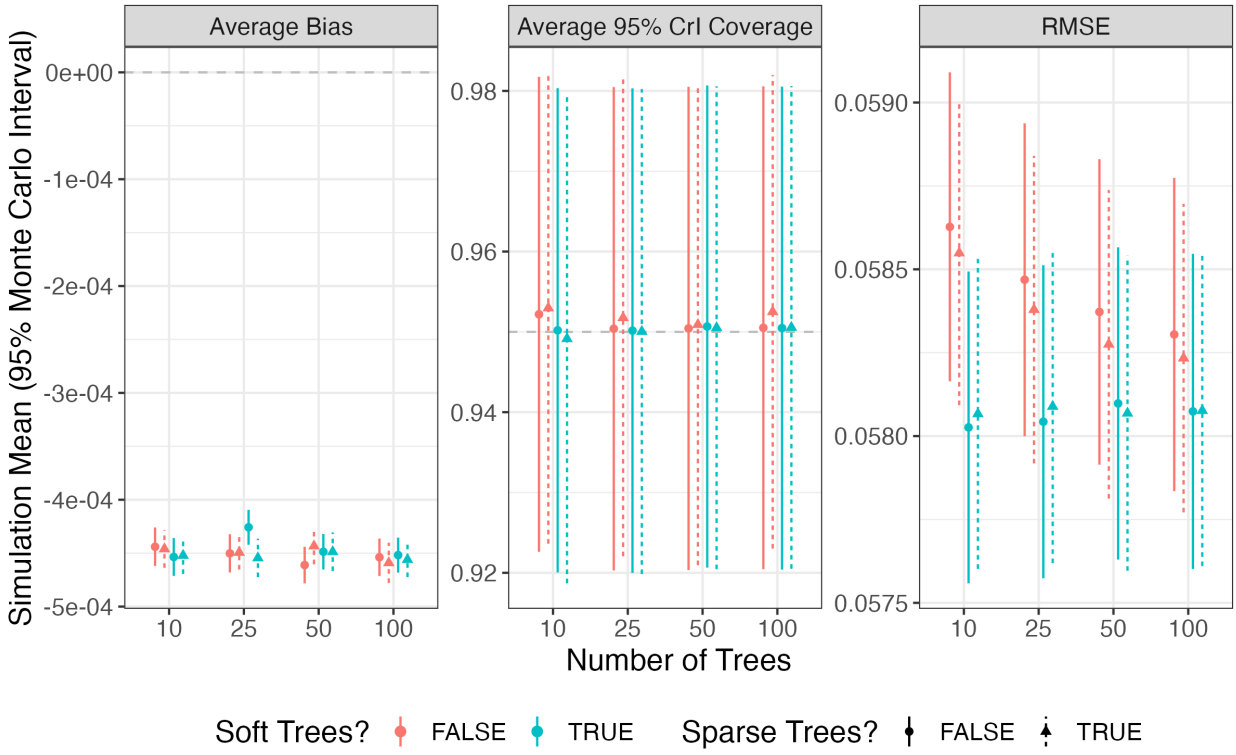

Figure 4.8: **Soft BART Simulation Results - Spatial Random Effects**: Simulation average bias, 95% credible interval coverage, and root mean squared error (RMSE) for spatial random effects. Estimates are plotted as the simulation mean +/- 1.96× the simulation Monte Carlo standard error.

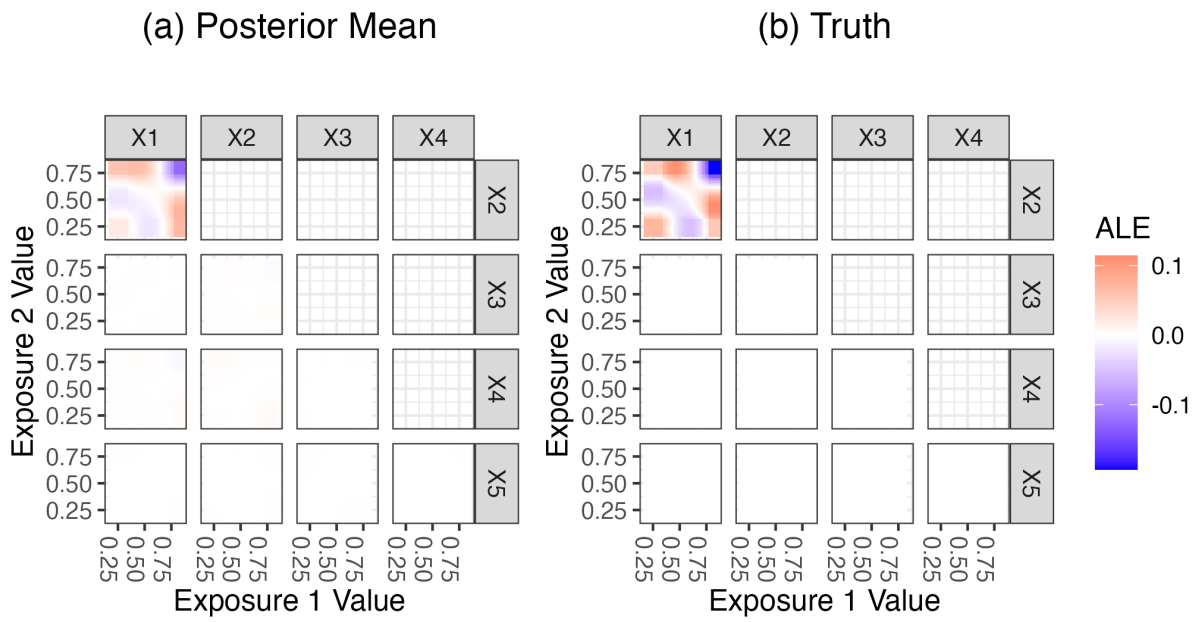

Figure 4.9: **Pairwise ALE Interaction Effects for the Soft BART Simulation Study**: Depicted are the ALE interaction (second-order only) effects from the simulation study with $T = 25$, soft, sparse trees. Plot (a) displays the simulation average posterior mean at each grid cell, while (b) displays the truth calculated per the data generating process based on Friedman [35] for comparison. ALEs are calculated using $K = 40$ quantile intervals for each exposures. Plots are trimmed so that only the central 95% of each exposure is displayed.

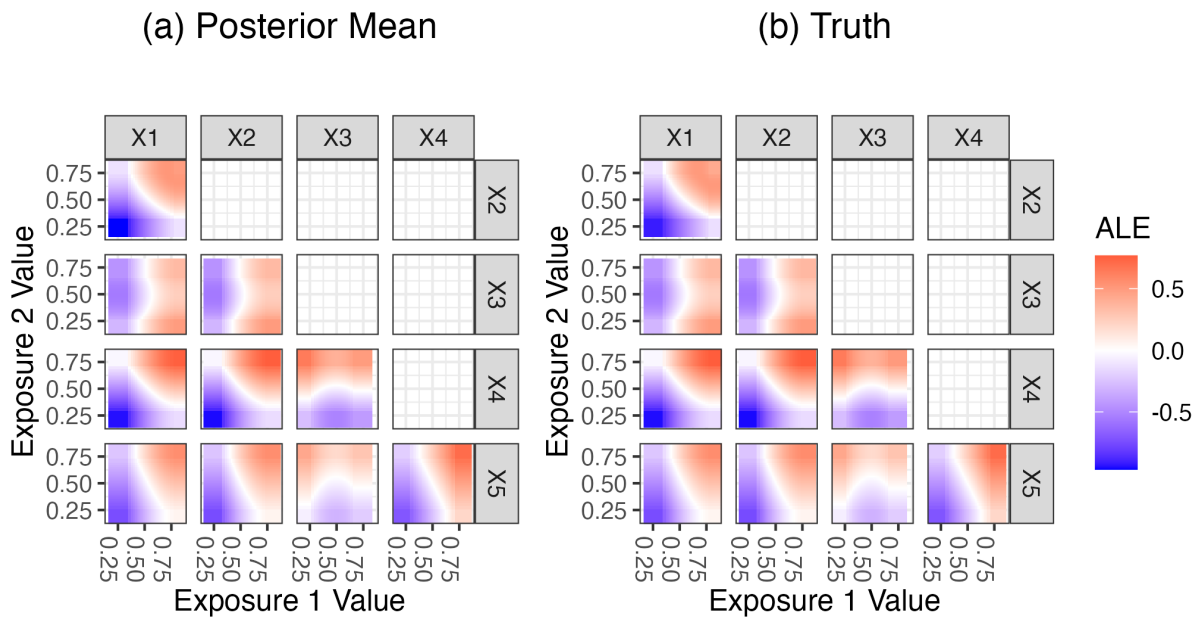

Figure 4.10: **Pairwise ALE Joint Effects for the Soft BART Simulation Study**: Depicted are the ALE pairwise joint effects (main effects + second-order effects) from the simulation study with $T = 25$, soft, sparse trees. Plot (a) displays the simulation average posterior mean at each grid cell, while (b) displays the truth calculated per the data generating process based on Friedman [35]. ALEs are calculated using $K = 40$ quantile intervals for each exposures. Plots are trimmed so that only the central 95% of each exposure is displayed.

4.7.4 Additional Application Materials

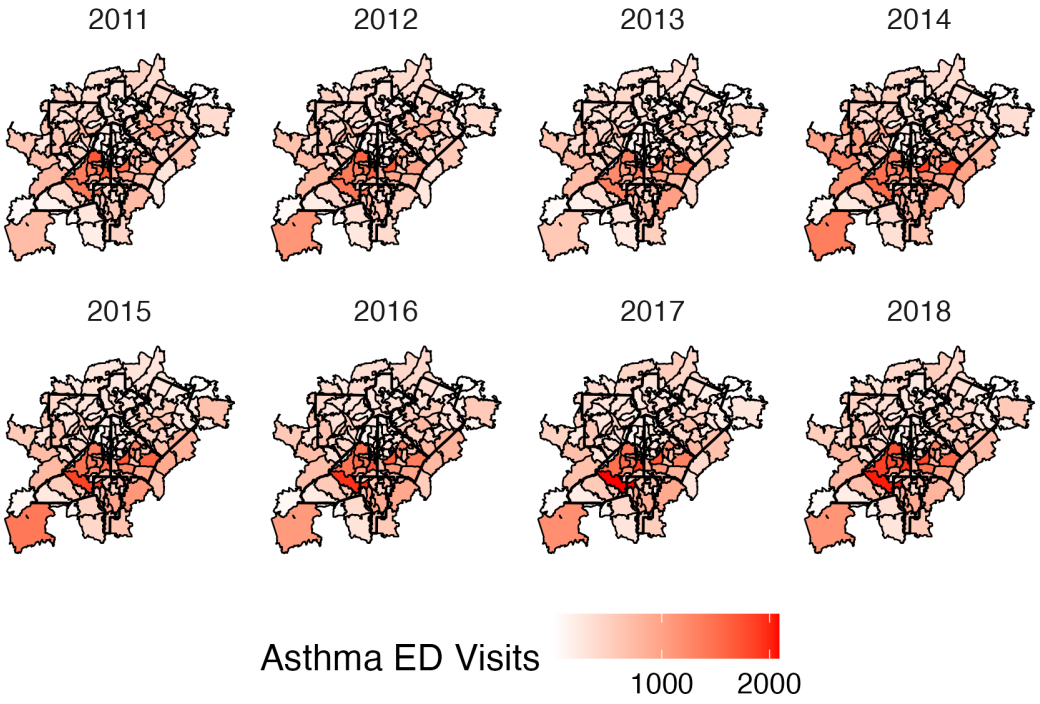

Figure 4.11: Annual ZIP Counts of Asthma-Related Emergency Department Visits in Metropolitan Atlanta by ZIP Code, 2011-2018 Warm Season

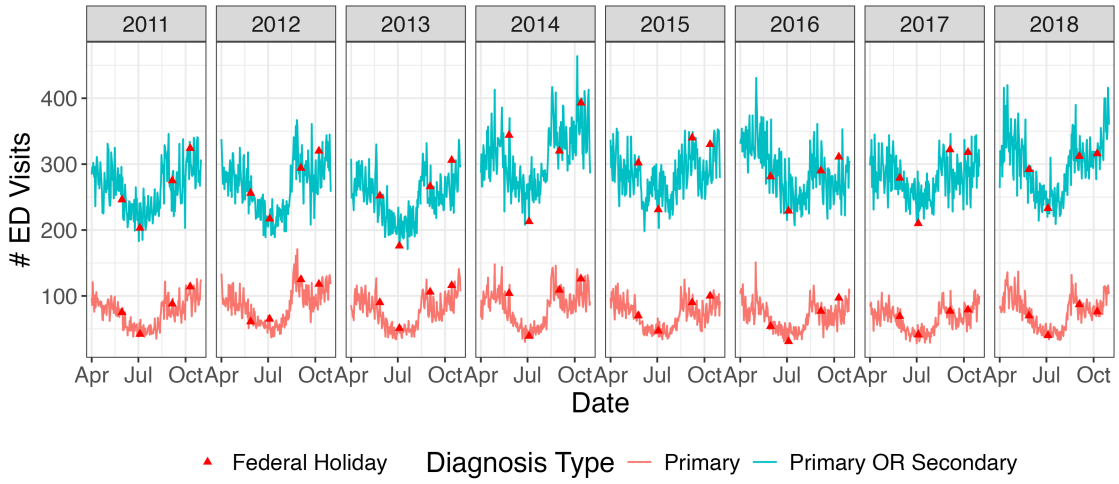

Figure 4.12: **Aggregated Daily Counts of Asthma-Related Emergency Department Visits in Metropolitan Atlanta, 2011-2018 Warm Season**: Counts of ED visits with asthma as the primary diagnosis are shown in red, while visits with any asthma diagnosis are shown in blue. All counts are aggregated over ZIP code.

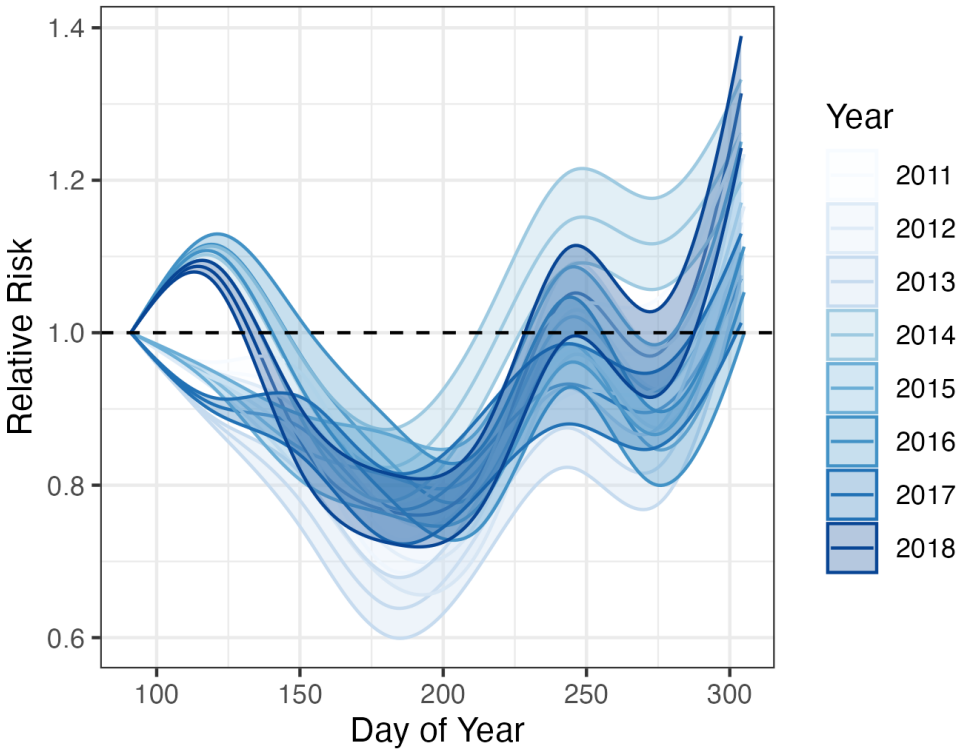

Figure 4.13: **Temporal Trends in Risk of Asthma-Related Emergency Department Visits in the Asthma Application**: Depicted is the estimated day-of-year spline (posterior mean +/- 95% credible interval) from the 25-tree mixture model fitted to the Atlanta asthma-related emergency department data.

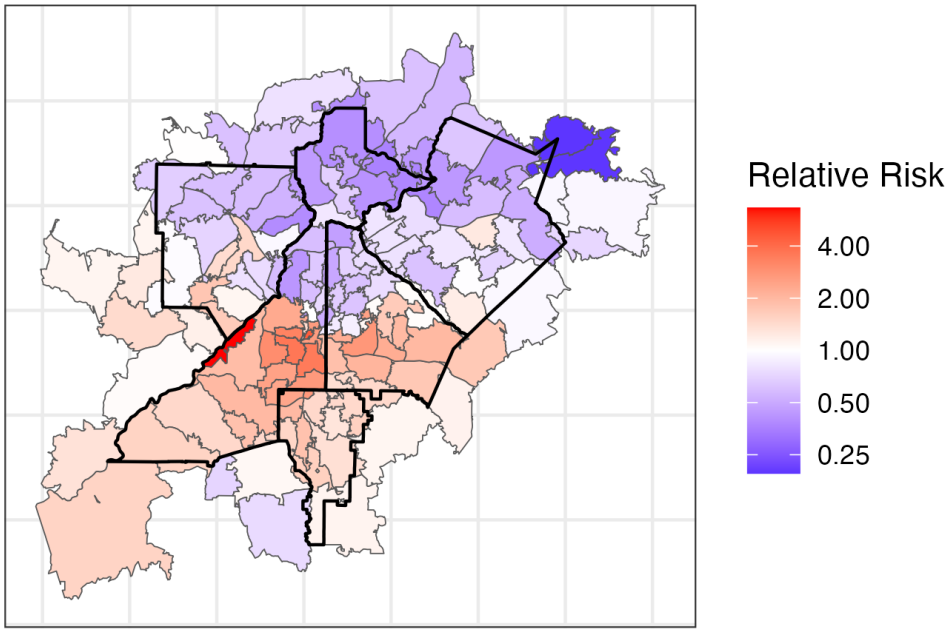

Figure 4.14: **Spatial Trends in Risk of Asthma-Related Emergency Department Visits in the Asthma Application**: Depicted are the posterior means of the spatial random effects from the 25-tree mixture model fitted to the Atlanta asthma-related emergency department data.

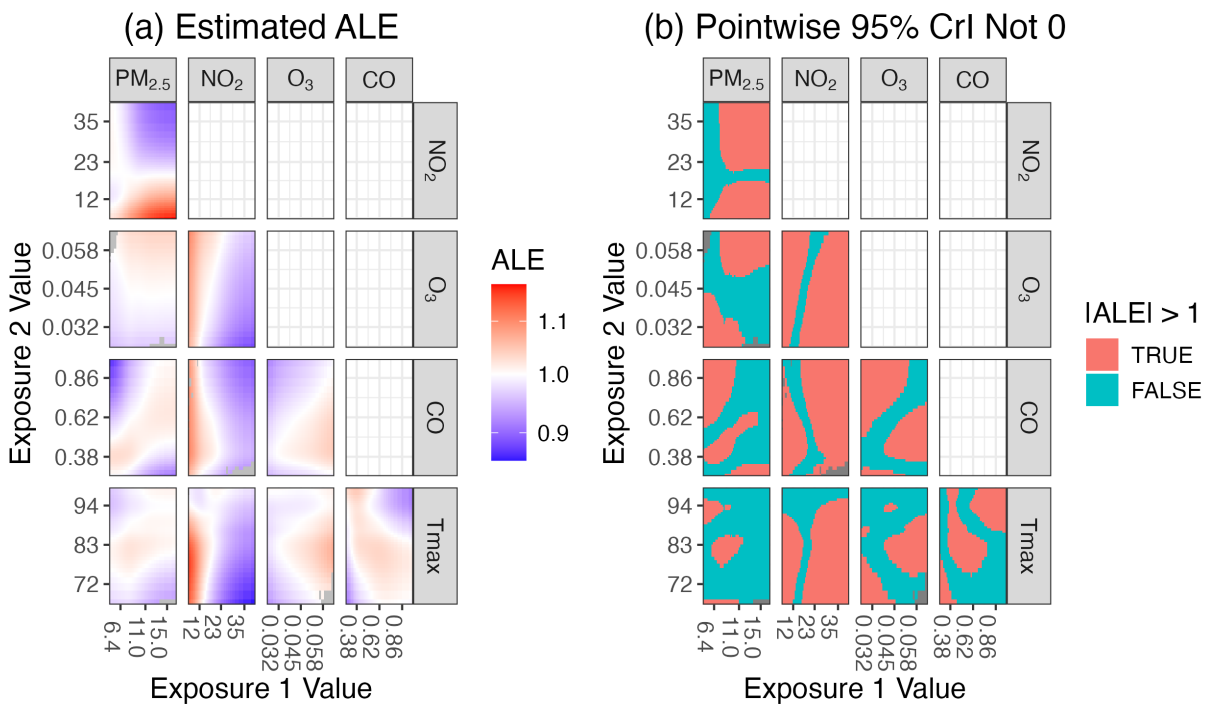

Figure 4.15: **All Mixture Model Pairwise ALEs for the Asthma Application**: Depicted are all of the pairwise second-order ALEs for the mixture model with $T = 25$, with the corresponding main effect ALEs added on. The posterior means for each observed pairwise combination of exposures are plotted in (a), while (b) indicates whether or not the corresponding pointwise 95% credible intervals contain 0. ALEs are calculated using $K = 40$ quantile intervals for each exposure. Plots are trimmed so that only the central 95% of each exposure is displayed.

Chapter 5

Spatially Varying Coefficient Models for Estimating Heterogeneous Mixture Effects

5.1 Introduction

In 2022, the National Center for Health Statistics reported that an estimated 8.60% of infants born in the United States had low birth weight (less than 2,500 grams) [87]. Low birth weight has a strong association with infant mortality and morbidity. In the same year, infant mortality rate in the United States was 42.36 per 1,000 live births among low birth weight infants, compared to just 2.10 per 1,000 live births among infants greater than 2,500 grams [32]. Identifying risk factors of reduced birth weight, particularly those due to modifiable environmental risk factors, is an important research priority.

In environmental and perinatal epidemiology, there is a rich literature supporting the

associations between various air pollutants and birth outcomes including, but not limited to, reduced birth weight [61, 66, 97, 101]. These studies have commonly identified associations between low birth weight and elevated concentrations of $PM_{2.5}$, nitrogen dioxide, sulfur dioxide, ozone, and carbon monoxide, among others. We have previously reported these associations in Atlanta, Georgia [28, 100].

In the past, studies of association between air pollution and health outcomes have utilized single-exposure models. While useful, these models may be inadequate for describing the combined effect of multiple air pollutants that individuals are simultaneously exposed to. In recent years, research has shifted toward developing and applying mixture models that attempt to quantify this joint association between multiple exposures and health. Of the many modeling strategies introduced, quantile g-computation (QGCOMP) proposed by Keil et al. [58] has been the most widely used approach in population-based epidemiologic studies. QGCOMP is favored for its simple definition, computational speed and interpretation of the overall mixture effect, as well as its straightforward implementation via the well-maintained `qgcomp` R package [57].

In many studies of environmental mixtures where QGCOMP or alternative approaches might be used, it is common to have health and exposure data acquired from a large geographical study region. While compiling data from all regions within the study area increases sample size and the ability to detect small mixture effects, it also provides an opportunity to explore spatial heterogeneity in health effects.

In this work, we consider a varying coefficient model based on Bayesian additive regression trees (BART) [25, 30] to estimate spatially heterogeneous mixtures effects within the QGCOMP framework. BART is a flexible modeling approach that has consistently performed

well on a variety of prediction, classification, and causal inference tasks [25, 51, 46]. An additional benefit of using BART is that, unlike most other machine learning models, BART is fully Bayesian and thus offers natural uncertainty quantification via the posterior distribution. We conduct a simulation study to evaluate the method in the presence of spatially varying mixture effects, and then apply the method to an analysis of birth weight from vital records in the state of Georgia.

5.2 Data

5.2.1 Air Pollution Data

We considered five air pollutants: fine particulate matter with diameter 2.5 µm and smaller ($PM_{2.5}$, 24-hr average, $\mu m/m^3$), nitrogen dioxide (NO_2, 1-hr max, ppb), sulfur dioxide (SO_2, 1-hr max, ppb), ozone (O_3, 8-hr max, ppm), and carbon monoxide (CO, 1-hr max, ppb). Daily estimates of the concentrations of each pollutant were derived from a data fusion model which utilized simulations from the Community Multiscale Air Quality Model and monitoring data from the Environmental Protection Agency's Air Quality System database [93]. The original data product is available at a 12km x 12km gridded spatial resolution. We used area-weighted averaging to obtain exposures at the ZIP code level.

5.2.2 Health Data

We obtained birth records from the Office of Health Indicators for Planning of the Georgia Department of Public Health. We restricted the data to only include singleton pregnancies

with gestational age greater than 27 weeks and an estimated date of conception between January 1st, 2005 and December 31st, 2016. There were a total of 1,468,531 births meeting this criteria. Additional covariates collected on the birth mothers included age (years), race, level of educational attainment, marital status, and parity. Pregnancy-wide air pollution exposures were estimated by linking maternal residential address ZIP code and calculating the average concentration of each pollutant from the date of conception to the date of birth.

5.3 Methods

5.3.1 Review of Quantile g-Computation for Mixture Modeling

The goal of QGCOMP, like its predecessor weighted quantile sum (WQS), is to provide a more interpretable mixture effect. For this reason, QGCOMP is sometimes referred to as a *summary index* method. This stands in contrast to *response surface* methods, such as Bayesian kernel machine regression (BKMR), which provide a more flexible approach to modeling complex exposure-response surfaces, but generally are not as easily implemented or interpreted.

In the QGCOMP framework, the target parameter(s) quantify the expected change in the outcome due to an increase of one quantile in all exposures of interest. Because the implementation of QGCOMP leverages model fitting procedures from standard regression models, it can be run efficiently and has been extended to a variety of models. This is particularly important for studies which make use of administrative datasets, for which computationally burdensome methods such as BKMR are impractical.

The linear and additive conditional model for QGCOMP is given by (5.1):

$$Y_i = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{w}_i^T\boldsymbol{\gamma} + \epsilon_i \qquad \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2), \tag{5.1}$$

where Y_i is some continuous outcome (e.g., birth weight), and $\boldsymbol{x}_i$ is a $P_x \times 1$ vector of quantized exposures for observation i. Here, the term quantized exposure refers to an originally continuous exposure whose values have been recoded to $0, 1, \ldots, Q-1$, where the new value represents which of the Q quantiles the original observed value belonged to. The model may also include $\boldsymbol{w}_i$, a $P_w \times 1$ vector containing confounders for adjustment for observation i (note these are not generally quantized). Thus, each element of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{P_x})^T$ represents the expected change in the outcome for a one quantile increase in the corresponding exposure, while the elements of $\boldsymbol{\gamma}$ retain typical interpretations for regression coefficients.

The reasoning behind this treatment of the exposures of interest is to define a *mixture effect* as $\Psi = \sum_{p=1}^{P_x} \beta_p$. When the model is linear and additive in terms of the quantized exposures, Ψ represents the expected change in the outcome for a one quantile increase in every exposure simultaneously. In this scenario, Ψ also coincides with the slope parameter from a simple linear marginal structural model (MSM) in which the sole predictor, denoted S_q, represents the quantized exposure mixture. Specifically, S_q takes on values $0, \ldots, Q-1$, corresponding to when all quantized exposures are simultaneously set to $0, \ldots, Q-1$. When framing the problem in this manner, Ψ may alternatively be estimated via g-computation with the joint exposure quantile S_q. Under certain identifiability assumptions, Ψ might be interpreted as a causal parameter [58].

In this work we focus on the linear and additive model (5.1), but in general, QGCOMP is

not restricted to models which are linear and additive in the quantized exposures. However, if interactions or nonlinearities are included between or among exposures, Ψ is no longer simply the sum of the regression coefficients corresponding to the quantized exposures. This is because the effect of increasing an exposure by a quantile will either depend on the baseline value of the exposure, or the value of another exposure in the case of interaction. Different specifications for the MSM are also possible, with the current version of the `qgcomp` R package supporting polynomial functions of S_q.

5.3.2 Review of Spatially Varying Coefficient Models

Many spatially varying coefficient models have been proposed over the years. Casetti [20] originally described an expansion method for generating improved models by taking the parameters from an initial model and making them a function of variables. Later the geographically weighted regression (GWR) [34] and spatially varying coefficient (SVC) [39] models were proposed. GWR is a frequentist approach that involves estimating a separate weighted least squares regressions at each location, where the weights are determined by proximity between locations as measured by some kernel function. The SVC model is a Bayesian approach which places Gaussian process (GP) priors on the individual regression coefficients, where again a kernel function is used to estimate the distance between observations. Comparisons of the two approaches have found similar performance in many settings, but note that GWR may occasionally struggle in the presence of correlated covariates [110, 111, 33]. The SVC model provides a richer framework for making predictions on new spatial locations and drawing inference for all model parameters, however has large computational overhead

given the use of GP priors, and in simpler settings it may be unnecessarily flexible. An alternative to GP priors for the regression coefficients in applications with areal data are conditional autoregressive (CAR) priors [10]. These reduce the dimension of the coefficients to the number of unique spatial locations and thus are more computationally convenient.

Tangential to these approaches are methods for detecting spatial clusters in varying coefficient models. These methods include formal Monte Carlo hypothesis testing for clustering [63, 64], spatially explicit penalized objective functions [65], and even tree-structured clustering of regression coefficients [9]. Recently, estimation of clustered spatially varying coefficients has made use of spanning trees [65, 67, 30]. Not to be confused with the binary trees that serve as the base learners in a BART model, spanning trees are rooted in graph theory and refer to efficient, non-cyclical paths through a network of vertexes. In the context of areal spatial units, these vertexes are often taken to be centroids of counties, ZIP codes, or other types of regions.

5.3.3 Spatially Varying Quantile g-Computation with BART

We propose allowing the individual exposure coefficients $\boldsymbol{\beta}$ to vary across space, which in turn implies the mixture effect Ψ also varies across space. The result is the following model:

$$Y_i = \beta_0(z_i) + \sum_{p=1}^{P_x} \beta_p(z_i) x_{i,p} + \boldsymbol{w}_i^T \boldsymbol{\gamma} + \epsilon_i \qquad \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2), \tag{5.2}$$

where the intercept and quantized exposure coefficients depend on the spatial location of observation i, z_i. The local mixture effect specific to location z is then defined as $\Psi(z) = \sum_{p=1}^{P_x} \beta_p(z)$.

There are a few reasons allowing for a spatially varying air pollutant mixture effect is warranted. For example, the mixture of $PM_{2.5}$ components may be different spatially due to differences in local emission sources and meteorology. Spatially varying population characteristics that impact the relationships between personal exposure and ambient concentration may also result in effect heterogeneity. When the target estimand is the overall mixture effect, differences may also be attributable to specific exposure levels due to a nonlinear exposure-response relationship. For example, NO_2 may be a more important component in the mixture for regions near highways where levels are high. Additionally, if the true exposure-response surface contains any interactions, then the effects of individual exposures and the overall mixture effect is likely different in regions with different exposure concentrations. Allowing for spatially varying weights allows for capturing locally linear mixture effects, even when the overall mixture effect is more complex.

The spatially varying intercept and exposure coefficients in model (5.2) can be estimated in various ways; we suggest using BART priors for each of these parameters. Deshpande et al. [30] recently developed a varying coefficient BART (VCBART) model and demonstrated its use for studying time series of crime rates across census tracts in Philadelphia, Pennsylvania. When supplied with a list of which sub-regions are spatially adjacent to one another (i.e., share a border), VCBART uses efficient proposal mechanisms based on sampling spanning trees to repeatedly subdivide the study area into contiguous sub-regions which the data suggest are heterogeneous (see Figure 5.1) [29]. This process is done separately for each of the spatially varying parameters in model (5.2), which allows for different spatial clusters for each mixture component. We will henceforth refer to the SVC model with parameters estimated using VCBART as SVC BART.

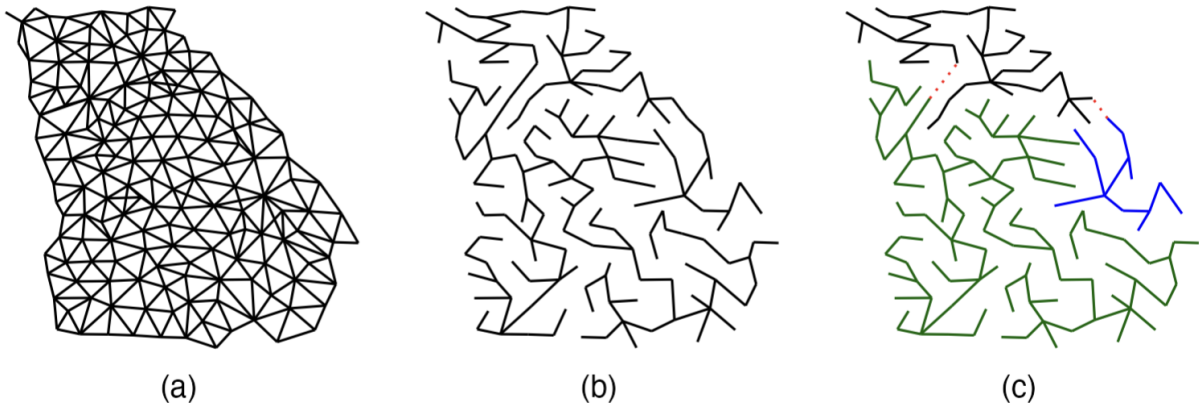

Figure 5.1: **An Illustration of the SVC BART Spatial Branching Process Applied to Georgia Counties**: In panel (a), edges are drawn between centroids of adjacent counties. In panel (b), a random spanning tree is drawn from the graph in panel (a). In panel (c), three groups of contiguous counties are formed by randomly deleting two edges from the spanning tree in panel (b).

Estimation of the SVC BART model is carried out using Markov chain Monte Carlo (MCMC), with each sample from the posterior distribution partitioning the study area differently. The posterior distribution of mixture effects might then be summarized for each location using their posterior means and 95% credible intervals. BART priors function similarly to *boosting*, as each tree contributes a small portion to the overall output, allowing for fine tuning of the spatial branching process. Consistent with other BART implementations, regularization priors are used to encourage homogeneity across the entire study area to prevent over-fitting. In general, BART priors are more computationally feasible than Gaussian process priors in large sample sizes, while retaining much of the same flexibility.

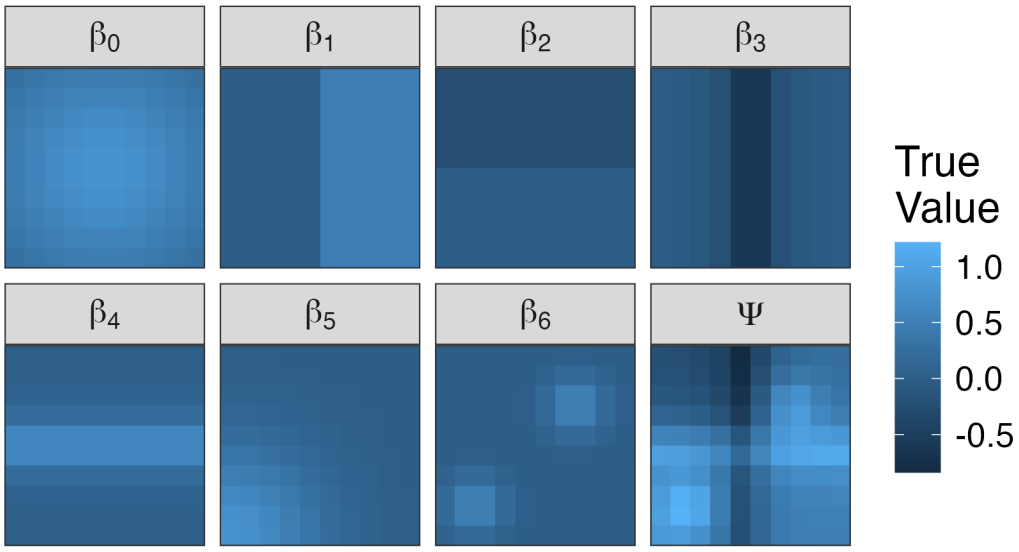

Figure 5.2: **True Spatially Varying Parameter Surfaces for the Spatial QGCOMP Simulation Study**: Spatially varying intercept (β_0), regression coefficients (β_1-β_6), and overall mixture effect (Ψ) surfaces used for the simulation study. All are defined on a 10 x 10 grid.

5.4 Simulation Study

In this section we evaluate the ability of SVC BART to estimate the spatially varying parameters of model (5.2). We generate a spatially varying intercept, $\beta_0(z)$, and six spatially varying regression coefficients, $\beta_1(z), ..., \beta_6(z)$, across a 10 x 10 grid using various smooth and rigid functions (see Section 5.7.1 of the Chapter 5 Supplementary Materials for details of the functions). The surfaces are plotted in Figure 5.2, along with the true mixture effect $\Psi(z)$.

For the simulation, the exposures are generated from a mean zero multivariate normal distribution with covariance $\rho\mathbf{J}_6 + (1-\rho)\mathbf{I}_6$ (i.e., with an exchangeable correlation structure). Each of the exposures is then quantized using $Q = 4$ quantile bins. Finally, the outcome is drawn from a normal distribution with some noise variance σ^2. Parameters varied during the simulation study include the sample size within each grid cell ($n \in \{10, 50, 100, 250\}$), the degree of correlation between exposures ($\rho \in \{0.0, 0.5, 0.8\}$), and the amount of noise

($\sigma \in \{0.1, 1\}$). Each parameter setting is run for $B = 200$ unique datasets.

We fit all SVC BART models for the simulation using the `VCBART` R package publicly available on GitHub (`https://github.com/skdeshpande91/VCBART`). The default hyperparameter settings from the package are used, including 50 trees per BART ensemble. Each model is run for 2,000 MCMC iterations, discarding the first 1,000 samples as burn-in. The most natural comparison might be the SVC GP model, which uses GP priors in place of BART priors [39]. Others have described the connection between BART and GP priors [69]. However, due to the computational burden presented by GP priors, studies of areal data often make use of CAR priors [10]. For this reason, we compare SVC BART to a model with proper Besag CAR priors on the intercept and each of the regression coefficients (SVC CAR). We fit these models using the integrated nested Laplace approximation (INLA) available in the `INLA` R package and simulate 1,000 draws from the posterior distribution (`https://www.r-inla.org`).

To evaluate the models, we compute the global average 95% posterior credible interval coverage and root mean squared error (RMSE) for the 100 mixture effects (one for each grid cell). These quantities describe the average performance of the model across all grid cells and are calculated as in (5.3) and (5.4), where $\hat{\Psi}(z)$, $\hat{\Psi}(z)_{0.025}$, and $\hat{\Psi}(z)_{0.975}$ are the mean, 2.5th percentile, and 97.5th percentile of the posterior distribution of the estimated mixture effect at location z, and $\mathbb{I}$ is an indicator function.

$$\widehat{\text{Coverage}} = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{100}\sum_{z}\mathbb{I}\left[\Psi(z) \in [\hat{\Psi}(z)_{0.025}, \hat{\Psi}(z)_{0.975}]\right]\right) \tag{5.3}$$

$$\widehat{\text{RMSE}} = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{100}\sum_{z}\left[\Psi(z) - \hat{\Psi}(z)\right]^2\right) \tag{5.4}$$

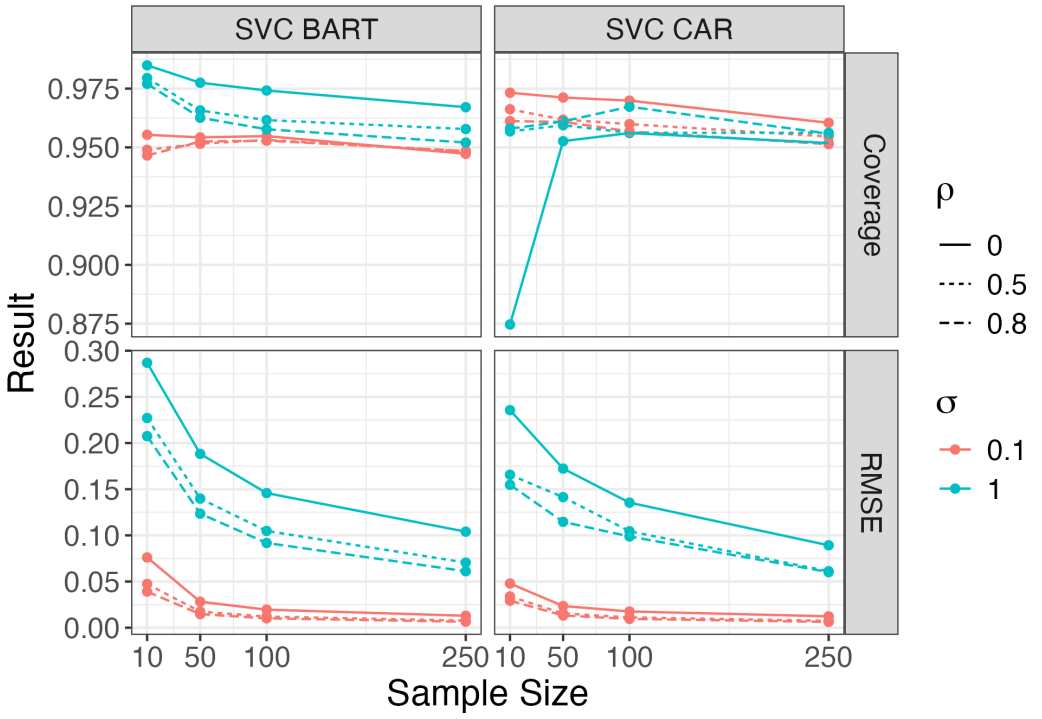

Figure 5.3: **Spatial QGCOMP Simulation Results - Global Mixture Effect**: Global average 95% credible interval coverage and RMSE for the local mixture effects using SVC CAR and SVC BART models in the simulation study.

Figure 5.3 contains a summary of the simulation results for both the SVC CAR and SVC BART models. As the sample size increases, global coverage tends toward 95% and RMSE decreases for both models. The CAR model has slightly better RMSE in small samples, but the difference is negligible in the $n = 250$ setting. In general, better global coverage and RMSE is observed when exposures exhibit stronger correlation. Despite this, the individual performance on any one of the spatially varying coefficients may decrease with increasing correlation (see Figure 5.8 in the Chapter 5 Supplementary Materials).

While the global statistics suggest the two models are performing at a somewhat similar level, there are differences in each model's ability to estimate the local mixture effects for each grid cell. The 95% credible interval coverage for the local mixture effects is shown

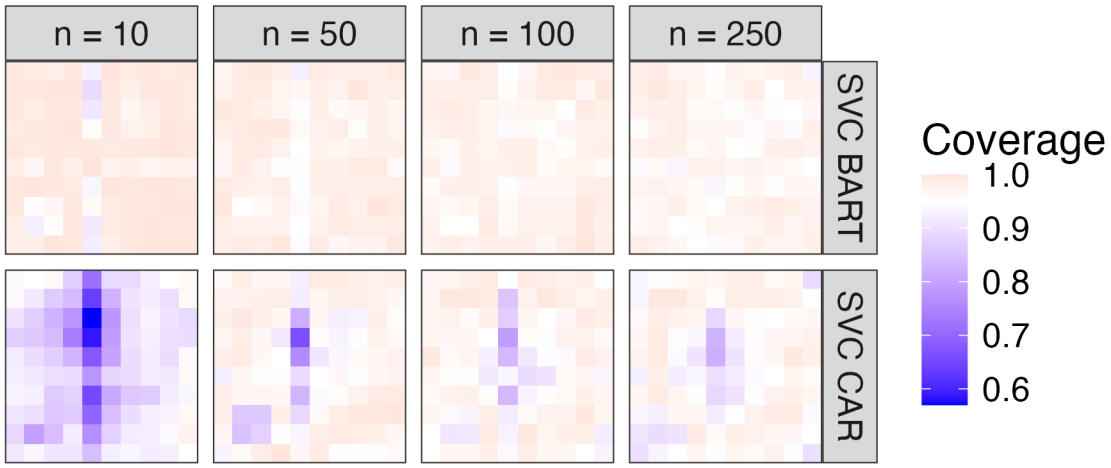

Figure 5.4: **Spatial QGCOMP Simulation Results - Local Mixture Effect Coverage**: 95% credible interval coverage for local mixture effects $\Psi(z)$ when sample size per cell is 10, 50, 100, and 250. Fixed settings: $\rho = 0, \sigma = 1$.

in Figure 5.4 for the high noise setting ($\sigma = 1$) with uncorrelated exposures ($\rho = 0$). The coverage for SVC CAR is very poor for many of the cells in the lowest sample size ($n = 10$), but improves some as the sample size increases. On the other hand, SVC BART generally has better coverage across all grid cells, with near 95% coverage even in small sample sizes.

Spatial patterns of poor coverage for local mixture effects might be attributable to poor coverage for one or more of the constituent local exposure coefficients. We found that the SVC CAR model struggles most with the spatial patterns used to generate β_1, β_2, and β_6 (see Figure 5.10 in the Chapter 5 Supplementary Materials for simulation average coverage for each spatially varying coefficient for the $n = 100$ setting, corresponding to the third column of Figure 5.4). These surfaces contain some of the sharpest contrasts between neighboring cells, which presents difficulties for models which rely on spatial smoothing. SVC BART also struggles to capture the two hot spots in the β_6 surface, but not to the same extent as the

SVC CAR model, and generally has as good or better coverage across the other parameters.

These results suggest that SVC BART may be preferable to SVC CAR in settings with high-noise or small local sample sizes. In general, we found that as ρ increases, coverage and bias for $\Psi(z)$ improves or changes little, while coverage and bias for the spatially varying regression parameters worsens. The latter was particularly noticeable for the SVC CAR model. We also noticed that coverage for the SVC CAR model was substantially worse in the high noise variance setting, whereas the amount of noise had little effect on coverage for SVC BART.

5.5 Application: Reduced Birth Weight and Air Pollution in Georgia

5.5.1 Descriptive Statistics

In an application of SVC BART, we analyzed 1,468,531 live singleton births to mothers residing in Georgia with an estimated conception date between January 1st, 2005 and December 31st, 2016. In this sample, the majority of mothers were white (58.1%), and in terms of educational attainment about half (50.7%) reported at least some college experience. Additional demographic information is provided in Table 5.1.

5.5.2 Model Considerations

We fit an SVC BART model with the default 50-trees-per-ensemble setting and set aside 2,000 posterior samples after discarding the first 2,000 as burn-in. A spatially varying intercept, as

Table 5.1: Maternal Demographic Characteristics

Characteristic	N	%
Race		
White	853,575	58.1
Black	505,304	34.4
Asian or Pacific Islander	58,706	4.0
American Indian or Alaskan Native	2,460	0.2
Other	48,486	3.3
Ethnicity		
Non-Hispanic	1,252,268	85.3
Hispanic	216,263	14.7
Age		
Less than 25 years	519,590	35.4
25-31 years	568,344	38.7
More than 31 years	380,597	25.9
Education		
Less than 9th grade	73,584	5.0
9th-11th grade	204,226	13.9
12th grade	446,449	30.4
Some college	744,272	50.7

well as spatially varying coefficients for quantized versions of $PM_{2.5}$, NO_2, SO_2, O_3, and CO were included. For this analysis, we chose to quantize each exposure into 10 quantile bins, i.e., deciles. Additional covariates modeled using fixed effects included estimated conception date, gestational age, tobacco use, and the parity, age, race, ethnicity, level of educational attainment, and marital status of the mother. We also adjusted for socioeconomic status using Census tract-level estimates of the percentage below the poverty level. The continuous covariates age, tract poverty level, and conception date were modeled used natural cubic splines with 5 degrees of freedom, while gestational age was modeled using indicator variables for the number of weeks.

5.5.3 Results

As previously mentioned, one of the reasons a spatially varying coefficient model might be appropriate for this analysis is that pollutant concentrations may vary across space. We calculated the average (mean) pregnancy-wide concentration of each pollutant within each Georgia county. The distribution of these county-level averages are summarized in Table 5.2. Most notably, mothers in counties at the 90th percentile of NO_2 and SO_2 were, on average, exposed to more than double the concentration of these pollutants compared to mothers in counties at the 10th percentile. Figure 5.5 displays the median pregnancy-wide pollutant concentrations for each county, after the exposures have been quantized into deciles. While pollutant concentration typically varies seasonally, some trends are clear, such as CO and NO_2 concentrations being highest in the Atlanta metropolitan area, and NO_2 following the path of Interstate 75.

Table 5.2: Percentiles of County-level Mean Pregnancy-wide Pollutant Exposures

	Percentile				
Pollutant	10th	25th	50th	75th	90th
24-hr average $PM_{2.5}$ ($\mu m/m^3$)	9.00	9.59	10.22	10.65	10.65
1-hr max NO_2 (ppb)	4.40	5.12	6.60	9.53	9.53
1-hr max SO_2 (ppb)	1.90	2.25	3.14	3.99	3.99
8-hr max O_3 (ppb)	38.98	39.66	40.50	41.22	41.22
1-hr max CO (ppb)	0.30	0.31	0.33	0.37	0.37

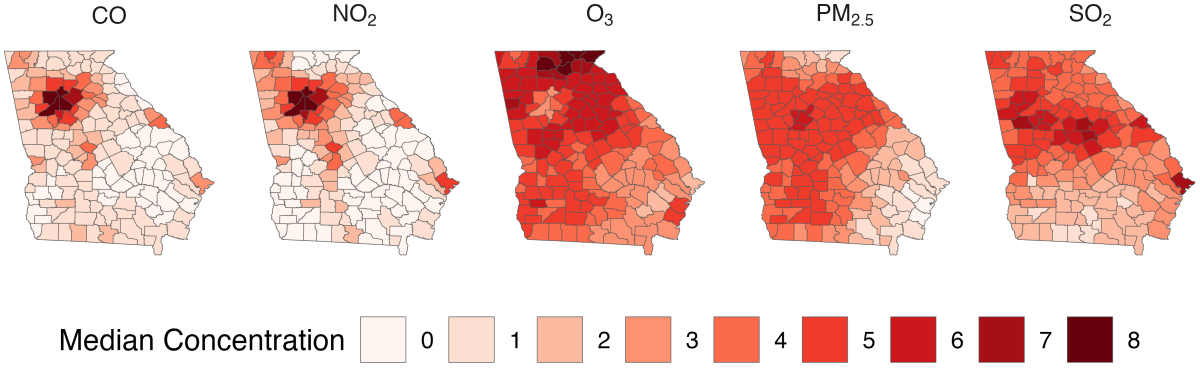

Figure 5.5: **Spatial Distribution of Quantized Air Pollutants**: Five maps depicting the county-level median pregnancy-wide concentration for each air pollutant after having been quantized into deciles.

Posterior means of the mixture effects using the SVC BART model are plotted in Figure 5.6. In general, we found stronger negative mixture effects in counties making up the central and eastern portion of the state. For a one decile increase in all five exposures, the county-specific estimates range from an expected reduction in birth weight of -16.65 grams (95% CrI: -33.93, -0.40) in Decatur county to an increase of 13.28 grams (95% CrI: 0.06, 27.19) in Wheeler county. Of the counties with 95% credible intervals that exclude zero, 17/23 are in a negative direction. A forest plot of these 23 county-level mixture effects is provided in Figure 5.11 of the Chapter 5 Supplementary Materials.

As a comparison, we also estimated a common mixture effect using a linear model with only a spatial CAR random intercept and no spatially varying coefficients. This common mixture effect was estimated to be a reduction of -1.81 grams (95% CrI: -2.84, -0.70) per decile increase in all pollutants. This estimate is slightly attenuated compared to a weighted average of the local mixture effects from the SVC BART model (reduction of 2.27 grams). SVC BART outperformed the CAR random intercept model, as well as an SVC CAR model akin to that which was fit in the simulation study, in terms of the Widely Applicable Information Criterion [108] (see Table 5.3 in the Chapter 5 Supplemental Materials). In fact, both CAR models (random intercept and SVC) performed worse than a non-spatial ordinary least-squares regression in terms of WAIC. Locally, SVC BART ranked the best in terms of in-sample RMSE for nearly all counties, and ranked the best in terms of out-of-sample performance (approximated using WAIC) for the greatest number of counties compared to other approaches (see Figure 5.12 of the Chapter 5 Supplementary Materials).

In Figure 5.7, we plot the estimated local mixture effects against county-level summaries of the exposures and confounders included in the SVC BART model. For the most part, there is

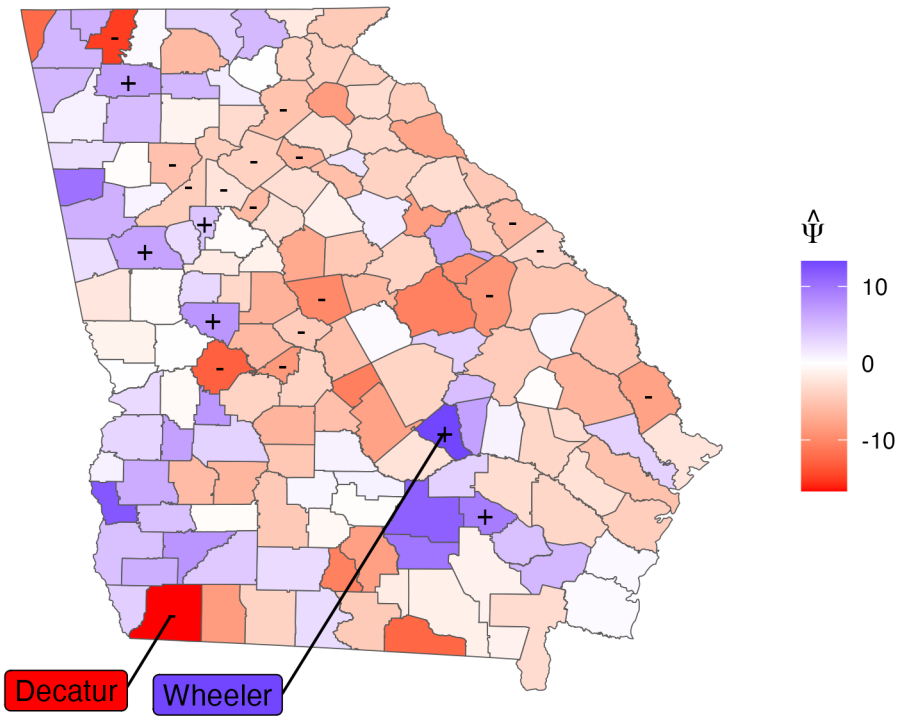

Figure 5.6: **County-Level Mixture Effects for the Birth Weight Application**: The posterior means of the local mixture effects $\Psi(z)$ estimated using SVC BART are displayed. Counties marked with a "–" have a 95% credible interval entirely below zero, and counties marked with a "+" have a 95% credible interval entirely above zero.

no discernible pattern in the estimated mixture effects when compared to the confounders. In terms of the exposures, for CO and NO_2, the local mixture effects tend to shift in the negative direction as the level of exposure increases at the lower end of the observed concentrations.

5.6 Discussion

We describe varying coefficient models as a useful extension to the popular QGCOMP method to account for when heterogeneous exposure-response relationships are present in the data. We have shown through simulation and an analysis of birth records in Georgia how one might estimate spatially varying parameters in such a model via a Bayesian approach which uses CAR or BART priors. In our analysis of birth records, we found that for many Georgia counties there exists an association between elevated concentrations of a mixture of $PM_{2.5}$, NO_2, SO_2, O_3, and CO and reduced birth weight.

A limitation of the analysis is the measurement of exposures. Not only is there potential for error in the exposure measurements themselves, but the mechanism by which we assign pregnancy-averaged pollutant concentrations to each mother is imperfect. The residential address on file may not be reflective of where the mother spent most her time during the pregnancy, and even when it is, the amount of exposure two individuals from the same neighborhood experience could be very different due to unmeasured factors such as occupation or personal lifestyle behaviors. Additionally, the mixture effect definition in QGCOMP is not always of interest. QGCOMP runs the risk of extrapolating to unseen exposure profiles during the estimation process, and so even when exposures are measured accurately, the method may not always be reliable.

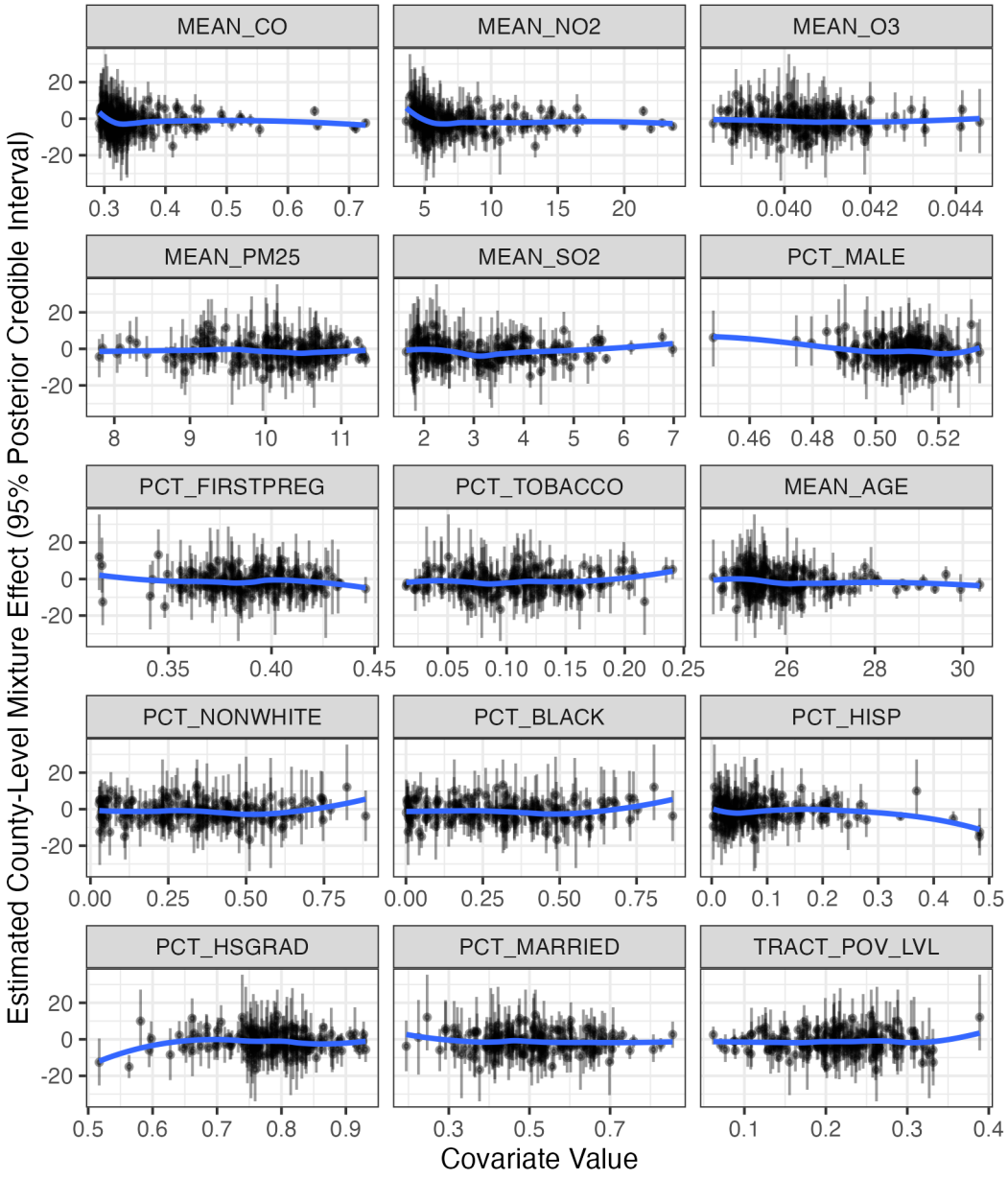

Figure 5.7: **Scatter Plots of Local Mixture Effects by Covariates**: Posterior means and 95% credible intervals for the local mixture effects $\Psi(z)$ estimated using SVC BART are plotted on the y-axis, and selected county-level average exposures and other covariates are plotted on the x-axis. A smooth loess curve is overlaid.

Generally, ambient air pollution contributes little explanatory power for birth weight. The R^2 value for the SVC BART model in the application is 39.90%. It is possible that all or some subset of these pollutants may be more informative if averaged during a critical window of the pregnancy instead of the entire duration. Previous studies have observed different associations between air pollution and birth weight in Atlanta, Georgia for specific months or trimesters of pregnancy [28, 100]. Various data driven methods have been developed for identifying windows of pregnancy particularly susceptible to air pollution, including some based on BART [113, 22, 80, 81]. On a similar note, spatially varying distributed lag models, such as the one proposed in Warren et al. [107], might be another QGCOMP extension worth exploring.

BART is not the only option for fitting the SVC regression, as any of the models described in Section 5.3.2 might be useful. In terms of the methods we have considered, SVC BART is more computationally burdensome than the SVC CAR model, particularly when an efficient implementation like INLA is used. This is due to the overhead required for managing tree structures. However, we have found that this tree-based approach is advantageous over the SVC CAR model to estimate local mixture effects, particularly in high-noise settings such as our birth weight analysis. Also, we only consider a spatially varying coefficient model for areal data, but alternative approaches that use GP priors [39] or a mixture of BART and GP priors [77] might perform well for estimating smooth mixture effects over a region from point-referenced data, with the main limitation being computational burden.

In this work, we focused on a spatially heterogeneous approach to quantile g-computation that made use of SVC BART's graph-structured branching process. However, in practice one could also use the traditional BART branching process to model heterogeneity in the

exposure coefficients as a function of demographic or clinical covariates as we have done in Chapter 3. For instance, Darrow et al. [28] reported higher estimates of the associations between various pollutants and birth weight for Hispanic and Non-Hispanic Black infants compared to Non-Hispanic White infants. The current implementation of the `VCBART` R package requires the BART ensembles to all use the same set of covariates, and a future extension would also be to select different covariates for each exposure.

Finally, while we have restricted our attention to the Gaussian regression setting, it is also worth noting that the QGCOMP framework has been implemented for binary, count, and survival outcomes, with the resulting interpretations changing to those used for logistic/probit, Poisson, and Cox proportional hazards regression, respectively. While BART has been extended to these settings [25, 82, 14, 96], developing VCBART models is nontrivial.

5.7 Supplementary Material

5.7.1 Additional Simulation Materials

For the simulation study in Section 5.4, we generate a spatially varying intercept and six spatially varying regression coefficients across a 10 x 10 grid using the following functions:

$$\beta_0(x,y) = 100\phi\left((x,y)^T \mid (5.5,5.5)^T, 20\mathbf{I}_2\right) \tag{5.5}$$

$$\beta_1(x,y) = 0.50 \times \mathbb{I}\left[x > 5\right] \tag{5.6}$$

$$\beta_2(x,y) = -0.25 \times \mathbb{I}\left[y > 5\right] \tag{5.7}$$

$$\beta_3(x,y) = -\exp\left(-|x-5.5|\right) \tag{5.8}$$

$$\beta_4(x,y) = \exp\left(-|y-5.5|\right) \tag{5.9}$$

$$\beta_5(x,y) = 50 \times \phi\left((x,y)^T \mid (1,1)^T, 10\mathbf{I}_2\right) \tag{5.10}$$

$$\beta_6(x,y) = 4 \times \left[\phi\left((x,y)^T \mid (7.5,7.5)^T, \mathbf{I}_2\right) + \phi\left((x,y)^T \mid (2.5,2.5)^T, \mathbf{I}_2\right)\right] \tag{5.11}$$

where x and y correspond to the integer dimensions of the cells in the grid, ϕ is the probability density function of a bivariate normal distribution, $\mathbb{I}$ is an indicator function, and $\mathbf{I}$ is a diagonal identity matrix.

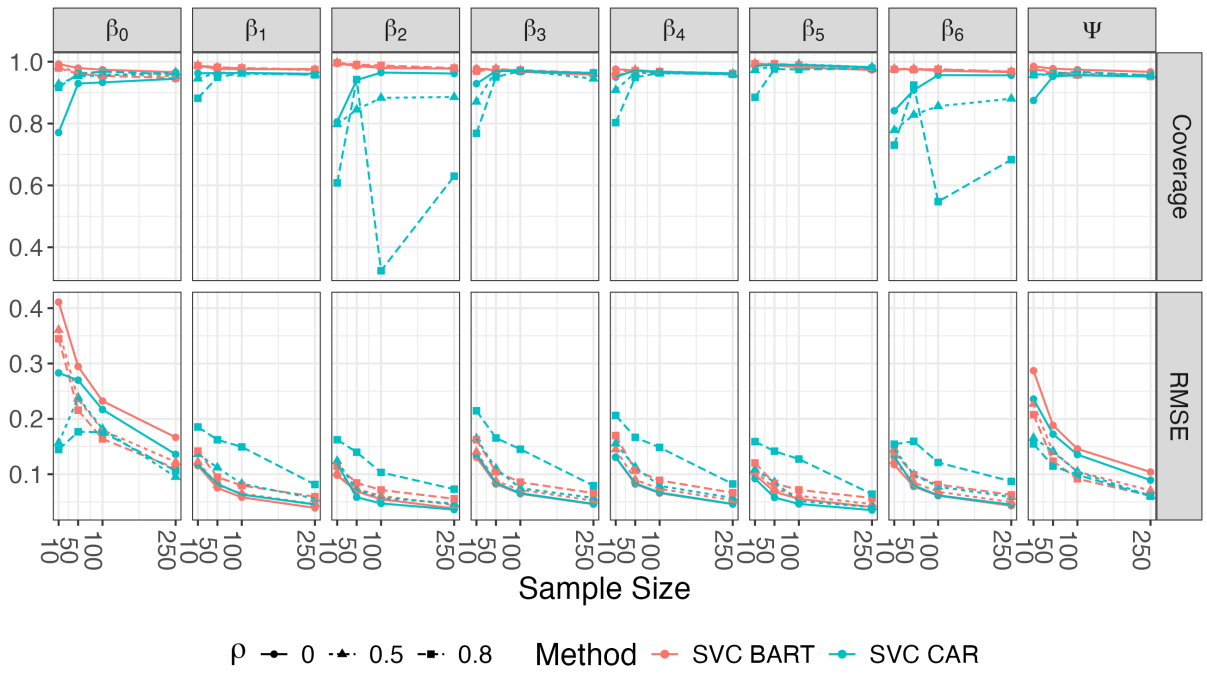

Figure 5.8: **Spatial QGCOMP Simulation Results - Global Performance for All Spatially Varying Parameters (High Noise Setting)**: Global average 95% credible interval coverage and root mean squared error (RMSE) for the local mixture effects $\Psi(z)$ and individual exposure coefficients. Fixed settings: $\sigma = 1$.

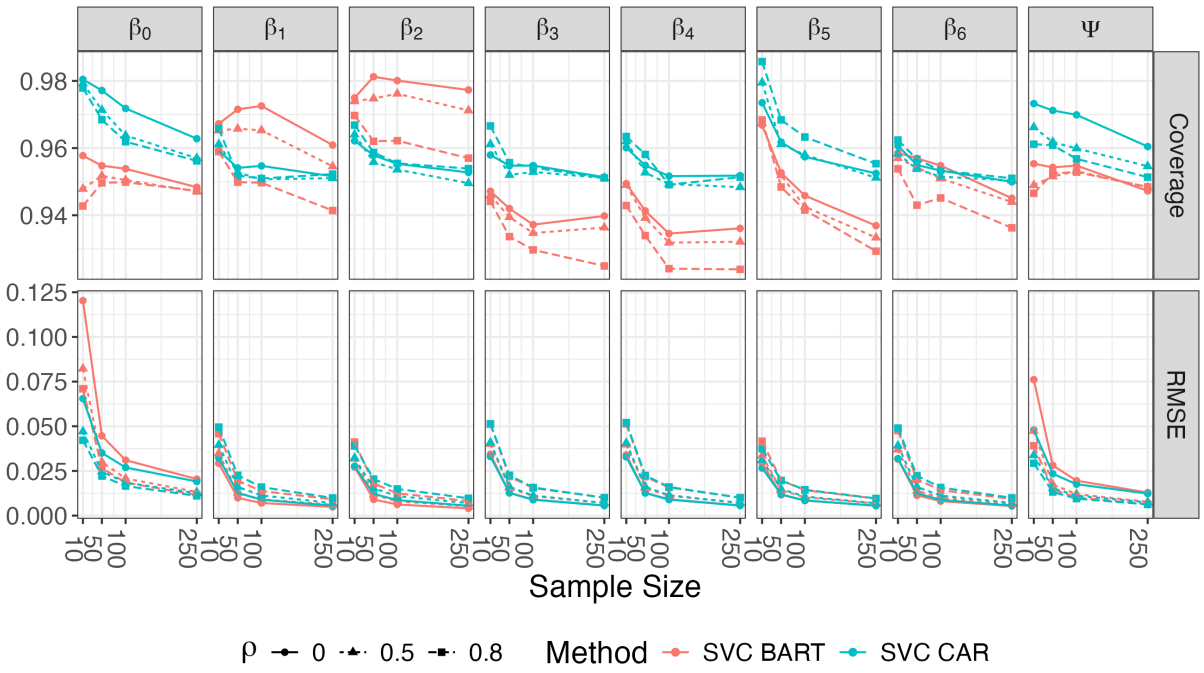

Figure 5.9: **Spatial QGCOMP Simulation Results - Global Performance for All Spatially Varying Parameters (Low Noise Setting)**: Global average 95% credible interval coverage and root mean squared error (RMSE) for the local mixture effects $\Psi(z)$ and individual exposure coefficients. Fixed settings: $\sigma = 0.1$.

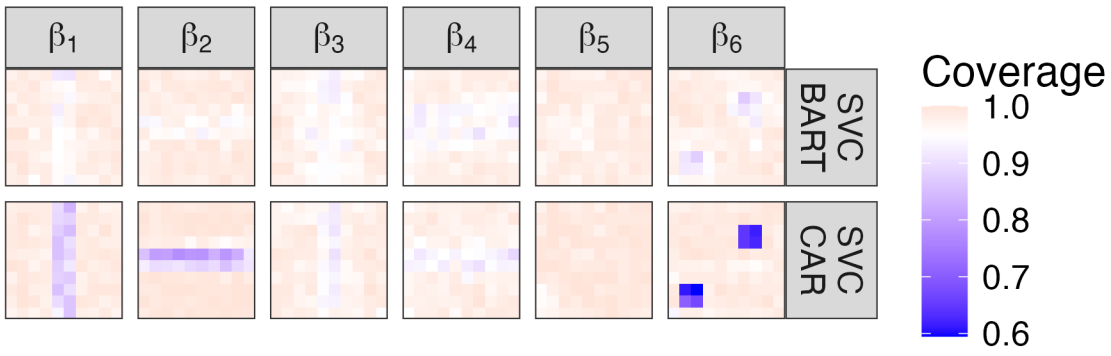

Figure 5.10: **Spatial QGCOMP Simulation Results - Local Coverage for All Spatially Varying Parameters**: 95% credible interval coverage for all spatially varying regression coefficients. Fixed settings: $n = 100, \rho = 0, \sigma = 1$.

5.7.2 Additional Application Materials

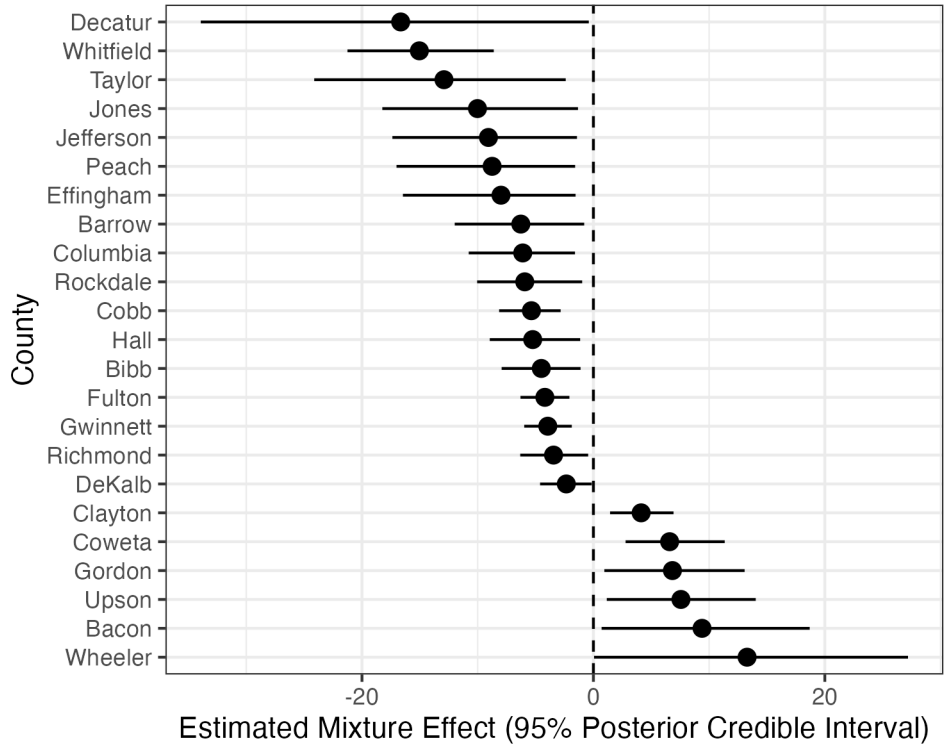

Figure 5.11: **Notable Local Mixture Effects for the Birth Weight Application**: Local mixture effects estimated using SVC BART. Only the 23 counties with 95% posterior credible intervals excluding zero are shown.

Table 5.3: WAIC for Candidate Models for the Birth Weight Application

Model	WAIC
SVC BART	21,937,450
Ordinary Least Squares Regression	21,939,763
Spatial Random Intercept (CAR)	21,943,089
SVC CAR	21,943,136

WAIC: Widely applicable information criterion.

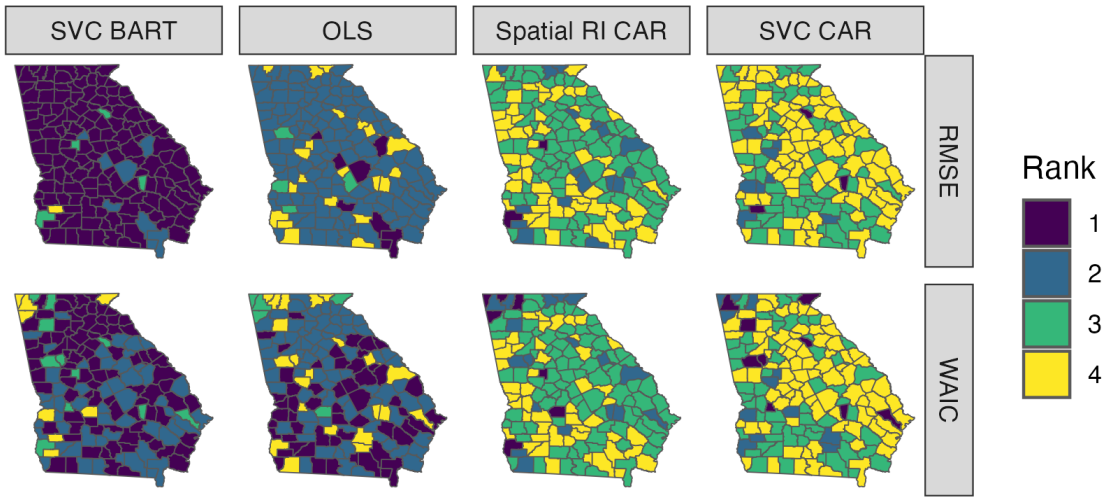

Figure 5.12: **Local RMSE and WAIC Rank for Candidate Models for the Birth Weight Application**: Lower rank corresponds to lower values and hence better performance. In-sample performance is measured using RMSE, while WAIC is used as an approximation for out-of-sample performance.

Chapter 6

Conclusion

6.1 Summary of Contributions

This dissertation advances the application of Bayesian nonparametric models in environmental health research by extending Bayesian Additive Regression Trees (BART) to address key challenges in estimating exposure-response relationships. Across three distinct study designs, this work demonstrates how BART can improve flexibility, interpretability, and inference in settings where traditional parametric models face limitations.

First, a varying coefficient BART model was introduced to estimate individual-level exposure-response relationships for acute exposures and binary health outcomes within the case-crossover design. By incorporating patient-specific covariates, this approach provides a data-driven assessment of vulnerability to environmental stressors, which is particularly relevant for studying populations with underlying health conditions.

Next, a negative binomial BART framework was implemented for modeling count-based health outcomes in environmental mixture studies. By leveraging data augmentation, this

approach supports flexible estimation of smooth exposure-risk surfaces while accounting for spatial, temporal, and covariate confounding. Beyond methodological advances, this work highlights the challenges associated with interpreting mixture models and describes the use of accumulated local effects as a tool for improving interpretability.

Finally, this work contributes to the growing literature on spatially varying mixture effects by extending quantile g-computation to account for heterogeneity across diverse geographic regions. By utilizing a spatially explicit varying coefficient BART model, this approach relaxes the assumption of a common mixture effect, allowing for location-specific insights that may be useful for public health decision-making.

Collectively, these contributions broaden the scope of Bayesian tree-based methods in environmental health applications.

6.2 Future Directions

The discussions at the end of Chapters 3, 4, and 5 provide suggestions for future work relating to the specific methods presented in this dissertation. More generally, BART has potential for an expanded role in environmental health research.

6.2.1 Exposure Modeling

The work presented in this dissertation focuses exclusively on modeling associations between one or more environmental exposures and various health outcomes. BART has proven useful for us and for many others for this task. While significant, health outcome modeling only represents a portion of the statistical challenges faced in environmental health. In our work

we take exposure data as a known quantity, when it is often the case that the exposures we work with are estimated using sophisticated models of their own. Tree-based models such as random forests tend to perform well for exposure modeling [93], so future work might explore applying BART in this setting.

6.2.2 Causal Inference for Environmental Mixtures

One of the areas BART shines most is the estimation of (conditional) average treatment effects in various populations of interest [52, 31, 46]. Bayesian nonparametric models like BART offer a highly flexible interface for estimating these effects [85, 4, 72].

A key challenge in applying BART to causal inference problems in environmental health is defining the treatment variable in a meaningful way. When it comes to studies of environmental exposure mixtures, there are multiple continuous exposures. Summary index methods such quantile g-computation attempt to simplify the definition of a mixture effect, but this definition is not without its faults. New causal inference frameworks might consider environmental mixtures as a multivariate treatment, where shifts in the individual exposures define causal effects within a statistical model. A more policy-relevant approach would define the treatment in terms of external interventions on one or more of the exposures constituting the mixture.

A fundamental challenge in environmental epidemiology is that both association and causal inference studies depend critically on accurate exposure measurement. Exposure assignments are often estimated using environmental models rather than directly observed at the individual level. The process of linking exposures to individuals – whether through ambient monitoring, geospatial interpolation, or modeled predictions – introduces error. This

measurement error can lead to bias, increased variability, or even spurious causal conclusions, particularly when exposures are estimated at coarse spatial or temporal resolutions.

6.3 Software

Several R packages were developed to conduct the analyses discussed in this dissertation. All packages are available along with vignettes/tutorials on GitHub (`https://github.com/jacobenglert`).

- `clbart`: Conditional Logistic Regression with BART. Implements the CL-BART approach from Chapter 3.
- `jrpg`: Just Random Pólya-Gamma Variables. Facilitates sampling of latent variables for the Gibbs sampler in Chapter 4.
- `pdpd`: Partial Dependence for Posterior Distributions. Calculates partial dependence and accumulated local effects statistics for Bayesian regression models.
- `VCBART`*: Varying Coefficient Bayesian Additive Regression Trees. Fits the spatially varying coefficient model from Chapter 5. *Note that as of the time of dissertation submission, this is a fork of the original VCBART package available at `https://github.com/skdeshpande91/VCBART` that allows for modeling a subset of covariates without traditional fixed regression coefficients.

Code for implementing the Soft BART Gibbs sampler in Chapter 4 is provided in the Chapter 4 Supplementary Materials.

Appendix A

Appendix

A.1 Selection of Controls in the Case-Crossover Design

The case-crossover design used in Chapter 3 is a default choice for analyzing data where only cases (or events) are observed and estimating short-term associations of acute exposure on (relatively) rare events is a priority. The general concept of the design is that if the event were experienced on a given day, then it was likely not experienced on other days near in time. Under this assumption, a referent window of controls is selected for each event, essentially matching the events to themselves. The selection of the referent window is very important to ensure bias due to time trend and seasonality is properly controlled and adjusted for [18, 53].

Consider we observe an event on Tuesday, March 18th, 2025. Some popular strategies for selecting control days are illustrated in Figure A.1. In Figure A.1a, one control observation is selected exactly one week prior to the event. This is a *unidirectional* scheme. This scheme does not adequately control for time trend. The *symmetric bidirectional* scheme is illustrated in Figure A.1b. This scheme selects two control days for the observation, exactly

one week before and after the event occurred. This strategy controls for bias due to time trend and day of the week. Both the unidirectional and symmetric bidirectional schemes are *nonlocalizable*, meaning that they suffer from *overlap bias* as a result of there not being an unbiased estimating equation when restricting to the referent windows. This is important to because researchers often prefer to use conditional logistic regression to avoid needing to adjust for confounders which are matched on (e.g., the $\boldsymbol{v}$ covariates in (3.2)). In both of these designs, the index time point for the event is known given the referent window [54].

Another option is the *semi-symmetric bidirectional* scheme, which randomly selects one of the control days given by the symmetric bidirectional scheme for each event. This results in a *localizable* design, meaning an unbiased estimating equation exists. However, such a design is *nonignorable*, meaning that conditional logistic regression cannot be used blindly. The actual unbiased estimating equation depends on how the controls were selected and thus requires additional modification [53, 54].

Figure A.1c and A.1d depict the *full stratum bidirectional* and *time-stratified* schemes. Both of these are localizable and ignorable, meaning that unbiased inference can be made using the standard conditional logistic regression estimating equations using only the referent windows. Both control for time trend and day of the week. Generally speaking, these approaches avoid overlap bias because the referent windows represent disjoint strata. This property allows for simplification in the estimating equations because summations can be expressed over strata as opposed to all index time points. The time-stratified scheme is preferred because it also controls for seasonality and introduces far fewer control days to the analysis. The strata in Figure A.1d matches control days to event days based on day of the week, month, and calendar year, but other approaches are also possible.

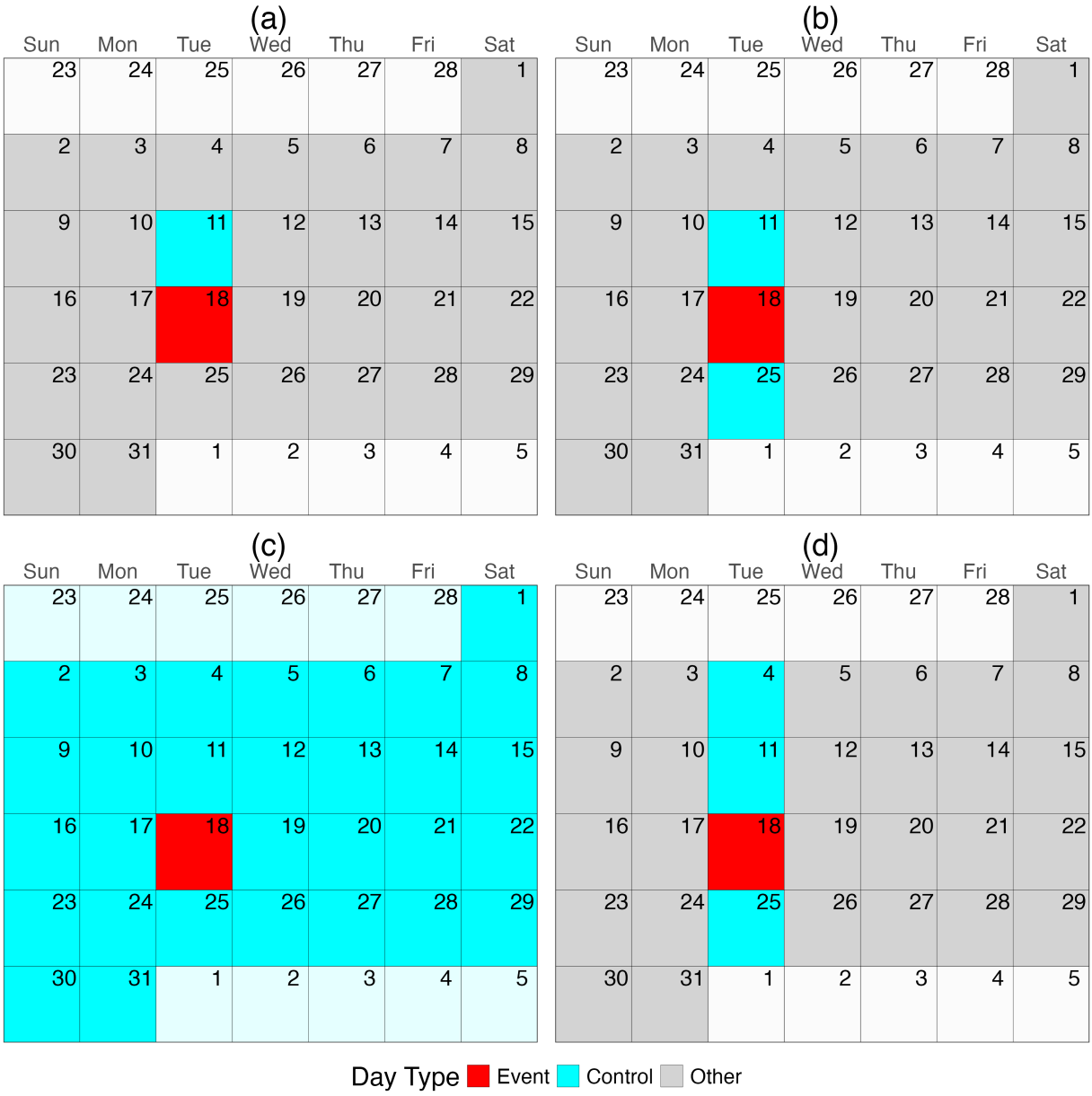

Figure A.1: **Common Referent Window Selection Schemes for the Case-Crossover Design**: Illustrated are the (a) unidirectional, (b) symmetric bidirectional, (c) full stratum bidirectional, and (d) time-stratified schemes.

A.2 Bayesian Computation

A.2.1 Metropolis-Hastings Algorithm

The Metropolis-Hastings (M-H) [78, 50] algorithm is a Markov chain Monte Carlo method used to generate samples of a parameter, θ from some target distribution, $p(\theta)$. To generate a sample from $p(\theta)$, start with some initial value θ. Then sample a value of θ' from a proposal distribution denoted $q(\theta' \mid \theta)$. Finally, accept this proposed value as a sample from the target distribution with probability $\min(1, r)$, where r is defined as in (A.1).

$$r = \frac{p(\theta')q(\theta \mid \theta')}{p(\theta)q(\theta' \mid \theta)} \tag{A.1}$$

If the proposed value is accepted, update the current value of θ to θ'. Otherwise, do not change the current value of θ. This process can be repeated S times to generate a sample of S observations from the target distribution $p(\theta)$. This procedure is formalized in Algorithm A.6. Depending on the quality of the starting value and proposal distribution, some amount of initial draws may need to be discarded as burn-in samples.

Algorithm A.6 Metropolis-Hastings Algorithm

1: Initialize $\theta^{(0)}$ (starting point)
2: **for** $s = 1$ to S **do**
3: Sample θ' from proposal distribution $q(\theta'|\theta^{(s-1)})$
4: Compute acceptance ratio:

$$r = \frac{p(\theta')}{p(\theta^{(s-1)})} \times \frac{q(\theta^{(s-1)}|\theta')}{q(\theta'|\theta^{(s-1)})}$$

5: Set $\theta^{(s)} = \begin{cases} \theta' & \text{with probability } \min(1, r) \\ \theta^{(s-1)} & \text{otherwise} \end{cases}$
6: **end for**
7: **Return** $\{\theta^{(s)}\}_{s=1}^{S}$

In the context of Bayesian inference, the posterior distribution serves as the target distribution. Mathematically speaking, $p(\theta) = \pi(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)\pi(\theta)$, where $\pi(\theta)$ is the prior distribution for θ and $p(\mathcal{D} \mid \theta)$ is the likelihood of the data given the current value of θ. The choice of proposal distribution is up to the researcher. It is common to specify a symmetric proposal distribution, i.e. $q(\theta' \mid \theta) = q(\theta \mid \theta')$. This causes the proposal distribution to cancel out in the calculation of the acceptance probability. Examples of this might include a uniform distribution or a normal distribution centered at the parameter upon which q is conditioned. Despite this cancellation, the specification of the proposal distribution is quite important as it determines the step size used when exploring the posterior distribution. It may be necessary to tune some hyperparameters of the proposal distribution, such as the variance in a normal distribution, to achieve some desired proportion of accepted proposals.

A.2.2 Reversible Jump Metropolis-Hastings Algorithm

The M-H algorithm is useful for updating a single parameter or a group of parameters (e.g., a vector of regression coefficients). In some cases it is desirable to have an algorithm that explores not only the posterior distribution of a set of parameters, but also the dimension of the parameters in question. This is particularly relevant for tree-based models such as BART, where the number of parameters required to adequately model the data is unknown. Every time a tree structure is modified, it can be viewed as changing the model itself to an entirely different model with the same number of, more, or less parameters, each with potentially new meaning. Green [44] and Godsill [43] introduce and expand upon a M-H

algorithm that is capable of exploring the posterior distribution of models, in addition to the posterior distribution of parameters. This approach is known as reversible jump Markov chain Monte Carlo.

Godsill [43] discusses the reversible jump sampler in terms of a Metropolis-Hastings proposal in the composite model space. Here, the composite model space refers to the set of all potential parameters and models in the target distribution. In this case, the model itself is also a parameter of target distribution $p(k, \theta)$ being sampled from. The proposal distribution q then proposes model a new model structure k' and new parameter θ'. The proposal distribution can be factored as in (A.2).

$$q(k', \theta' \mid k, \theta) = q_1(k' \mid k) q_2(\theta'_{k'} \mid \theta_k) p(\theta'_{-k'} \mid \theta'_{k'}, k') \tag{A.2}$$

Essentially, the structural component of the proposal from k to k' depends only on the current structure k, and the parameter proposal from θ_k to $\theta'_{k'}$ depends only on the current parameters θ_k used in the current model k. Finally, $p(\theta'_{-k'} \mid \theta'_{k'}, k')$ is the so-called *pseudo-prior* over the parameters in the composite model space not present within the proposed model k'. This last component is required because not all parameters in the composite model space are used in every model state.

The reversible jump M-H acceptance ratio is calculated as in (A.3).

$$\begin{aligned} r &= \frac{p(k', \theta' \mid \mathcal{D})}{p(k, \theta \mid \mathcal{D})} \times \frac{q(k, \theta \mid k', \theta')}{q(k', \theta' \mid k, \theta)} \\ &= \frac{p(k', \theta'_{k'} \mid \mathcal{D}) p(\theta'_{-k'} \mid \theta'_{k'}, k')}{p(k, \theta_k \mid \mathcal{D}) p(\theta_{-k} \mid \theta_k, k)} \times \frac{q_1(k \mid k') q_2(\theta_k \mid \theta'_{k'}) p(\theta_{-k} \mid \theta_k, k)}{q_1(k' \mid k) q_2(\theta'_{k'} \mid \theta_k) p(\theta'_{-k'} \mid \theta'_{k'}, k')} \\ &= \frac{p(k', \theta'_{k'} \mid \mathcal{D})}{p(k, \theta_k \mid \mathcal{D})} \times \frac{q_1(k \mid k') q_2(\theta_k \mid \theta'_{k'})}{q_1(k' \mid k) q_2(\theta'_{k'} \mid \theta_k)} \end{aligned} \tag{A.3}$$

Note that for the posterior (target) distribution in the numerator and denominator, the parameters which are excluded from the model are factored out into their respective pseudo-priors. The result is that all pseudo-priors cancel out in the ratio, so that only the posterior density (up to a constant) need to be computed in addition to the proposal densities. The algorithm is provided in Algorithm A.7.

Algorithm A.7 Reversible Jump Metropolis-Hastings Algorithm

1: Initialize $k^{(0)}, \theta^{(0)}$ (starting point)
2: **for** $s = 1$ to S **do**
3: Sample k' from proposal distribution $q_1(k'|k^{(s-1)})$
4: Sample θ' from proposal distribution $q_2(\theta^*|\theta^{(s-1)})$
5: Compute acceptance ratio:

$$r = \frac{p(k', \theta'_{k'} \mid \mathcal{D})}{p(k, \theta_k \mid \mathcal{D})} \times \frac{q_1(k \mid k')q_2(\theta_k \mid \theta'_{k'})}{q_1(k' \mid k)q_2(\theta'_{k'} \mid \theta_k)}$$

6: Set $k^{(s)} = \begin{cases} k' & \text{with probability } \min(1, r) \\ k^{(s-1)} & \text{otherwise} \end{cases}$
7: (Optionally) Sample $\theta^{(s)}$ targeting its full conditional distribution.
8: **end for**
9: **Return** $\{\theta^{(s)}\}_{s=1}^{S}$

The proposal distribution q_1 and q_2 are very important. q_1 is very structural in nature. For BART, this includes the probability of selecting a specific node to GROW, PRUNE or CHANGE, and the probability for selecting the new decision rule. Naturally, this depends on the current model state k. For q_2, the decision is also quite involved. For the reversible jump sampler in Chapter 3, a Laplace approximation is used. If q_2 does not effectively target the full conditional of θ, then an additional M-H step or adaptive rejection sampling (ARS) step might be required. If the full conditional distribution for θ is available in closed form, one could use this for q_2, in which case cancellation would occur with the posterior density resulting in a standard M-H proposal using the marginal distribution $p(k \mid \mathcal{D})$ as the target

distribution. This is exactly how the original BART algorithm samples model states (tree structures) (see (2.6)).

A.2.3 Gibbs Sampling

Gibbs sampling [41] can be viewed as a special case of the M-H algorithm, where the proposal distribution q is the full conditional distribution for the parameter θ. When this is the case, we can derive the acceptance ratio as in (A.4).

$$r = \frac{p(\theta')q(\theta \mid \theta')}{p(\theta)q(\theta' \mid \theta)} = \frac{\pi(\theta' \mid \mathcal{D})\pi(\theta \mid \mathcal{D})}{\pi(\theta \mid \mathcal{D})\pi(\theta' \mid \mathcal{D})} = 1 \tag{A.4}$$

This result means that the proposal is always accepted. This type of sampling is most efficient when the full conditional distribution is available in closed form such that $\pi(\theta \mid \mathcal{D})$ can be sampled from directly, as is the case when conjugate priors to the data likelihood are used. Gibbs sampling is particularly useful for sampling from a joint posterior distribution of multiple parameters. In this case, the parameters are updated sequentially, where the effective target distribution at each step is the full conditional distribution of the parameter being updated, given the current state of the other parameters in the model. While some burn-in period may still be necessary, proposal tuning is not required.

A.2.4 Adaptive Rejection Sampling

Gibbs samplers are generally the most efficient for obtaining samples from posterior distributions, but their implementation is only straightforward when the posterior distribution has a recognizable closed form that can be sampled from directly. When this isn't the case, the

M-H algorithm usually performs well in its place. However, the M-H algorithm requires a good proposal distribution, which may not be easy to specify for certain densities or may require extensive tuning. This is where sampling approaches such as adaptive rejection sampling (ARS) come into play. Such methods directly target the posterior distribution, like conjugate Gibbs samplers, but instead use known properties of the target density to construct an *envelope* from which samples may be obtained. ARS requires the target density $p(\theta)$ to be log-concave (i.e. $\frac{\partial^2}{\partial\theta^2}\log p(\theta) < 0$), continuous, and differentiable. For such densities, i.i.d. samples may be obtained using Algorithm A.8.

Algorithm A.8 Adaptive Rejection Sampling (ARS) Algorithm

1: **Input:** Target log-density $h(\theta) \propto \log p(\theta)$ and initial support points Θ.
2: Compute the initial piecewise linear upper hull $\bar{h}(\theta)$ and lower hull $\underline{h}(\theta)$ using $h(\theta)$ and (potentially) $h'(\theta)$ at each $\theta \in \Theta$.
3: Initialize: Sample set $\mathcal{S} = \emptyset$.
4: **while** $|\mathcal{S}| < S$ **do**
5: Sample θ' from the normalized exponential of the piecewise linear upper hull $\bar{h}(\theta)$.
6: Sample $U \sim$ Uniform(0,1).
7: **if** $U \leq \exp\left(\underline{h}(\theta') - \bar{h}(\theta')\right)$ **then**
8: Accept θ' and add it to $\mathcal{S}$.
9: **else**
10: **if** $U \leq \exp\left(h(\theta') - \bar{h}(\theta')\right)$ **then**
11: Accept θ^* and add it to $\mathcal{S}$.
12: **end if**
13: Add θ' to the set of support points Θ.
14: Reconstruct $\bar{h}(\theta)$ and $\underline{h}(\theta)$ using all support points in Θ.
15: **end if**
16: **end while**
17: **Return** $\mathcal{S}$.

Note how in Algorithm A.8, the log-density $h(\theta)$ need only be evaluated if the initial acceptance condition in step 7 fails (this evaluation occurs in step 10). Since this evaluation has already occurred, updating the lower and upper hulls is a quick adjustment. ARS is adaptive in the sense that the envelope approximation to h improves whenever the squeeze

sampling step fails to accept the proposed value.

There are multiple options for specifying the lower and upper hulls. If the derivative of $h(\theta)$ is known, then tangent lines to $h(\theta)$ may be drawn at each of the support points and connected in a piecewise fashion to from the upper hull $\bar{h}(\theta)$. Without using the derivative, secant lines between each pair of support points may be drawn and their portions above $h(\theta)$ can be connected to form a piecewise upper hull. In either case, the lower hull $\underline{h}(\theta)$ can be formed by forming a piecewise linear function between points $(\theta, h(\theta))$ for all $\theta \in \Theta$.

A.3 Bayesian Model Comparison

A.3.1 Widely Applicable Information Criterion

The widely applicable information criterion (WAIC) is commonly used as a metric for comparing Bayesian models [108, 40]. It is calculated as in (A.5)

$$\text{WAIC}(\boldsymbol{y}, \Theta) = -2\left(\text{lppd} - p_{\text{WAIC}}\right), \tag{A.5}$$

where

$$\text{lppd} = \sum_{i=1}^{N} \log \mathbb{E}_{\theta}\left[p(y_i \mid \theta)\right] \tag{A.6}$$

and

$$p_{\text{WAIC}} = \sum_{i=1}^{N} \text{Var}_{\,\theta}\left[\log p(y_i \mid \theta)\right] \tag{A.7}$$

Once S posterior samples have been generated, these quantities can be estimated with

(A.8) and (A.9)

$$\hat{\text{lppd}} = \sum_{i=1}^{N} \log \left(\frac{1}{S} \sum_{s=1} p\left(y_i \mid \theta^{(s)}\right) \right) \tag{A.8}$$

$$p_{\hat{\text{WAIC}}} = \sum_{i=1}^{N} \left[\frac{1}{S} \sum_{s=1} \left(\log p\left(y_i \mid \theta^{(s)}\right) \right)^2 - \left(\frac{1}{S} \sum_{s=1} \log p\left(y_i \mid \theta^{(s)}\right) \right)^2 \right] \tag{A.9}$$

The WAIC is asymptotically equivalent to leave-one-out cross-validation [40]. This is particularly useful when working with computationally expensive Bayesian models on large datasets because it removes the need to manually cross-validate.

Since the WAIC is a sum over all observation, on might also reference the average pointwise WAIC for individual observations or groups of observations to evaluate model performance on subsets of the data.

A.4 Conditional Autoregressive Models for Spatial Data

Originally devised by Besag [10], conditional autoregressive (CAR) models are widely used in spatial statistics to model areal spatially-referenced data. These models refer to those which use CAR prior distributions to model spatially dependent random effects such as those present in Chapters 4 and 5. The areal units may represent different political or geographical boundaries such as ZIP codes and counties. These models capture spatial dependence by specifying conditional distributions for each spatial unit given its neighbors, thus allowing for nearby regions to be more similar than regions far apart.

Let $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_N)^T$ represent spatially indexed random effects over N regions. Define $\mathbf{A}$ as the $N \times N$ spatial adjacency matrix with element A_{ij} representing the connection between regions i and j (usually 1 if i and j are neighbors, 0 otherwise). Diagonal elements

of $\mathbf{A}$ are usually set to 0 as well. We also define $\mathbf{D} = \text{diag}(A_{1+}, \ldots, A_{N+})$, where A_{i+} is the number of connected neighbors for area i.

Model Specification

The CAR model defines the distribution of ν_i conditional on its neighbors as in (A.10)

$$\nu_i \mid \nu_{-i} \sim \text{Normal}\left(\rho \frac{1}{A_{i+}} \sum_{j=i}^{N} A_{ij}\nu_j, \frac{\tau^2}{A_{i+}}\right), \tag{A.10}$$

where ρ is the spatial autocorrelation parameter and τ^2/A_{i+} is the conditional variance. This model suggests the conditional mean of ν_i is equal to the mean of its neighbors under perfect spatial autocorrelation ($\rho = 1$), while the conditional variance is proportional to the number of neighboring units.

The full joint distribution implied by the CAR model is multivariate normal:

$$\boldsymbol{\nu} \sim \text{Normal}_n(\mathbf{0}, \tau^2(\mathbf{D} - \rho\mathbf{A})^{-1}). \tag{A.11}$$

In this dissertation, we exclusively use *proper* CAR models with $0 < \rho < 1$. For $\rho = 1$, the joint distribution is improper because $\mathbf{D} - \mathbf{A}$ is not invertible. Models which fix $\rho = 1$ are sometimes called *intrinsic* CAR models, and while the joint distribution is not valid, these may still be used as a prior for random effects in Bayesian models.

A.5 Updating the BART Terminal Node Prior Variance

In the original BART model, the outcome y is scaled to lie in the interval $[-0.5, 0.5]$ prior to training the model. Recall the prior on terminal node parameters $\pi(\mu_{tl}) = \text{Normal}(\mu_\mu, \sigma_\mu^2)$. This in turn implies that the prior over $\sum_{t=1}^{T} \text{Tree}(\boldsymbol{x}; \mathcal{T}_t, \mathcal{M}_t)$ is $\text{Normal}(T\mu_\mu, T\sigma_\mu^2)$. To encourage the range of predictions to fall within k standard deviations, we specify the prior by solving for σ_μ^2 (and μ_μ) using the following equations:

$$T\mu_\mu - 2k\sigma_\mu\sqrt{T} = y_{min}$$

$$T\mu_\mu + 2k\sigma_\mu\sqrt{T} = y_{max}$$

Solving for μ_μ and σ_μ gives:

$$\mu_\mu = \frac{y_{min} + y_{max}}{2T} \qquad \text{and} \qquad \sigma_\mu = \frac{y_{max} - y_{min}}{2k\sqrt{T}}$$

In these equations, k is typically set to 2, indicating prior belief that 95% of predictions will fall within the interval (y_{min}, y_{max}). When y_{min} and y_{max} are known a priori to be, say, -0.5 and 0.5, then the result simplifies to $\mu_\mu = 0$ and $\sigma_\mu = \frac{1}{2k\sqrt{T}}$.

The challenge with using BART as a single component of more complicated models, such as the ones presented in this dissertation, is that there is no observed y to inform these calculations. In Chapter 3, the true range of heterogeneous exposure effects is unknown and in Chapter 4, the BART model is trained on latent (unobserved) variables, making it impossible to determine the range of values ahead of time. In these settings, it is beneficial to introduce a hyperprior distribution over σ_μ^2. The goal for the rest of this section is to describe options

for estimating σ_μ. While we ended up choosing the method in Section A.5.2, we provide all strategies that were tried in case they become useful for future BART implementations.

A.5.1 Inverse-Gamma Approach

$$\sigma_\mu^2 \sim IG(\alpha, \beta)$$

A natural first step is specifying an inverse-gamma hyperprior for σ_μ^2 since it is conditionally conjugate to the prior for σ_μ^2. This results in the following:

$$\left[\sigma_\mu^2 \mid -\right] \sim IG\left(\alpha + \frac{\sum_{t=1}^{T} \mid \mathcal{L}(\mathcal{T}_t) \mid}{2}, \beta + \frac{\sum_{t=1}^{T} \sum_{l \in \mathcal{L}(\mathcal{T}_t)} (\mu_{tl} - \mu_\mu)^2}{2}\right)$$

where $\mid \mathcal{L}(\mathcal{T}_t) \mid$ is the number of leaf nodes in $\mathcal{T}_t$. As it turns out, this specification can be very sensitive to how the hyperparameters α and β are specified.

A.5.2 Half-Cauchy Approach

$$\sigma_\mu \sim \mathcal{C}_+\left(0, \frac{k}{\sqrt{T}}\right)$$

This approach was suggested in Linero [71]. It is also what is used in Linero [70], and what we use for Chapters 3 and 4. The prior has median $\frac{k}{\sqrt{T}}$. The general strategy is as follows:

- First, we make a proposal on the precision (τ) scale. Assuming a flat prior over τ, we

propose from the full conditional

$$\tau \sim Ga\left(1 + \frac{\sum_{t=1}^{T} \mid \mathcal{L}(\mathcal{T}_t) \mid}{2}, \frac{\sum_{t=1}^{T}\sum_{l\in\mathcal{L}(\mathcal{T}_t)}(\mu_{tl} - \mu_\mu)^2}{2}\right)$$

- Since the proposal was done on the precision scale, we must adjust the prior to also be on the precision scale. Since $\sigma_\mu = g(\tau)$ where $g(x) = x^{-1/2}$, by the method of transformation we have:

$$\pi(\tau) = \pi(g(\tau)) \mid g'(\tau) \mid = \mathcal{C}_+\left(\sigma_\mu \mid 0, \frac{k}{\sqrt{T}}\right)\frac{1}{2}\sigma_\mu^3$$

- The M-H ratio is given by:

$$r = \frac{\mathcal{C}_+\left(\sigma'_\mu \mid 0, \frac{k}{\sqrt{T}}\right)(\sigma'_\mu)^3}{\mathcal{C}_+\left(\sigma_\mu \mid 0, \frac{k}{\sqrt{T}}\right)\sigma_\mu^3}$$

thanks to perfect cancellation between the likelihoods and proposal densities.

A.5.3 Marginal Half-Cauchy Approach

Consider the hierarchical representation:

$$\left[\sigma_\mu^2 \mid \sigma_0\right] \sim IG\left(\frac{1}{2}, \frac{1}{\sigma_0}\right)$$

$$\left[\sigma_0\right] \sim IG\left(\frac{1}{2}, \frac{1}{C^2}\right)$$

It follows that marginally, $\sigma_\mu \sim \mathcal{C}_+(0, C)$. In agreement with the direct half-Cauchy prior

in the previous section, it makes sense to set $C = \frac{k}{\sqrt{T}}$. The benefit of this specification is that a convenient two-stage Gibbs sampler is available.

$$[\sigma_0 \mid -] \sim IG\left(1, \frac{1}{C^2} + \frac{1}{\sigma_\mu^2}\right)$$

$$\left[\sigma_\mu^2 \mid -\right] \sim IG\left(\frac{\sum_{t=1}^{T} \mid \mathcal{L}(\mathcal{T}_t) \mid +1}{2}, \frac{\sum_{t=1}^{T}\sum_{l\in\mathcal{L}(\mathcal{T}_t)}(\mu_{tl} - \mu_\mu)^2}{2} + \frac{1}{\sigma_0}\right)$$

A.5.4 Horseshoe Approach

The horseshoe prior applies both global and local (tree-specific) shrinkage, so the prior on terminal nodes can be expressed as:

$$\mu_{tl} \sim \text{Normal}(\mu_\mu, \omega^2\tau_t^2)$$

$$[\omega] \sim \mathcal{C}_+(0, C)$$

$$[\tau_t] \sim \mathcal{C}_+(0, 1)$$

We can apply the same method above to obtain a Gibbs sampler for the updates. The priors can be rewritten as follows:

$$\left[\omega^2 \mid \omega_0\right] \sim IG\left(\frac{1}{2}, \frac{1}{\omega_0}\right) \qquad [\omega_0] \sim IG\left(\frac{1}{2}, \frac{1}{C^2}\right)$$

$$\left[\tau_t^2 \mid \tau_{t0}\right] \sim IG\left(\frac{1}{2}, \frac{1}{\tau_{t0}}\right) \qquad [\tau_{t0}] \sim IG\left(\frac{1}{2}, 1\right)$$

Mork and Wilson [79] set $C = 1$, however their model differs from the model in Chapter

3 in that there is also residual variance to consider. Without this consideration, it may once again prove beneficial to set the value to $\frac{k}{\sqrt{T}}$. The updates proceed from the full conditional distributions as follows:

$$[\tau_{t0} \mid -] \sim IG\left(1, 1 + \frac{1}{\tau_t^2}\right)$$

$$[\tau_t \mid -] \sim IG\left(\frac{\mid \mathcal{L}(\mathcal{T}_t) \mid +1}{2}, \frac{\sum_{l \in \mathcal{L}(\mathcal{T}_t)} (\mu_{tl} - \mu_\mu)^2}{2\omega^2} + \frac{1}{\tau_{t0}}\right)$$

$$[\omega_0 \mid -] \sim IG\left(1, \frac{1}{C^2} + \frac{1}{\omega^2}\right)$$

$$[\omega \mid -] \sim IG\left(\frac{\sum_{t=1}^{T} \mid \mathcal{L}(\mathcal{T}_t) \mid +1}{2}, \frac{\sum_{t=1}^{T} \sum_{l \in \mathcal{L}(\mathcal{T}_t)} (\mu_{tl} - \mu_\mu)^2 / \tau_t^2}{2} + \frac{1}{\omega_0}\right)$$

These parameters can all be updated after the other updates of tree structures and parameters have been completed.

A.5.5 Inverse-Gamma Approach (k)

Suppose $\sigma_\mu^2 = \left(\frac{k}{\sqrt{T}}\right)$. Since we know that σ_μ^2 should probably be scaled by T, we can instead place an inverse-gamma prior on k. This approach is otherwise identical to that presented in Section A.5.1. If the prior is

$$k^2 \sim IG(\alpha, \beta),$$

the full conditional is given by:

$$\left[k^2 \mid -\right] \sim IG\left(\alpha + \frac{\sum_{t=1}^{T} \mid \mathcal{L}(\mathcal{T}_t) \mid}{2}, \beta + \frac{T\sum_{t=1}^{T}\sum_{l\in\mathcal{L}(\mathcal{T}_t)}(\mu_{tl}-\mu_{\mu})^2}{2}\right)$$

After updating k, we can set $\sigma_{\mu}^2 = \left(\frac{k}{\sqrt{T}}\right)$. α and β can be chosen to encourage values of k to be in the interval $[0.1, 1]$ as suggested by Linero [71].

A.5.6 Marginal Half-Cauchy Approach (k)

Similar to the previous section, we can also specify a marginal half-Cauchy prior for k. The hierarchical prior of the form:

$$\left[k^2 \mid k_0\right] \sim IG\left(\frac{1}{2}, \frac{1}{k_0}\right) [k_0] \sim IG\left(\frac{1}{2}, \frac{1}{C^2}\right)$$

implies $k \sim \mathcal{C}_+(0, C)$. The full conditional distributions are given by:

$$[k_0 \mid -] \sim IG\left(1, \frac{1}{C^2} + \frac{1}{k^2}\right)$$

$$\left[k^2 \mid -\right] \sim IG\left(\frac{\sum_{t=1}^{T} \mid \mathcal{L}(\mathcal{T}_t) \mid +1}{2}, \frac{T\sum_{t=1}^{T}\sum_{l\in\mathcal{L}(\mathcal{T}_t)}(\mu_{tl}-\mu_{\mu})^2}{2} + \frac{1}{k_0}\right)$$

Bibliography

[1] Agency for Healthcare Research and Quality, Rockville, MD. HCUPnet, Healthcare Cost and Utilization Project, 2021. `https://datatools.ahrq.gov/hcupnet`.

[2] Brooke A Alhanti, Howard H Chang, Andrea Winquist, James A Mulholland, Lyndsey A Darrow, and Stefanie Ebelt Sarnat. Ambient air pollution and emergency department visits for asthma: a multi-city assessment of effect modification by age. *Journal of Exposure Science & Environmental Epidemiology*, 26(2):180–188, 2016. doi:10.1038/jes.2015.57.

[3] Alzheimer's Association. 2023 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 19(4):1598–1695, 2023. doi:10.1002/alz.13016.

[4] Joseph Antonelli, Georgia Papadogeorgou, and Francesca Dominici. Causal inference in high dimensions: A marriage between Bayesian modeling and good frequentist properties. *Biometrics*, 78(1):100–114, 2022. doi:10.1111/biom.13417.

[5] Joseph Antonelli, Ander Wilson, and Brent A Coull. Multiple exposure distributed lag models with variable selection. *Biostatistics*, 25(1):1–19, 2023. doi:10.1093/biostatistics/kxac038.

[6] Daniel W. Apley and Jingyu Zhu. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1059–1086, 2020. doi:10.1111/rssb.12377.

[7] Ben G Armstrong. Fixed Factors that Modify the Effects of Time-Varying Factors: Applying the Case-Only Approach:. *Epidemiology*, 14(4):467–472, 2003. doi:10.1097/01.ede.0000071408.39011.99.

[8] Adrian G. Barnett, Susana Sans, Veikko Salomaa, Kari Kuulasmaa, and Annette J. Dobson. The effect of temperature on systolic blood pressure. *Blood Pressure Monitoring*, 12(3):195–203, 2007. doi:10.1097/MBP.0b013e3280b083f4.

[9] Moritz Berger and Gerhard Tutz. Tree-Structured Clustering in Fixed Effects Models. *Journal of Computational and Graphical Statistics*, 27(2):380–392, 2018. doi:10.1080/10618600.2017.1371030.

[10] Julian Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 36(2):192–225, 1974. doi:10.1111/j.2517-6161.1974.tb00999.x.

[11] Jianzhao Bi, Rohan R. D'Souza, Shannon Moss, Niru Senthilkumar, Armistead G. Russell, Noah C. Scovronick, Howard H. Chang, and Stefanie Ebelt. Acute Effects of Ambient Air Pollution on Asthma Emergency Department Visits in Ten U.S. States. *Environmental Health Perspectives*, 131(4):047003, 2023. doi:10.1289/EHP11661.

[12] Justin Bleich and Adam Kapelner. Bayesian Additive Regression Trees With Parametric Models of Heteroskedasticity, 2014. doi:10.48550/arXiv.1402.5397.

[13] Jennifer F. Bobb, Linda Valeri, Birgit Claus Henn, David C. Christiani, Robert O. Wright, Maitreyi Mazumdar, John J. Godleski, and Brent A. Coull. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, 16(3):493–508, 2015. doi:10.1093/biostatistics/kxu058.

[14] Vinicius Bonato, Veerabhadran Baladandayuthapani, Bradley M. Broom, Erik P. Sulman, Kenneth D. Aldape, and Kim-Anh Do. Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics*, 27(3):359–367, 2011. doi:10.1093/bioinformatics/btq660.

[15] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification And Regression Trees*. Routledge, 1 edition, 2017. ISBN 978-1-315-13947-0. doi:10.1201/9781315139470.

[16] R. P. Brent. *Algorithms for minimization without derivatives*. Prentice-Hall series in automatic computation. Prentice-Hall, Englewood Cliffs, N.J, 1972. ISBN 978-0-13-022335-7.

[17] Alberto Caron, Gianluca Baio, and Ioanna Manolopoulou. Shrinkage Bayesian Causal Forests for Heterogeneous Treatment Effects Estimation. *Journal of Computational and Graphical Statistics*, 31(4):1202–1214, 2022. doi:10.1080/10618600.2022.2067549.

[18] Eduardo Carracedo-Martínez, Margarita Taracido, Aurelio Tobias, Marc Saez, and Adolfo Figueiras. Case-Crossover Analysis of Air Pollution Health Effects: A Systematic Review of Methodology and Application. *Environmental Health Perspectives*, 118(8): 1173–1182, 2010. doi:10.1289/ehp.0901485.

[19] Caroline Carrico, Chris Gennings, David C. Wheeler, and Pam Factor-Litvak. Characterization of Weighted Quantile Sum Regression for Highly Correlated Data in a Risk Analysis Setting. *Journal of Agricultural, Biological, and Environmental Statistics*, 20 (1):100–120, 2015. doi:10.1007/s13253-014-0180-3.

[20] Emilio Casetti. Generating Models by the Expansion Method: Applications to Geographical Research*. *Geographical Analysis*, 4(1):81–91, 1972. doi:10.1111/j.1538-4632.1972.tb00458.x.

[21] Centers for Disease Control and Prevention. Most Recent National Asthma Data, 2021. `https://www.cdc.gov/asthma/most_recent_data.htm`.

[22] Howard H. Chang, Joshua L. Warren, Lnydsey A. Darrow, Brian J. Reich, and Lance A. Waller. Assessment of critical exposure and outcome windows in time-to-event analysis with application to air pollution and preterm birth study. *Biostatistics*, 16(3):509–521, 2015. doi:10.1093/biostatistics/kxu060.

[23] Song Chi, Chong Wang, Teng Jiang, Xi-Chen Zhu, Jin-Tai Yu, and Lan Tan. The Prevalence of Depression in Alzheimer's Disease: A Systematic Review and Meta-Analysis. *Current Alzheimer Research*, 12(2):189–198, 2015. doi:10.2174/1567205012666150204124310.

[24] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bayesian CART Model Search. *Journal of the American Statistical Association*, 93(443):935–948, 1998. doi:10.1080/01621459.1998.10473750.

[25] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian

additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010. doi:10.1214/09-AOAS285.

[26] D R Culqui, C Linares, C Ortiz, R Carmona, and J Díaz. Association between environmental factors and emergency hospital admissions due to Alzheimer's disease in Madrid. *Science of The Total Environment*, 592:451–457, 2017. doi:10.1016/j.scitotenv.2017.03.089.

[27] Michael J. Daniels, Antonio Linero, and Jason Roy. *Bayesian Nonparametrics for Causal Inference and Missing Data*. Chapman and Hall/CRC, Boca Raton, 1 edition, 2023. ISBN 978-0-429-32422-2. doi:10.1201/9780429324222.

[28] Lyndsey A. Darrow, Mitchel Klein, Matthew J. Strickland, James A. Mulholland, and Paige E. Tolbert. Ambient Air Pollution and Birth Weight in Full-Term Infants in Atlanta, 1994–2004. *Environmental Health Perspectives*, 119(5):731–737, 2011. doi:10.1289/ehp.1002785.

[29] Sameer K. Deshpande. flexBART: Flexible Bayesian regression trees with categorical predictors. *Journal of Computational and Graphical Statistics*, pages 1–18, 2024. doi:10.1080/10618600.2024.2431072.

[30] Sameer K. Deshpande, Ray Bai, Cecilia Balocchi, Jennifer E. Starling, and Jordan Weiss. VCBART: Bayesian Trees for Varying Coefficients. *Bayesian Analysis*, pages 1–28, 2024. doi:10.1214/24-BA1470.

[31] Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated

versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition. *Statistical Science*, 34(1), 2019. doi:10.1214/18-STS667.

[32] Danielle M. Ely and Anne K. Driscoll. Infant mortality in the United States, 2022: Data from the period linked birth/infant death file. Technical report, National Center for Health Statistics (U.S.), Hyattsville, MD, 2024. `https://stacks.cdc.gov/view/cdc/157006`.

[33] Andrew O. Finley. Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution*, 2(2):143–154, 2011. doi:10.1111/j.2041-210X.2010.00060.x.

[34] A. Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. J. Wiley, Chichester, 2002. ISBN 978-0-471-49616-8.

[35] Jerome H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–67, 1991. doi:10.1214/aos/1176347963.

[36] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi:10.1214/aos/1013203451.

[37] Thomas Fritze. The Effect of Heat and Cold Waves on the Mortality of Persons with Dementia in Germany. *Sustainability*, 12(9):3664, 2020. doi:10.3390/su12093664.

[38] A. Gasparrini, B. Armstrong, and M. G. Kenward. Distributed lag non-linear models. *Statistics in Medicine*, 29(21):2224–2234, 2010. doi:10.1002/sim.3940.

[39] Alan E Gelfand, Hyon-Jung Kim, C. F Sirmans, and Sudipto Banerjee. Spatial Modeling With Spatially Varying Coefficient Processes. *Journal of the American Statistical Association*, 98(462):387–396, 2003. doi:10.1198/016214503000170.

[40] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014. doi:10.1007/s11222-013-9416-2.

[41] Stuart Geman and Donald Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984. doi:10.1109/TPAMI.1984.4767596.

[42] W. R. Gilks and P. Wild. Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics*, 41(2):337–348, 1992. doi:10.2307/2347565.

[43] Simon J Godsill. On the Relationship Between Markov chain Monte Carlo Methods for Model Uncertainty. *Journal of Computational and Graphical Statistics*, 10(2):230–248, 2001. doi:10.1198/10618600152627924.

[44] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995. doi:10.1093/biomet/82.4.711.

[45] Yuming Guo, Adrian G Barnett, Xiaochuan Pan, Weiwei Yu, and Shilu Tong. The Impact of Temperature on Mortality in Tianjin, China: A Case-Crossover Design with a Distributed Lag Nonlinear Model. *Environmental Health Perspectives*, 119(12): 1719–1725, 2011. doi:10.1289/ehp.1103598.

[46] P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, 15(3):965–1056, 2020. doi:10.1214/19-BA1195.

[47] A. L Hansen, P. Bi, P. Ryan, M. Nitschke, D. Pisaniello, and G. Tucker. The effect of heat waves on hospital admissions for renal disease in a temperate city of Australia. *International Journal of Epidemiology*, 37(6):1359–1365, 2008. doi:10.1093/ije/dyn165.

[48] David G Harper, Ladislav Volicer, Edward G Stopa, Ann C McKee, and et al. Disturbance of Endogenous Circadian Rhythm in Aging and Alzheimer Disease. *The American Journal of Geriatric Psychiatry*, 13(5):359–368, 2005.

[49] Trevor Hastie and Robert Tibshirani. Bayesian backfitting (with comments and a rejoinder by the authors. *Statistical Science*, 15(3):196–223, 2000. doi:10.1214/ss/1009212815.

[50] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. doi:10.1093/biomet/57.1.97.

[51] Jennifer Hill, Antonio Linero, and Jared Murray. Bayesian Additive Regression Trees: A Review and Look Forward. *Annual Review of Statistics and Its Application*, 7(1): 251–278, 2020. doi:10.1146/annurev-statistics-031219-041110.

[52] Jennifer L. Hill. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi:10.1198/jcgs.2010.08162.

[53] Holly Janes, Lianne Sheppard, and Thomas Lumley. Case-Crossover Analyses of Air Pollution Exposure Data: Referent Selection Strategies and Their Implications for Bias. *Epidemiology*, 16(6):717–726, 2005. doi:10.1097/01.ede.0000181315.18836.9d.

[54] Holly Janes, Lianne Sheppard, and Thomas Lumley. Overlap bias in the case-crossover design, with application to air pollution exposures. *Statistics in Medicine*, 24(2):285–300, 2005. doi:10.1002/sim.1889.

[55] Meng Ji, Daniel S Cohan, and Michelle L Bell. Meta-analysis of the Association between Short-Term Exposure to Ambient Ozone and Respiratory Hospital Admissions. *Environmental Research Letters*, 6(2):024006, 2011. doi:https://doi.org/10.1088/1748-9326/6/2/024006.

[56] Richard J. Johnson, Laura G. Sánchez-Lozada, Lee S. Newman, Miguel A. Lanaspa, Henry F. Diaz, Jay Lemery, Bernardo Rodriguez-Iturbe, Dean R. Tolan, Jaime Butler-Dawson, Yuka Sato, Gabriela Garcia, Ana Andres Hernando, and Carlos A. Roncal-Jimenez. Climate Change and the Kidney. *Annals of Nutrition and Metabolism*, 74 (Suppl. 3):38–44, 2019. doi:10.1159/000500344.

[57] Alexander Keil. qgcomp: Quantile G-Computation, 2019. doi:10.32614/CRAN.package.qgcomp. `https://CRAN.R-project.org/package=qgcomp`.

[58] Alexander P. Keil, Jessie P. Buckley, Katie M. O'Brien, Kelly K. Ferguson, Shanshan Zhao, and Alexandra J. White. A Quantile-Based g-Computation Approach to Addressing the Effects of Exposure Mixtures. *Environmental Health Perspectives*, 128(4): 047004, 2020. doi:10.1289/EHP5838.

[59] Andis Klegeris, Michael Schulzer, David G. Harper, and Patrick L. McGeer. Increase in Core Body Temperature of Alzheimer's Disease Patients as a Possible Indicator

of Chronic Neuroinflammation: A Meta-Analysis. *Gerontology*, 53(1):7–11, 2007. doi:10.1159/000095386.

[60] Kim Knowlton, Miriam Rotkin-Ellman, Galatea King, Helene G. Margolis, Daniel Smith, Gina Solomon, Roger Trent, and Paul English. The 2006 California Heat Wave: Impacts on Hospitalizations and Emergency Department Visits. *Environmental Health Perspectives*, 117(1):61–67, 2009. doi:10.1289/ehp.11594.

[61] Dirga Kumar Lamichhane, Jong-Han Leem, Ji-Young Lee, and Hwan-Cheol Kim. A meta-analysis of exposure to particulate matter and adverse birth outcomes. *Environmental Health and Toxicology*, 30:e2015011, 2015. doi:10.5620/eht.e2015011.

[62] Brooke L. Lappe, Stefanie Ebelt, Rohan R. D'Souza, Arie Manangan, Claudia Brown, Shubhayu Saha, Drew Harris, Howard H. Chang, Adam Sole, and Noah Scovronick. Pollen and asthma morbidity in Atlanta: A 26-year time-series study. *Environment International*, 177:107998, 2023. doi:10.1016/j.envint.2023.107998.

[63] Junho Lee, Ronald E. Gangnon, and Jun Zhu. Cluster detection of spatial regression coefficients. *Statistics in Medicine*, 36(7):1118–1133, 2017. doi:10.1002/sim.7172.

[64] Junho Lee, Ying Sun, and Howard H. Chang. Spatial cluster detection of regression coefficients in a mixed-effects model. *Environmetrics*, 31(2):e2578, 2020. doi:10.1002/env.2578.

[65] Furong Li and Huiyan Sang. Spatial Homogeneity Pursuit of Regression Coefficients for Large Datasets. *Journal of the American Statistical Association*, 114(527):1050–1062, 2019. doi:10.1080/01621459.2018.1529595.

[66] Xiangyu Li, Shuqiong Huang, Anqi Jiao, Xuhao Yang, Junfeng Yun, Yuxin Wang, Xiaowei Xue, Yuanyuan Chu, Feifei Liu, Yisi Liu, Meng Ren, Xi Chen, Na Li, Yuanan Lu, Zongfu Mao, Liqiao Tian, and Hao Xiang. Association between ambient fine particulate matter and preterm birth or term low birth weight: An updated systematic review and meta-analysis. *Environmental Pollution*, 227:596–605, 2017. doi:10.1016/j.envpol.2017.03.055.

[67] Fangzheng Lin, Yanlin Tang, Huichen Zhu, and Zhongyi Zhu. Spatially clustered varying coefficient model. *Journal of Multivariate Analysis*, 192:105023, 2022. doi:10.1016/j.jmva.2022.105023.

[68] Shao Lin, Ming Luo, Randi J. Walker, Xiu Liu, Syni-An Hwang, and Robert Chinery. Extreme High Temperatures and Hospital Admissions for Respiratory and Cardiovascular Diseases. *Epidemiology*, 20(5):738–746, 2009. doi:10.1097/EDE.0b013e3181ad5522.

[69] Antonio R. Linero. A review of tree-based Bayesian methods. *Communications for Statistical Applications and Methods*, 24(6):543–559, 2017. doi:10.29220/CSAM.2017.24.6.543.

[70] Antonio R. Linero. Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection. *Journal of the American Statistical Association*, 113(522):626–636, 2018. doi:10.1080/01621459.2016.1264957.

[71] Antonio R. Linero. Generalized Bayesian Additive Regression Trees Models: Beyond Conditional Conjugacy. *Journal of the American Statistical Association*, pages 1–14, 2024. doi:10.1080/01621459.2024.2337156.

[72] Antonio R. Linero and Joseph L. Antonelli. The how and why of Bayesian nonparametric causal inference. *WIREs Computational Statistics*, 15(1), 2023. doi:10.1002/wics.1583.

[73] Antonio R. Linero and Yun Yang. Bayesian Regression Tree Ensembles that Adapt to Smoothness and Sparsity. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(5):1087–1110, 2018. doi:10.1111/rssb.12293.

[74] Antonio R. Linero, Debajyoti Sinha, and Stuart R. Lipsitz. Semiparametric mixed-scale models using shared Bayesian forests. *Biometrics*, 76(1):131–144, 2020. doi:10.1111/biom.13107.

[75] Jingwen Liu, Blesson M. Varghese, Alana Hansen, Matthew A. Borg, Ying Zhang, Timothy Driscoll, Geoffrey Morgan, Keith Dear, Michelle Gourley, Anthony Capon, and Peng Bi. Hot weather as a risk factor for kidney disease outcomes: A systematic review and meta-analysis of epidemiological evidence. *Science of The Total Environment*, 801: 149806, 2021. doi:10.1016/j.scitotenv.2021.149806.

[76] Jaime Madrigano, Kazuhiko Ito, Sarah Johnson, Patrick L. Kinney, and Thomas Matte. A Case-Only Study of Vulnerability to Heat Wave–Related Mortality in New York City (2000–2011). *Environmental Health Perspectives*, 123(7):672–678, 2015. doi:10.1289/ehp.1408178.

[77] Mateus Maia, Keefe Murphy, and Andrew C. Parnell. GP-BART: A novel Bayesian additive regression trees approach using Gaussian processes. *Computational Statistics & Data Analysis*, 190:107858, 2024. doi:10.1016/j.csda.2023.107858.

[78] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H.

Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. doi:10.1063/1.1699114.

[79] Daniel Mork and Ander Wilson. Treed distributed lag nonlinear models. *Biostatistics*, 23(3):754–771, 2022. doi:10.1093/biostatistics/kxaa051.

[80] Daniel Mork and Ander Wilson. Estimating perinatal critical windows of susceptibility to environmental mixtures via structured Bayesian regression tree pairs. *Biometrics*, 79(1):449–461, 2023. doi:10.1111/biom.13568.

[81] Daniel Mork, Marianthi-Anna Kioumourtzoglou, Marc Weisskopf, Brent A. Coull, and Ander Wilson. Heterogeneous Distributed Lag Models to Estimate Personalized Effects of Maternal Exposures to Air Pollution. *Journal of the American Statistical Association*, pages 1–13, 2023. doi:10.1080/01621459.2023.2258595.

[82] Jared S. Murray. Log-Linear Bayesian Additive Regression Trees for Multinomial Logistic and Count Regression Models. *Journal of the American Statistical Association*, 116(534):756–769, 2021. doi:10.1080/01621459.2020.1813587.

[83] Fedelis Mutiso, Hong Li, John L Pearce, Sara E Benjamin-Neelon, Noel T Mueller, and Brian Neelon. Bayesian kernel machine regression for count data: modelling the association between social vulnerability and COVID-19 deaths in South Carolina. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(1):257–274, 2024. doi:10.1093/jrsssc/qlad094.

[84] Radford M. Neal. Slice sampling. *The Annals of Statistics*, 31(3), 2003. doi:10.1214/aos/1056562461.

[85] Arman Oganisian and Jason A. Roy. A practical introduction to Bayesian estimation of causal effects: Parametric and nonparametric approaches. *Statistics in Medicine*, 40 (2):518–551, 2021. doi:10.1002/sim.8761.

[86] Cassandra R O'Lenick, Andrea Winquist, James A Mulholland, Mariel D Friberg, Howard H Chang, Michael R Kramer, Lyndsey A Darrow, and Stefanie Ebelt Sarnat. Assessment of neighbourhood-level socioeconomic status as a modifier of air pollution–asthma associations among children in Atlanta. *Journal of Epidemiology and Community Health*, 71(2):129–136, 2017. doi:10.1136/jech-2015-206530.

[87] Michelle Osterman, Brady Hamilton, Joyce Martin, Anne Driscoll, and Claudia Valenzuela. Births: Final Data for 2022. Technical report, National Center for Health Statistics (U.S.), Hyattsville, MD, 2024. `https://stacks.cdc.gov/view/cdc/145588`.

[88] Jonathan Pillow and James Scott. Fully Bayesian inference for neural models with negative-binomial spiking. In F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. `https://proceedings.neurips.cc/paper_files/paper/2012/file/b55ec28c52d5f6205684a473a2193564-Paper.pdf`.

[89] Nicholas G. Polson, James G. Scott, and Jesse Windle. Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013. doi:10.1080/01621459.2013.829001.

[90] Matthew T. Pratola. Efficient Metropolis–Hastings Proposal Mechanisms for Bayesian

Regression Tree Models. *Bayesian Analysis*, 11(3):885–911, 2016. doi:10.1214/16-BA999.

[91] Andrew Satlin, Ladislav Volicer, Edward G. Stopa, and David Harper. Circadian locomotor activity and core-body temperature rhythms in Alzheimer's disease. *Neurobiology of Aging*, 16(5):765–771, 1995. doi:10.1016/0197-4580(95)00059-N.

[92] Joel Schwartz. Who is Sensitive to Extremes of Temperature?: A Case-Only Analysis. *Epidemiology*, 16(1):67–72, 2005. doi:10.1097/01.ede.0000147114.25957.71.

[93] Niru Senthilkumar, Mark Gilfether, Howard H. Chang, Armistead G. Russell, and James Mulholland. Using land use variable information and a random forest approach to correct spatial mean bias in fused CMAQ fields for particulate and gas species. *Atmospheric Environment*, 274:118982, 2022. doi:10.1016/j.atmosenv.2022.118982.

[94] Toki Sherbakov, Brian Malig, Kristen Guirguis, Alexander Gershunov, and Rupa Basu. Ambient temperature and added heat wave effects on hospitalizations in California from 1999 to 2009. *Environmental Research*, 160:83–90, 2018. doi:10.1016/j.envres.2017.08.052.

[95] D. Sonntag. Important new values of the physical constants of 1986, vapor pressure formulations based on the ITS-90, and psychrometer formulae. *Z. Meteorol.*, 70:340–344, 1990.

[96] Rodney A. Sparapani, Brent R. Logan, Robert E. McCulloch, and Purushottam W. Laud. Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Statistics in Medicine*, 35(16):2741–2753, 2016. doi:10.1002/sim.6893.

[97] David M. Stieb, Li Chen, Maysoon Eshoul, and Stan Judek. Ambient air pollution, birth weight and preterm birth: A systematic review and meta-analysis. *Environmental Research*, 117:100–111, 2012. doi:10.1016/j.envres.2012.05.007.

[98] Matthew J. Strickland, Lyndsey A. Darrow, Mitchel Klein, W. Dana Flanders, Jeremy A. Sarnat, Lance A. Waller, Stefanie E. Sarnat, James A. Mulholland, and Paige E. Tolbert. Short-term Associations between Ambient Air Pollutants and Pediatric Asthma Emergency Department Visits. *American Journal of Respiratory and Critical Care Medicine*, 182(3):307–316, 2010. doi:10.1164/rccm.200908-1201oc.

[99] Matthew J. Strickland, Hua Hao, Xuefei Hu, Howard H. Chang, Lyndsey A. Darrow, and Yang Liu. Pediatric Emergency Visits and Short-Term Changes in PM $_{2.5}$ Concentrations in the U.S. State of Georgia. *Environmental Health Perspectives*, 124(5):690–696, 2016. doi:10.1289/ehp.1509856.

[100] Matthew J. Strickland, Ying Lin, Lyndsey A. Darrow, Joshua L. Warren, James A. Mulholland, and Howard H. Chang. Associations Between Ambient Air Pollutant Concentrations and Birth Weight: A Quantile Regression Analysis. *Epidemiology*, 30 (5):624–632, 2019. doi:10.1097/EDE.0000000000001038.

[101] Xiaoli Sun, Xiping Luo, Chunmei Zhao, Bo Zhang, Jun Tao, Zuyao Yang, Wenjun Ma, and Tao Liu. The associations between birth weight and exposure to fine particulate matter (PM2.5) and its chemical constituents during pregnancy: A meta-analysis. *Environmental Pollution*, 211:38–47, 2016. doi:10.1016/j.envpol.2015.12.022.

[102] Yaoyuan Vincent Tan and Jason Roy. Bayesian additive regression trees and the General BART model. *Statistics in Medicine*, 38(25):5048–5069, 2019. doi:10.1002/sim.8347.

[103] Terry Therneau and Beth Atkinson. rpart: Recursive Partitioning and Regression Trees, 1999. doi:10.32614/CRAN.package.rpart. `https://CRAN.R-project.org/package=rpart`.

[104] M.M. Thornton, R. Shrestha, Y. Wei, P.E. Thornton, and S-C. Kao. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4 R1, 2022. doi:10.3334/ORNLDAAC/2129. `https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=2129`.

[105] J. Van Hoof, H.S.M. Kort, J.L.M. Hensen, M.S.H. Duijnstee, and P.G.S. Rutten. Thermal comfort and the integrated design of homes for older people with dementia. *Building and Environment*, 45(2):358–370, 2010. doi:10.1016/j.buildenv.2009.06.013.

[106] Ladislav Volicer, David G. Harper, Barbara C. Manning, Rachel Goldstein, and Andrew Satlin. Sundowning and Circadian Rhythms in Alzheimer's Disease. *American Journal of Psychiatry*, 158(5):704–711, 2001. doi:10.1176/appi.ajp.158.5.704.

[107] Joshua L. Warren, Thomas J. Luben, and Howard H. Chang. 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, 2020. doi:10.1111/rssc.12407.

[108] Sumio Watanabe and Manfred Opper. Asymptotic Equivalence of Bayes Cross Valida-

tion and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11(116):3571–3594, 2010.

[109] Yaguang Wei, Yan Wang, Cheng-Kuan Lin, Kanhua Yin, Jiabei Yang, Liuhua Shi, Longxiang Li, Antonella Zanobetti, and Joel D. Schwartz. Associations between seasonal temperature and dementia-associated hospitalizations in New England. *Environment International*, 126:228–233, 2019. doi:10.1016/j.envint.2018.12.054.

[110] David C. Wheeler and Catherine A. Calder. An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. *Journal of Geographical Systems*, 9(2):145–166, 2007. doi:10.1007/s10109-006-0040-y.

[111] David C. Wheeler and Lance A. Waller. Comparing spatially varying coefficient models: a case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests. *Journal of Geographical Systems*, 11(1):1–22, 2009. doi:10.1007/s10109-008-0073-5.

[112] Ander Wilson, Ana G. Rappold, Lucas M. Neas, and Brian J. Reich. Modeling the effect of temperature on ozone-related mortality. *The Annals of Applied Statistics*, 8 (3):1728–1749, 2014. doi:10.1214/14-AOAS754.

[113] Ander Wilson, Yueh-Hsiu Mathilda Chiu, Hsiao-Hsien Leon Hsu, Robert O Wright, Rosalind J Wright, and Brent A Coull. Potential for Bias When Estimating Critical Windows for Air Pollution in Children's Health. *American Journal of Epidemiology*, 186(11):1281–1289, 2017. doi:10.1093/aje/kwx184.

[114] Ander Wilson, Hsiao-Hsien Leon Hsu, Yueh-Hsiu Mathilda Chiu, Robert O. Wright,

Rosalind J. Wright, and Brent A. Coull. 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, 2022. doi:10.1214/21-AOAS1533.

[115] Andrea Winquist, Ellen Kirrane, Mitch Klein, Matthew Strickland, Lyndsey A. Darrow, Stefanie Ebelt Sarnat, Katherine Gass, James Mulholland, Armistead Russell, and Paige Tolbert. Joint Effects of Ambient Air Pollutants on Pediatric Asthma Emergency Department Visits in Atlanta, 1998–2004. *Epidemiology*, 25(5):666–673, 2014. doi:10.1097/ede.0000000000000146.

[116] Spencer Woody, Carlos M. Carvalho, and Jared S. Murray. Model Interpretation Through Lower-Dimensional Posterior Summarization. *Journal of Computational and Graphical Statistics*, 30(1):144–161, 2021. doi:10.1080/10618600.2020.1796684.

[117] Zhiwei Xu, James Lewis Crooks, Deborah Black, Wenbiao Hu, and Shilu Tong. Heatwave and infants' hospital admissions under different heatwave definitions. *Environmental Pollution*, 229:525–530, 2017. doi:10.1016/j.envpol.2017.06.030.

[118] Zhiwei Xu, Shilu Tong, Jian Cheng, Yuzhou Zhang, Ning Wang, Yuqi Zhang, Alimila Hayixibayi, and Wenbiao Hu. Heatwaves, hospitalizations for Alzheimer's disease, and postdischarge deaths: A population-based cohort study. *Environmental Research*, 178: 108714, 2019. doi:10.1016/j.envres.2019.108714.

[119] Antonella Zanobetti, Marie S. O'Neill, Carina J. Gronlund, and Joel D. Schwartz. Susceptibility to Mortality in Weather Extremes: Effect Modification

by Personal and Small-Area Characteristics. *Epidemiology*, 24(6):809–819, 2013. doi:10.1097/01.ede.0000434432.06765.91.

[120] Yuzi Zhang, Stefanie T. Ebelt, Liuhua Shi, Noah C. Scovronick, Rohan R. D'Souza, Kyle Steenland, and Howard H. Chang. Short-term associations between warm-season ambient temperature and emergency department visits for Alzheimer's disease and related dementia in five US states. *Environmental Research*, 220:115176, 2023. doi:10.1016/j.envres.2022.115176.

[121] Mingyuan Zhou, Lingbo Li, David Dunson, and Lawrence Carin. Lognormal and Gamma Mixed Negative Binomial Regression. *Proceedings of the ... International Conference on Machine Learning. International Conference on Machine Learning*, 2012: 1343–1350, 2012.

[122] Ozan İrsoy, Olcay Taner Yıldız, and Ethem Alpaydın. Soft decision trees. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1819–1822, 2012.