Bayesian Causal Inference for Estimating Impacts of Air Pollution Exposure

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029762">http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029762</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Bayesian causal inference for estimating impacts of air pollution exposure

A dissertation presented
by
Shirley Liao
to
the Department of Biostatistics
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts
May 2019
Bayesian causal inference for estimating impacts of air pollution exposure

Abstract

Estimation of the causal effect of air pollution exposure on population health measures poses unique challenges. One commonly used method for estimating causal effects on such data is propensity score analysis (PSA), which controls for confounding in a “design” stage where propensity scores (PS) are estimated and implemented. Our first paper addresses uncertainty in the design stage of PSA and formulates a probability distribution for the design-stage output in order to lend a degree of formality to Bayesian methods for PSA (BPSA) that have gained attention in recent literature. A procedure for obtaining the posterior distribution of causal effects after marginalizing over a distribution of design-stage outputs is then deployed in an investigation of the association between levels of fine particulate air pollution and elevated exposure to emissions from coal-fired power plants. In order to address seasonality in air pollution emissions, as well as time-varying confounding which occurs from weather and climate variables, our second paper extends two procedures for estimating the average treatment effect on the overlap population (ATO), which may be estimated with less bias and less variability over replications than the average treatment effect over the general population (ATE) via inverse probability weighting (IPW) or stabilize weighting (SW) when low covariate overlap exists in the data. An analysis using these methods is performed on Medicare beneficiaries residing across 18,480 zip codes in the U.S. to evaluate the effect of coal-fired power plant emissions exposure on ischemic heart disease hospitalization, accounting for seasonal patterns that lead to change in treatment over time. Our third paper addresses non-linear confounding
Dissertation Advisor: Corwin Zigler and Francesca Dominici

and higher-order interactions which may exist in the relationship between ozone exposure and violent criminal activity by performing an analysis using Bayesian additive regression trees (BART), a powerful machine learning procedure able to model complex, non-linear relationships. This study employs time-series data from 6 cities in the US (Chicago, NYC, Atlanta, Philadelphia, Phoenix, LA) from 2009 to 2018 in order to estimate the causal effect of ozone exposure above NAAQS standards for air quality, as well as of a continuous causal effect of ozone exposure on violent crime rates.
Contents

1 Uncertainty in the Design Stage of Two-Stage Bayesian Propensity Score Analysis
   1.1 Introduction ................................................. 1
   1.2 Notation, Estimands, and Overview of Marginalizing over Design Uncertainty ................................. 4
   1.3 A Distribution of Designs Governing Design Uncertainty ......................................................... 7
   1.4 Sampling from the marginal posterior distribution of treatment effects ......................................... 10
   1.5 Simulation Study to Assess Design and Analysis Uncertainty under varying PS models and implementations ............................................... 16
   1.6 Investigating the Effect of High Power Plant Emissions on Ambient Pollution ................................. 21
   1.7 Discussion ...................................................... 24

2 Posterior Predictive Treatment Assignment Methods for Causal Inference in the Context of Time-Varying Treatments
   2.1 Introduction ...................................................... 36
   2.2 MSM and covariate overlap with time-varying treatments ......................................................... 39
   2.3 Estimating the ATO in the time-varying treatment setting ......................................................... 44
   2.4 Simulation ...................................................... 48
   2.5 Causal effect of treatment to power plant emission on ischemic heart disease hospitalization rate at the zip-code level ............................................... 53
   2.6 Discussion ...................................................... 62
List of Figures

1.5.1 Relationship between between-design variability and PROPDU under different PS models and PS implementations .................. 27
1.5.2 Relationship between within-design variability and PROPDU under different PS models and PS implementations .................. 28
1.5.3 Relationship between bias and PROPDU under different PS models and PS implementations ........................................... 29
1.6.1 Histogram of estimated propensity scores under treated and control observations ......................................................... 32
1.6.2 Estimated ATEs and constructed 95% confidence/credible intervals created via PSA and BPSA ............................................. 33
2.5.1 Dichotomized treatment maps for winter (top left), spring (top right), summer (bottom left) and winter (bottom right) 2012 .......... 56
2.5.2 COP as identified by PPTA ...................................................... 56
2.5.3 Distribution of overlap weights .............................................. 58
2.5.4 Distribution of IP weights ..................................................... 58
2.5.5 Distribution of stabilized weights ......................................... 60
3.4.1 Histograms of daily counts of AA ......................................... 75
List of Tables

<table>
<thead>
<tr>
<th>Table Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5.1</td>
<td>Simulation result table for BPSA</td>
<td>30</td>
</tr>
<tr>
<td>1.5.2</td>
<td>Simulation result table for PSA</td>
<td>31</td>
</tr>
<tr>
<td>1.6.1</td>
<td>Application data: table of unadjusted averages of covariates across treatment</td>
<td>34</td>
</tr>
<tr>
<td>1.6.2</td>
<td>Estimated treatment effects of high power plant emissions on ambient pollution</td>
<td>34</td>
</tr>
<tr>
<td>1.6.3</td>
<td>Between- and within-design variance under BPSA</td>
<td>35</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Simulation results for homogeneous treatment effect data generation</td>
<td>51</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Simulation results for heterogeneous treatment effect data generation</td>
<td>51</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Size of pruned sample and percent of samples used for PPTA</td>
<td>52</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Table of baseline covariates</td>
<td>55</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Treatment pattern prevalence after dichotomization</td>
<td>57</td>
</tr>
<tr>
<td>2.5.3</td>
<td>Estimates of the causal effect of an additional season of high exposure to air pollution on IHD admissions (per 10,000 person-years) at the zip-code level</td>
<td>59</td>
</tr>
<tr>
<td>2.5.4</td>
<td>Distribution of weights for IPW, SW and OW and posterior probability of inclusion for PPTA calculated on application data</td>
<td>59</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Summary statistics for exposure, outcome and a subset of confounder variables provided for all six cities</td>
<td>68</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Effect estimates and 95% CI for dichotomized and continuous ozone exposure analyses</td>
<td>73</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Measures of fit for BART and OLS: Adjusted/Pseudo $R^2$ and RMSE</td>
<td>73</td>
</tr>
</tbody>
</table>
3.4.1 Effect estimates for continuous and dichotomous ozone exposure on non-violent crimes, larceny and auto theft, by city with bolded significant effects. 

3.4.2 Rate ratio estimates from Poisson regressions
This dissertation is dedicated to all my patient, loving friends and family who have tolerated me through endless tribulations and celebrated with me through comparatively briefer triumphs.
Sincerest thanks to Cory Zigler and Francesca Dominici for their mentorship and academic guidance, Sebastien Haneuse, Aaron Fisher and Rachel Nethery for helpful feedback and discussion, Clas Linnman, Danielle Poole and Jake Crookall for interesting and productive collaborations, and finally Kevin Cummisky, Lucas Henneman and Benjamin Sabath for aiding in the gathering, cleaning and interpretation of air pollution data.
1

Uncertainty in the Design Stage of Two-Stage Bayesian Propensity Score Analysis

1.1 Introduction

Propensity score (PS) analysis refers to a wide range of strategies for estimating causal treatment effects with observational data. Rubin and others [44, 92] motivate this process by conceptualizing an observational study as having arisen from a randomized clinical trial where the rules of assignment have been lost and must be estimated. This perspective draws focus to the “design” stage of a propensity score analysis (PSA), in which PS are estimated then implemented to create a sub- or pseudo-population of the data representative of a hypothesized randomized trial [41, 92]. Following the design stage, observed outcome information is used to estimate a treatment effect in the “analysis” stage.
A well-conducted design stage is “absolutely essential for drawing objective inferences for causal effects,” and is typically conducted without access to outcome data [92]. Many analytic decisions are required to create a successful design which implements the estimated propensity score (e.g., via matching, weighting, or subclassification) to achieve treated and untreated groups that are “balanced” with respect to observed covariates [14, 16, 23, 57, 97]. In traditional PSAs, once the researcher has satisfactorily approximated the design of the hypothesized randomized study, the sub- or pseudo- population of observations created in the design stage is treated as fixed or known and estimation of causal effects in the analysis stage is conducted conditional on the design. In light of the importance of the design stage and the multi-faceted decisions made towards estimating and implementing the propensity score, it stands to reason that uncertainty associated with the design stage of PSA should be propagated into the estimation of causal effects in the analysis stage in order to more fully acknowledge all potential sources of PSA uncertainty.

Consideration of design uncertainty is often framed solely in terms of whether and how to incorporate uncertainty from the estimation of the PS into variance estimation of causal effects [20, 22, 38, 47, 48, 58, 106]. However, more sources of uncertainty may exist in the design stage. Dehejia and Wahba [30] illustrated that when implementing the PS with nearest neighbor matching without replacement, the ordering of observations in the matching procedure resulted in different estimates of treatment effect, even when using the same set of estimated PS [30]. Abadie and Imbens derive a variance estimator for causal effects that accounts for both PS estimation uncertainty and uncertainty in the construction of matches [12]. Further expansion of the idea of design uncertainty has appeared in Zigler and Dominici, who discuss uncertainty in choice of covariates included in the PS model [109], Spertus and Normand [96] who construe probabilistic design weights, and Zigler and Cefalu, who regard the subset of observations pruned (or truncated) from the design to be unknown [23].

Rather than conduct estimation of causal effects conditional on a single design stage, we offer an analysis that first generates a probability distribution for the space of all possible designs, the structure of which is determined by a) the PS estimation model and b) the specific type of implementation used in the design. To propagate design uncertainty into effect estimation, inference in the analysis stage is then marginalized over the distribution of possible designs while maintaining separation between the outcome data and the design stage. Developing a distribution for the space of designs clarifies the dis-
tinction between quantities we define as 1) “design estimation uncertainty” (DEU) corresponding to familiar notions of estimation uncertainty in the PS, and 2) “design decision uncertainty,” (DDU) corresponding additional stochasticity inherent to some PS implementations which lead to various design outputs for a fixed value of the PS. Uncertainty arising from treatment effect estimation in the analysis stage is defined as analysis estimation uncertainty (AEU). These elements of uncertainty are considered in a setting where the following PSA decisions are made a priori: the causal estimand of interest, the form of the PS model, the type of PS implementation (e.g., matching, weighting, etc.), and the method of treatment effect estimation in the analysis stage.

The method described herein is operationalized with the mechanics of Bayesian inference. While performing Bayesian inference separately in either the “design” or “analysis” stage is straightforward, combining them in a single Bayesian updating approach requires careful consideration. Bayesian inference that maintains the inherent two-stage nature of PSA gained attention beginning with McCandless [70, 71], and continued to appear in following literature [15, 53, 110–112]. These procedures fall within a broader class of methods for modularized Bayesian inference [50, 65], where a “model made of modules” may restrict the propagation of information from one module to another for reasons such as preventing the misspecification of one module from contaminating another, although there are other motivating considerations [50]. In the PS context, the design and analysis stages can be cast as different modules, with care taken so that the analysis does not contaminate the design. Such restricted flow of Bayesian updating has also been described as a “cut” or a “valve” to prevent feedback of information [68, 77]. One goal of this paper is to bring together many such issues that have appeared sporadically in literature towards improved understanding of the role of Bayesian methods for accounting for different elements of uncertainty in a propensity score analysis. As will be elaborated, deploying the mechanics of Bayesian inference across distinct “design” and “analysis” stages (or modules) has close ties to multiple imputation (MI) of missing data. As in MI, Bayesian ideas motivate considerations for marginalizing over uncertainty in the unknown design, but notions of formal Bayesian validity are elusive in this instance, and thus the procedures contained herein are evaluated on the basis of their “calibration” or Frequentist operating characteristics [63, 74].

We examine this framework within the context of several commonly-used PS implementations and corresponding analysis stages: quantile stratification, nearest-neighbor
matching with replacement, caliper matching with replacement, inverse probability of treatment weighting (IPTW), and a common doubly-robust estimator, although the formalization applies to a broader range of PSAs. We consider both settings where the analysis stage consists of a parametric “outcome model” (conditional on the PS), and also settings where no outcome-model likelihood exists, as would be the case with many common non- or semi-parametric (e.g., matching or weighted) estimators with known asymptotic properties. The latter case is particularly complicated for Bayesian inference, in which case we propose an approximation to the asymptotic posterior distribution of a causal effect. The collection of PS analyses considered here is not designed to be exhaustive and we make no argument for superiority of any given approach. Rather, these specific implementations and analysis stages are used to illustrate the different mechanistic sources of uncertainty that can arise when performing a PSA.

After distinguishing between different sources of design uncertainty and offering a formalization of integrating design uncertainty into the analysis stage, we outline a corresponding computational algorithm and illustrate the approach in a simulation study which demonstrates how the quantity of design uncertainty varies under choices a statistician makes in the design stage - from the variables included in the PS model to the form of PS implementation utilized. Arguments borrowed from MI literature will be used to compare between- and within-design variability in treatment effect estimates in order to quantify how design uncertainty impacts posterior inference on treatment effects. We then illustrate this method in an analysis of emissions from coal-fired power plants and ambient particulate air pollution collected over 22,723 zip codes in the Northeast, Southeast and Industrial Midwest regions of the United States. The paper concludes with a discussion of possible future directions.

1.2 Notation, Estimands, and Overview of Marginalizing over Design Uncertainty

1.2.1 Notation and Estimand

Let \( Y_i, X_i \) and \( T_i \) represent the observed outcome, covariate vector (of length \( p \)) and treatment indicator for unit \( i \), where \( i = 1, 2, \ldots, n \). \( T_i \in [0, 1] \) is dichotomous and \( Y_i \) is a continuous random variable. Let \( \mathbf{Y} \) and \( \mathbf{T} \) be vectorized representations of the
data and \( \mathbf{X} \) a matrix of observed covariates. We state without comment the following assumptions common to causal inference literature and defer readers to other work for details\[92\]: positivity (each subject has a non-zero probability of receiving either treatment) and SUTVA (units do not interfere with each other’s outcomes and potential outcomes are well-defined).

The additional key assumption of strongly ignorable treatment assignment states that potential outcomes which exist under either treatment are independent of observed treatment assignment after adjusting for observed covariates in \( \mathbf{X} \), which must include all confounders. Under ignorability, the average treatment effect (ATE) may be estimated from observed outcomes among units that exhibit the same distribution in background covariates:

\[
\Delta_{\text{ATE}} = E[E(Y_i | T_i = 1, X_i) - E(Y_i | T_i = 0, X_i)]
\] (1.1)

All relevant covariate information may be summarized in a “coarsened” manner with the propensity score, defined as the probability to receive the treatment rather than the control conditional on pre-treatment covariates of that individual \[88\]. Let \( e_i \) represent the propensity score for individual \( i \).

\[
e_i = P(T_i = 1 | X_i)
\] (1.2)

Units which are homogeneous with respect to \( e_i \) are said to be “balanced” with respect to \( X_i \), and under the assumption of strong ignorability, units with the same value of propensity score but assigned to different treatments have an expected difference in responses equal to the average treatment effect \[88\].

While \( \Delta_{\text{ATE}} \) is defined above as the marginal average treatment effect in the population, \( \Delta \) (without a subscript) will be used from this point to generically represent a causal estimand such as the ATE or the ATT(C) (Average Treatment Effect on the Treated (Control)). The choice of estimand considered is implied by the PS implementation, which is assumed to be chosen \textit{a priori}. Differences between estimands are not the focus of this paper, rather, we will consider estimation of each in the context of a corresponding implementation stage (e.g., matching will be implicitly taken to be estimating the ATT, whereas IPW estimates the ATE).
1.2.2 Overview of Marginalizing over Design Uncertainty

We begin with a heuristic description of the proposed framework for Bayesian PSA, deferring details to subsequent sections. Without the use of the propensity score, traditional Bayesian inference for \( \Delta \) would follow from specification of a likelihood for \((T, X, Y)\) conditional on unknown parameters, \(\theta\), a prior distribution for \(\theta\), and some function relating the data and \(\theta\) to the quantity \(\Delta\). However, this standard procedure for Bayesian inference to jointly estimate the propensity score and \(\Delta\) does not provide reliable causal estimates, as the PS is not intended to be part of the data-generating likelihood for \(Y\) \([85, 86, 111]\), thus motivating careful delineation of unknown quantities in the design and analysis stages.

Let \(\nu\) be a parameter summarizing the output of a PS implementation, that is, representing one propensity score “design” implied by implementing a set of estimated propensity scores. For example, \(\nu\) may be defined as a partition of observations into strata based on their estimated propensity scores (detailed definition of \(\nu\) is developed in the context of several PS implementations in Section 1.3.2). Under the assumptions outlined in Section 1.2.1 and a design stage that is successful in the sense that it appropriately balances observed covariates between treated and untreated units, inference for \(\Delta\) may be carried out with standard methodology (Bayesian or otherwise) conditional on a single value of \(\nu\), possibly with variance estimators adapted to account for uncertainty in \(\nu\) \([12, 93, 106]\). Instead of conditioning on a single realization of \(\nu\), the goal of this paper is to marginalize posterior inference for \(\Delta\) over the space of reasonable values of \(\nu\), each with an associated probability. This is represented heuristically as:

\[
f(\Delta|T, X, Y) = \int_{\nu} f(\Delta|T, X, Y, \nu)f(\nu|T, X)d\nu
\] 

The distribution \(f(\nu|T, X)\) is written to explicitly acknowledge that the design stage is conducted without outcome information \((Y)\). Integrating the posterior distribution of \(\Delta\) over the marginal distribution of \(\nu\) is a mathematical representation of propagating design uncertainty into the analysis stage by marginalizing over the space of all possible designs.

Expression (1.3) also highlights an important analog to the analysis of missing data: Construing \(\nu\) as a missing quantity closely parallels the Bayesian derivation of a multiple imputation procedure, where \((T, X, Y)\) represent the observed data and the expression \(f(\nu|T, X)\) represents a model for the missing data, intentionally specified in this case to
omit Y to preclude the outcome from contributing to the design. Integration over v represents the multiple imputation of “missing” designs, with the analysis stage conducted conditional on each and, as will be detailed in the subsequent, uncertainty in the parameters governing the construction of v (in addition to the value of v itself) being propagated into estimates of Δ. The modularity of BPSA means that, analogous to procedures for multiple imputation, f(Δ|T, X, Y) may not be regarded as a standard Bayesian posterior [63, 74]. Detailed specification of the modules will follow in Sections 1.3 and 1.4, with the corresponding computational procedure described in Section 1.4.2.

1.3 A DISTRIBUTION OF DESIGNS GOVERNING DESIGN UNCERTAINTY

1.3.1 BAYESIAN PROPENSITY SCORE ESTIMATION AND DESIGN ESTIMATION UNCERTAINTY

The parameter v represents the result of propensity score estimation and implementation for which ignorable treatment assignment is assumed, thus its probability distribution f(v|X, T) must be anchored to the PS estimation model. We expand the posterior distribution of v to reflect its dependence on parameters from the PS model, which represents the “design module” of expression (1.3):

\[
f(v|X, T) = \int f(v|a, X, T)f(a|X, T)da = \int f(v|a, X, T)L(T|X, a)\pi(a)da,
\]

where \(L(T|X, a)\) represents a likelihood function for the treatment assignment mechanism (i.e., a propensity score model) and \(\pi(a)\) represents a prior distribution for the unknown parameter a.

For illustration throughout this paper, we consider a simplistic propensity score model based on a generalized linear model \(L(T|X, a) = \prod_{i=1}^{n} g(aX_i)^{T_i}(1-g(aX_i))^{1-T_i}\), with a logit link function \(g(aX_i)\), although the ensuing discussion will be relevant to any parametric propensity score model specification with unknown parameter a. Note that the propensity score for each observation, \(e_i\), is a deterministic function of \((a, X, T)\). In the case of a logistic propensity score model, \(e_i = g(aX_i) = \frac{e^{aX_i}}{1+e^{aX_i}}\). A traditional Frequentist estimate of the propensity score may be obtained by replacing the unknown parameter a with its
maximum likelihood estimate. Instead, integrating with respect to $a$ in expression (1.4) may be thought of as simulating from the posterior “predictive” distribution of propensity scores for each individual [88]. From this point forward, we refer to $a$ and the PS interchangeably with regards to describing a posterior distribution of quantities derived from the treatment assignment model.

Variability in the predictive distribution of propensity scores captures uncertainty in the estimation of the propensity scores, which we refer to as “design estimation uncertainty” (DEU). Propagation of DEU into causal effect estimates has been previously considered in Bayesian and non-Bayesian contexts [14, 53, 69].

1.3.2 Propensity Score Implementation and Design Decision Uncertainty

Define $\nu$ as a vector of length $n$, where $\nu = [\nu_1, \nu_2, \ldots, \nu_n]$ and $\nu_i$ encodes the output of the PS implementation for individual $i$. The parameter space for each $\nu_i$ depends upon the choice of PS implementation procedure and the space of the entire vector $\nu$ may receive additional restrictions based on the details of the implementation. Delineation of the components of $f(\nu|T, X)$ as expressed in expression (1.4) draws a distinction between uncertainty derived from estimation of the propensity score (DEU) and any additional source of uncertainty owing to the stochasticity inherent to the implementation, to which we refer as “design decision uncertainty” (DDU). The possibility of DDU is denoted by the expression $f(\nu|a, X, T)$.

A PS implementation may or may not generate DDU. We refer to implementations that do not generate DDU as deterministic implementations, which produce only one possible value of $\nu$ for each value of $a$. In these implementations, $f(\nu|T, X, a)$ is a point-mass function and all design uncertainty derives from DEU. In contrast, probabilistic implementations can produce multiple values of $\nu$ for a single $a$, as dictated by the non-degenerate distribution $f(\nu|T, X, a)$. In the case of probabilistic implementations, both DDU and DEU contribute to design uncertainty.

We consider specifying $\nu$ and $f(\nu|T, X, a)$ in the context of five propensity score implementations: quintile stratification, nearest neighbor matching with replacement, caliper matching with replacement, and weighting for normalized inverse-probability-weighted (IPW) and doubly-robust (DR) estimators.
Stratification

Stratification is a type of sub-classification performed on quantiles of the estimated propensity score distribution. For this implementation, $\nu_i$ may be specified as a categorical variable, with $\nu_i = q$ if individual $i$ is assigned to strata $q$, when the sample is stratified into one of $q = 1, 2, \ldots, Q$ quantiles. We consider quintile stratification such that $Q = 5$, $\nu_i$ takes on a value in $[1, 2, 3, 4, 5]$ and $\nu$ is defined on the space of possible allocations of $n$ units into quintiles. Stratification is a deterministic PS implementation method.

Nearest Neighbor Matching With Replacement, implemented with a caliper

Nearest neighbor (NN) matching with replacement considers treated observations one at a time and matches each to the control observation(s) (ratio of matching decided a priori) with the closest propensity score. Controls are allowed to be matched to multiple treated observations, and thus may appear multiple times in the matched set. By implementing NN matching with a caliper, the pool of possible matches is limited by the caliper width, preventing unsuitable matches and instead pruning treated observations with propensity scores too removed from the distribution of control propensity scores. In this implementation (and all matching with-replacement implementations), $\nu_i$ may be defined as the frequency weight for observation $i$, which is calculated by standardizing the frequency of inclusion of each observation to sum to the number of unique observations within each treatment arm in the matched set. Thus $\nu_i$ is defined on the space of real numbers bounded above by the maximum (across treatments) number of observations within each treatment arm.

As the case with stratification, NN matching with replacement is a deterministic implementation yielding one matched set for each set of propensity scores and a point-mass function $f(\nu | T, X, a)$.

Caliper Matching With Replacement

In addition to NN matching, we also consider matching with replacement utilizing a caliper, which we refer to as “caliper matching.” While this implementation is not widely used in propensity score literature, we incorporate it specifically as a stochastic counterpart.
to the NN algorithm of Section 1.3.2. The difference between the two matching algorithms is that while NN chooses matches deterministically (closest based on PS distance) caliper matching chooses matches randomly among the candidates contained within the caliper.

Within the context of caliper matching with replacement, $v$ is defined in the same manner as NN matching with replacement. However, the random choice of control matches within the caliper renders this a probabilistic implementation; even for a fixed $\alpha$, there may still be variability in $v$ owing to the random selection of matches. $f(v|\mathbf{T}, \mathbf{X}, \alpha)$ is a probability distribution without an easily obtainable closed-form solution, but draws may be taken via iteratively performing the caliper matching algorithm multiple times conditional on the same $\alpha$.

**Weighting for IPW and DR estimators**

In contrast to matching and subclassification, weighting implementations utilize the propensity score to create a “pseudo-population” of observations in the treatment group, control group, or both in order to balance covariate distributions between the two groups. Inverse probability weighting (IPW) assigns weights based the following transformation of the propensity score:

$$w_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i} \tag{1.5}$$

For IPW, $v_i = w_i$, living on the space of positive real numbers. Since weights are created from a one-to-one transformation of a given a set of PS, the implementation is deterministic and $f(v|\mathbf{T}, \mathbf{X}, \alpha)$ is point-mass. This applies whether the weights are deployed in a standard IPW analysis or in tandem with an outcome model specification towards construction of a doubly-robust estimator.

**1.4 Sampling from the Marginal Posterior Distribution of Treatment Effects**

This section puts the distribution of $v$ formulated in Section 1.3 within the modularized Bayesian framework presented in Section 1.2.2 with the goal of obtaining posterior
inference on the marginalized posterior distribution of treatment effects. Incorporating the relationship between design output $v$ and the PS model, we expand expression (1.3) as:

$$f(\Delta|T, X, Y) = \int_v f(\Delta|T, X, Y, v)f(v|T, X)dv \quad (1.6)$$

$$\propto \int_v f(\Delta|T, X, Y, v) \int_a f(v|a, X, T)f(a|X, T)da \quad (1.7)$$

$$\propto \int_v f(\Delta|T, X, Y, v) \int_a f(v|a, X, T) \pi(a) \prod_{i=1}^n L(T_i|X_i, a) da \quad (1.8)$$

The posterior distribution of treatment effects conditional on the design is denoted with $f(\Delta|T, X, Y, v)$, representing the “analysis module” of BPSA, with variability in this distribution defined as “Analysis Estimation Uncertainty” (AEU). The functional form of $f(\Delta|T, X, Y, v)$ is specified a priori, depending in part on the structure of $v$ implied by the design stage. Some PS implementations permit $f(\Delta|T, X, Y, v)$ to be specified based on a parametric outcome model, while others dictate specification of $f(\Delta|T, X, Y, v)$ using weighted estimators with known asymptotic properties, which are typically not motivated on Bayesian grounds. Sections 1.4.1 and 1.4.1 outline development in each case.

### 1.4.1 Formalizing the Conditional Posterior Distribution of $\Delta$

#### Analysis stages where $\Delta$ is estimated with a parametric model

We first consider settings where the analysis stage consists of estimating a parametric outcome model, conditioned on $v$. Following propensity score stratification, for example, the analysis stage may be performed with a parametric outcome model having the likelihood function $L(Y|T, X, Y, v, \theta)$, where $\Delta$ is defined as a function of the data and unknown parameters $\theta$. Allowing definition of the conditional posterior distribution of $\Delta$ from expression (1.6) as:

$$f(\Delta|T, X, Y, v) = \int f(\Delta|T, X, Y, \theta)f(\theta|T, X, Y, v)d\theta \quad (1.9)$$

where $f(\theta|T, X, Y, v) = L(Y|T, X, Y, v, \theta)\pi(\theta)$ is the posterior distribution of the out-
come model parameters under prior distribution $\pi(\theta)$. For illustration, we consider $f(\cdot)$ to be a linear regression model with $\theta = (\beta, \sigma^2)$ representing, respectively, a vector of regression coefficients and a conditional variance parameter. In this paper, we only consider a parametric outcome model in the case of stratification, where $\nu$, a factor variable indicating stratum membership, is included as a covariate in the outcome model, along with an interaction with the treatment. In this case, the causal effect $\Delta$ is a function of $\theta$ and observed data.

**Analysis stages where $\Delta$ is estimated with an estimator with known asymptotic properties**

Some PS implementations, in particular those based on weighting, do not lend themselves to estimation of $\Delta$ with a parametric outcome model. The form of the estimator depends on the definition of the weights, which we illustrate with implementations of caliper matching with replacement, NN matching with replacement, IPW, and a simple DR estimator (see Appendix 1 for weights and weighted estimators of these implementations). The lack of a parametric model for $Y$ and corresponding likelihood expression requires a different approach for specifying the conditional distribution $f(\Delta|T, X, Y, \nu)$ from expression (1.6).

One option utilized in past literature \[53\] is specification of $f(\Delta|T, X, Y, \nu)$ with the known asymptotic distribution of the estimator for $\Delta$, conditional on one realization of $\nu$. Let $\hat{Q}(\nu)$ represent an estimator of $\Delta$ and $\hat{V}(\nu)$ the sampling variance, both computed conditional on design $\nu$. Then an asymptotic approximation takes the form:

$$ f(\Delta|T, X, Y, \nu) \sim N(\hat{Q}^{\nu}, \hat{V}(\nu)) $$

(1.10)

Appendix 1 lists the specific expressions used to approximate $f(\Delta|T, X, Y, \nu)$ for BPSA performed with caliper matching, NN matching, IPW and DR estimation.

When using asymptotic distributions to approximate $f(\Delta|T, X, Y, \nu)$, evaluation of the marginal posterior distribution in expression (1.6) can be construed as evaluation of the “asymptotic posterior distribution” of $\Delta$, marginalized over design uncertainty. Note that, since the derivation of $\hat{Q}(\nu)$ and $\hat{V}(\nu)$ is not typically motivated on Bayesian grounds, they do not easily incorporate prior information on $\Delta$. Asymptotic estimators with poor finite-
sample performance might propagate similar properties into posterior estimates of $\Delta$.

1.4.2 Outline of Computational Procedure for Marginalizing over Design Uncertainty

Here we outline a sequential Markov-chain Monte Carlo (MCMC) procedure for evaluating the posterior distribution of causal effects, marginalized over design uncertainty as defined in expression (10), building off of previous literature $[14, 23, 53, 96]$.

Drawing from the posterior predictive distribution of $\nu$

In the first stage of BPSA, multiple draws are taken from $f(\nu|T, X)$ using the following procedure:

1. Obtain a sample of $K$ draws from the posterior distribution of the parameters of the propensity score model, $\alpha$. This can be accomplished with standard MCMC routines, for example, with the R package MCMCpack.

2. For each of the $K$ draws from posterior distribution of $\alpha$, calculate propensity scores for each observation with the use of covariate information $X$. Steps 1 and 2 in combination may be conceived as taking $K$ samples from the posterior predictive distribution of propensity scores.

3. For deterministic PS implementations, use the $K$ draws from the posterior predictive distribution of propensity scores to perform $K$ propensity score implementations, each representing a design stage $\nu_k$, indexed by $k = 1, 2, ..., K$. For stochastic PS implementations, perform the implementation $R$ times conditional on each set of PS for all $K$ draws, resulting in $R \times K$ values of $\nu$.

The output of the first stage (the design module), which may be performed in its entirety independent of outcome information, produces either $K$ or $R \times K$ draws of $\nu$ from its marginal distribution, $f(\nu|T, X)$, each representing a different design. For the analysis stage, we simplify notation and use $\nu_k$ to describe a single simulated output from the design stage, omitting the possible dependence on $R$ (in the case of probabilistic designs). If the PS distribution is explored well enough ($K$ is large), $R$ does not need to be too large.
(and may even be set to 1) as the MCMC algorithm re-visits the same sets of PS multiple times and the variability of \( f(\nu|T, X, a) \) is explored in that way.

**Drawing from the conditional posterior distribution of \( \Delta \)**

Estimation in the analysis stage is conducted conditional on each of the simulated values of \( \nu_k \) from the design stage. Specifically, for each value of \( \nu_k, k = 1, 2, \ldots, K \):

4. Draw \( S \) samples from the conditional posterior distribution \( f(\Delta|T, X, Y, \nu_k) \)

   When the analysis stage entails a parametric distribution for \( f(\theta|T, X, Y, \nu_k) \)

   **label=4.0a.** Take \( S \) draws from the posterior distribution of \( \theta = [\beta, \sigma^2] \), a vector of parameters with the posterior distribution \( f(\theta|T, X, Y, \nu) \propto L(Y|T, X, \theta, \nu)\pi(\theta) \).
   Let \( \theta_{sk} \) represent one such draw, with \( s = 1 \ldots S \). Again, this may be accomplished with standard MCMC procedure and a packages such as MCMCpack.

   **label=4.0b.** Each draw of \( \theta_{sk} \) may be transformed into \( \Delta_{sk} \) (ex: for stratification, \( \Delta \) is a linear combination of coefficients of a regression fit where treatment assignment is interacted with strata membership).

   When an asymptotic approximation is used for the distribution of \( f(\Delta|T, X, Y, \nu_k) \)

   **label=4.0b.** Calculate \( \hat{Q}(\nu_k) \) and \( \hat{V}(\nu_k) \).

   **label=4.0b.** Draw \( S \) samples from \( N(E(\hat{Q}(\nu_k), \hat{V}(\nu_k)) \) of \( \Delta_{sk}, s = 1 \ldots S \).

With either step 4a or 4b of the analysis, the end result will be \( K \times S \) draws from the posterior distribution \( f(\Delta|T, X, Y) \), which is marginalized over design uncertainty. If interest lies primarily in estimating the posterior mean and variance of \( \Delta \), instead of retaining \( S \) draws from \( f(\Delta|T, X, Y, \nu_k) \) for each \( k \), it is only necessary to save \( \Delta_k = E(\Delta|T, X, Y, \nu_k) \) and \( \sigma_k^2 = Var(\Delta|T, X, Y, \nu_k) \), which can be used with Rubin’s combining rules [64], to estimate the posterior mean and variance of \( \Delta \).
1.4.3 Between and Within Design Uncertainty

The apparent parallel between the sequential modularized procedure in Section 1.4.2 and multiple imputation provides a useful analog that relates design uncertainty and analysis uncertainty to notions of “between” and “within” imputation variance. Construing \( v \) as a missing quantity, estimating the mean and variance of \( \Delta \) with Rubin’s combining rules as

\[
E(\Delta|T, X, Y) = E(E(\Delta|T, X, Y, v)) \approx \frac{\sum_{k=1}^{K} \Delta_k}{K} \text{ and:}
\]

\[
\text{Var}(\Delta|T, X, Y) = E(\text{Var}(\Delta|T, X, Y, v)) + \text{Var}(E(\Delta|T, X, Y, v)) \approx \tilde{\sigma}_K^2 + (1 + \frac{1}{K}) \frac{\sum_{k=1}^{K} \sigma_k^2}{K} + \frac{(1 + \frac{1}{K}) \sum_{k=1}^{K} (\Delta_k - \bar{\Delta}_K)^2}{K-1}
\]

(1.11)

(1.12)

Where \( K, \Delta_k \) and \( \sigma_k^2 \) are as defined in Section 1.4.2 and can be calculated whether a parametric distribution or asymptotic approximation is used in the analysis stage.

The quantity AEU relates to within-design variability (\( \tilde{\sigma}_K^2 \)), quantifying the variability in the distribution of \( \Delta \), conditional on \( v \). Design uncertainty (both DDU and DEU) relate to between-design variability (\( B^2_K \)) across the samples from the distribution of \( v \). Plummer [77] draws parallels between the sequential (or modularized) MCMC procedure of Section 1.4.2 and concepts in the MI literature, where the reliability of posterior inference relates to the interplay between these two elements of uncertainty. Equating \( K \) to the number of “imputed” designs, we define an expression for the proportion of total variability in \( \Delta \) attributable to the between-design uncertainty, \( PROP_{DU} \), which is closely related to the proportion of missing information in a multiple imputation procedure [91]:

\[
PROP_{DU} = \frac{B^2_K}{B^2_K + \tilde{\sigma}_K^2}
\]

(1.13)

In MI, when the percentage of missing information is high, inference on \( \Delta \) must be treated with caution. The same principle applies to design uncertainty and the relative influence of between-design variability as quantified in \( PROP_{DU} \). Section 2.4 examines the relationship between \( PROP_{DU} \) and BPSA performance in finite samples across several scenarios with varying amounts of design uncertainty. Examining BPSA through this lens also points towards expected behavior in large samples. As \( n \) approaches \( \infty \), the posterior
distribution of the propensity score model parameters ($f(\alpha|X, T)$) will converge to point mass at the MLE. For implementations with no DDU (i.e., point mass for $f(\nu|X, T, \alpha)$), this would result in between-design variability approaching 0 and the within-design variance concentrating around the asymptotic $\text{Var}(\Delta|X, Y, T, \nu)$, which, in combination, would result in BPSA estimators exhibiting the same asymptotic properties (e.g., coverage) as their Frequentist PSA counterparts.

1.5 Simulation Study to Assess Design and Analysis Uncertainty under Varying PS Models and Implementations

Here we provide a simulation study to evaluate how different “design choices” made in PSA impact design and analysis uncertainty. Specifically, we implement BPSA to examine design and analysis uncertainty under the different PS implementations from Section 1.3.2 and across PS models that include various types of covariates (i.e., confounders, instrumental, prognostic, and noise variables). To focus on the impact of design choices, the different PS models and implementations will be considered in Monte Carlo replicates from a single data generating mechanism.

While the primary goal is to examine the different components of uncertainty within BPSA and how quantities such as $\text{PROPDU}$ relate to performance, standard (i.e., Frequentist) PSA estimators will be included for comparison, without claims as to the superiority of the Bayesian vs. Frequentist approach.

1.5.1 Data Generating Mechanism

200 replicated data sets are simulated to have $n = 1000$ observations and 20 uncorrelated normally-distributed covariates with mean 0 and variance 1. 5 of these covariates are true confounders associated with both $T$ and $Y$, while 5 are instruments (associated with $T$ only), 5 are prognostic (associated with $Y$ only) and 5 are noise (associated with neither $T$ nor $Y$) variables. Treatment assignment $T$ is simulated from a Bernoulli distribution with probability of treatment specified by the logistic regression described in Section 1.3.1, with intercept set to zero and coefficients $\alpha$ set to 0.75 for all confounders and instrumental variables and 0 for all prognostic and noise variables. Outcomes, $Y$, are simulated from a linear regression model of the form $Y = 1 + \Delta T + \beta X + \epsilon$, with $\Delta$ set to 1.5, and $\beta$
set to $[0.1, 0.2, 0.3, 0.4, 0.5]$ for the five confounders, $0.5$ for all prognostic variables and $0$ for noise and instrumental variables. $\epsilon$ is a normally-distributed random error with mean $0$ and variance $1$.

1.5.2 BPSA and PSA procedures

We analyzed simulated datasets with BPSA (procedure described in Section 1.4.2 with $K = 1000$, $S = 200$ and $R = 1$ for caliper matching) and standard PSA (which conditions analysis on the MLE of the PS) in the context of four different PS model specifications and six PS implementations.

The four PS model specifications were logistic regressions including the following sets of variables:

1. Only confounders
2. Confounders and instrumental variables
3. Confounders, instrumental and prognostic variables
4. Confounders, instrumental, prognostic and noise variables

Note that all scenarios correctly specify the functional form of the PS model and all include at least the confounders necessary to satisfy the ignorability assumption. When calculating the DR estimator, the potential outcome model utilizes the same covariates as the PS model.

With each of the two PS model specifications, we employ the following PS implementations: 1-1 caliper matching with a caliper of $0.5$, 1-5 caliper matching with a caliper of $0.5$, 1-1 NN matching with a caliper of $0.5$, quintile stratification, and weighting with IPW and DR estimators. Different ratios used for caliper matching were designed to influence the amount of DDU in the implementation; increasing the number of control matches decreases the amount of DDU.

Under BPSA, the analysis stage following stratification is performed with parametric modeling of the conditional distribution of $\Delta$ (see Section 1.4.1) while the analysis stage following all other implementations involve asymptotic approximation of the conditional distribution of $\Delta$ (see Section 1.4.1).
Under PSA, the analysis stage after all implementations utilizes asymptotic estimators for both the treatment effect and variance, where “robust” variance estimators are employed after caliper matching (with both ratios), NN matching and IPW estimation, and a standard OLS regression is performed for stratification utilizing the same linear model as BPSA (which interacts treatment and strata membership). The variance estimator for DR under PSA is the conditional variance estimator for $\Delta$ which is utilized in BPSA. Details of all treatment effect and variance estimators may be found on Appendix 1.

1.5.3 Simulation Results

A complete table of simulation results, including bias, empirical variance, and mean squared error (MSE) of posterior mean estimates of $\Delta$ across replications, average between-design, within-design and total estimated variance across replications, average $PROP_{DU}$ across replications, and coverage of 95% posterior intervals for the 24 combinations of PS model and implementation can be found in Table 1.5.1. Plots depicting how between-design variance (log-transformed), within-design variance (log-transformed) and bias vary across levels of $PROP_{DU}$ appear in Figures 1.5.1, 1.5.2, and 1.5.3, respectively.

Compared to a PS model that includes only the confounders, adding additional instrumental and prognostic variables increases average between-design variability for all implementations (Figure 1.5.1), with less pronounced impact on within-design variability (Figure 1.5.2). As a consequence, adding these variables to the PS model leads to higher average values of $PROP_{DU}$ within all implementations.

Under all PS models, the implementations display clear patterns in the amount of $PROP_{DU}$ they exhibit. DR displays the lowest $PROP_{DU}$ under all models, followed by stratification and 1-5 caliper matching. This aligns with the expectation that these implementations are relatively insensitive to moderate perturbations in the PS. On the other hand, IPW and NN
matching display the highest values in $PROP_{DU}$, indicating that even mild perturbations in the PS may result in very different treatment effect estimates.

Estimates of $\Delta$ using IPW display the clearest trend in performance, where high $PROP_{DU}$ is associated with more erratic effect estimates in terms of large bias (Figure 1.5.3), poor coverage and high variance across replications (Table 1.5.1). For IPW, adding instrument and prognostic variables has the most impact on the between-design variance of IPW estimates (Figure 1.5.1), while within-design variance is small and largely unaffected by changes to the PS model specification (Figure 1.5.2), resulting in large differences in average $PROP_{DU}$ across PS models. Performance of $\Delta$ estimates is consistent with the expectation that adding many covariates into the PS model can result causal estimates that are sensitive to extreme weights and low covariate overlap[84].

The performance of NN matching, on the other hand, does not suffer as substantially as IPW under PS models which generate high $PROP_{DU}$. Under PS models with instrumental, prognostic and noise variables, NN matching estimates $\Delta$ with less empirical variability across replications than IPW (Table 1.5.1) under all PS models since frequency weights are less likely to take on extreme values compared to IP weights. Coverage levels for NN matching are closer to nominal than IPW under PS models 2-4 due to the comparatively higher average total variance. This is primarily due to NN matching’s between-design variability (Table 1.5.1) which increases as more covariates (regardless of type) are added to the PS model, created by averaging estimates over designs that consist of different subsets of the data (on average 71% under the confounders-only PS model and on average 63% under other PS models).

Both caliper matching implementations show similar trends to NN matching in terms of the relationship between $PROP_{DU}$ and $\Delta$’s estimation performance, but with lower average between-design variability, empirical variability, and within-design variability within each PS model. Comparing NN matching and 1-1 caliper matching, 1-1 caliper matching displays lower between-design variability because the pool of possible matches displays few changes across draws from the distribution of PS due to the generous caliper (Section 1.3.2), whereas the “nearest neighbor” of a given observation may change with every PS draw, resulting in greater differences between designs. 1-5 caliper matching displays less between-design variability than 1-1 caliper matching since the average of 5 matches naturally has less variability across designs than a single match, an example of decreasing DDU. Among all the matching procedures, 1-5 caliper matching shows the lowest within-design
variability (and tends to include the highest proportion of the observed data - 92% on average - in the design), and 1-1 caliper matching exhibits slightly lower within-design variability than NN matching. Both caliper matching procedures are similar to NN matching in their relative insensitivity to $PROP_{DU}$ in terms of within-design variability. Importantly, the additional DDU associated with the caliper matching does not translate into increased between-design variability in estimates of $\Delta$, relative to NN matching (which exhibits no DDU).

Stratification exhibits the lowest average between-design variability among virtually all implementations and across all PS models, as observations generally remain in the same PS strata across multiple draws from the posterior distribution of the PS (Table 1.5.1). Accordingly, stratification displays low $PROP_{DU}$ across all PS models, and lower empirical variability estimates of $\Delta$ than the matching or IPW procedures. The stratification approach exhibits slightly more bias than the other procedures, likely due to the relatively small number of strata.

The DR estimator is unique among our considered implementations as it is the only one which incorporates an explicit model for potential outcomes that includes direct covariate adjustment. Thus, it is not surprising that its performance deviates from the general trend of the other implementations. First, within each PS model most replications under DR exhibit much lower $PROP_{DU}$ compared to other implementations. While most implementations display only small differences in within-design variability across PS models, DR displays large increases in average within-design variability in response to adding instrumental, then prognostic variables to the PS model. For most of the PS models considered, DR displays the largest within-design variability on average. Between-design variability, however, appears to only moderately increase as instrumental and prognostic variables are added to the PS model, explaining DR’s relatively smaller $PROP_{DU}$ under all PS models. The inclusion of the outcome model in the DR estimate leads to the comparatively low between-design variability, as perturbations in the propensity score have less impact in effect estimation than the (correctly specified) outcome model.

With the exception of DR, empirical variance of posterior mean estimates of $\Delta$ across replications demonstrate a positive relationship with $PROP_{DU}$ (Table 1.5.1) within each PS model, indicating a correspondence between between-design variability and variability of point estimates in repeated samples. Designs (combinations of PS models and implementations) which exhibit high $PROP_{DU}$ also exhibit more variability in treatment effect
estimates across replication and thus require larger estimated variances in order to achieve nominal coverage.

Results of a standard PSA analysis are presented in Table 1.5.2. For the three matching implementations and stratification, BPSA results in lower empirical variability and MSE due to the averaging over multiple possible matched or stratified designs. This trend is more evident under PS models which include more than the necessary confounders. In contrast, BPSA versions of IPW and DR exhibit slightly higher empirical variability and MSE than PSA, with this difference also most pronounced in the PS models that include prognostic, instrument, and/or noise variables. This may be a consequence of extreme weights, where averaging over multiple designs includes those with weights extreme enough to provide erratic estimates and increased variability.

Standard PSA under the three matching implementations and IPW achieve nearer to nominal coverage than BPSA under most PS models. For DR, PSA performs similarly to BPSA under all PS models except for the PS model with confounders and instruments, where BPSA achieves nominal coverage while PSA intervals do not. Finally, with stratification both PSA performs similarly to BPSA with regard to coverage under all PS models.

1.6 INVESTIGATING THE EFFECT OF HIGH POWER PLANT EMISSIONS ON AMBIENT POLLUTION

An ongoing analysis in Henneman et. al. [28] is evaluating the association between long-term exposure to emissions from coal-fired power plants and Ischemic Heart Disease (IHD) hospitalizations among Medicare beneficiaries. One important feature of this analysis is the way in which the analysis incorporates ambient fine particulate matter (particles less than 2.5 micrometers in diameter, denoted PM). Ambient PM concentrations are expected to be derived in part from coal-fired power plant emissions in some regions, and PM is known to be associated with a variety of adverse health outcomes [79]. While the primary analysis in Henneman et. al. [28] investigates the link between coal emissions and IHD without adjusting for PM, a secondary analysis regards PM as an adjustment covariate and briefly evaluates whether PM could be ruled out as a possible mediator of the relationship between coal emissions and IHD hospitalizations. Here we revisit the secondary analysis of Henneman et. al. [28], deploying the methods described in Section 1.5.2 to in-
vestigate the extent to which a binary metric of high/low coal emissions exposure causally impacts annual average PM. Details in Henneman et. al. [28] describe the creation of a binary metric of coal power plant exposure derived using a reduced-complexity chemical transport model (InMap) [100], as well as the data-fusion derived estimates of PM [32].

We examine the effect of elevated coal emissions on ambient particle concentrations measured at 22,723 zip codes in the Northeast, Southwest and Industrial Midwest regions of the United States. The analysis adjusts for 17 possible confounders, including location of zip code (latitude and longitude), climate (temperature and humidity), population density, demographics (e.g. racial makeup, education, household income, gender) as well as the makeup of the residential areas in each zip code (e.g. urban/rural, real estate value). A full list of covariates considered can be found in Table 1.6.1. Further details on consolidating the zip-code-level covariates (retrieved from 2000 US Census data) may be found in Henneman et. al [28].

Among the 22,723 observations in this dataset, 7,211 are classified as “exposed” (to high coal emissions) and 15,512 are control observations. Figure 1.6.1 visualizes the overlap of estimated propensity scores among treated and control observations using maximum likelihood estimates of the PS, and Table 1.6.1 contains average values of included confounders compared across exposure status. On average, control zip codes are located further West, have lower population densities, are less urban and more rural, and exhibit a lower median household income and higher percent poverty.

Analysis was performed with the PSA and BPSA procedures described in Section 1.5.2, compared within the contexts of quintile stratification, caliper matching with replacement, NN matching with replacement (both with caliper = 0.5 and a 1:1 ratio), and weighting with IPW and DR estimators. In this analysis, we focus on estimating the ATE, but employ caliper and NN matching methods which measure the ATT for purposes of illustration.
Estimated treatment effects, 95% uncertainty intervals and analysis details may be found in Table 1.6.2, and are depicted graphically in Figure 1.6.2. Presented on the same scale for comparison are also estimates and intervals constructed via fully adjusted and unadjusted linear regressions. Between-, within- and total variance for all implementations utilized with BPSA may be found in Table 1.6.3. NN matching, 1-1 caliper matching and IPW all display high $PROP_{DU}$ at 0.59, 0.49 and 0.46, respectively, while DR and stratification display relatively lower $PROP_{DU}$ at 0.11 and 0.13, respectively (Table 1.6.3). This mirrors our simulation findings, as DR and stratification display less between-design variability compared to the matching and IPW methods.

Both the fully adjusted and unadjusted linear regressions estimate lower treatment effects than any of the propensity score analysis methods. IPW and stratification produced similar point estimates for the average treatment effect ($\approx 1.75$) while both matching implementations display a lower estimation of the ATT. The DR estimator displays the highest $ATE$ estimate, and its confidence intervals do not overlap with that of any other implementation. The discrepancy between the DR estimate and the IPW estimate may arise due to the relatively high $PROP_{DU}$, to which IPW estimator was shown (in the simulation) to be sensitive. The discrepancy with the stratification estimator may relate to the relatively small number of strata used.

While point estimates of treatment effects are nearly identical between PSA and BPSA procedures in the context of DR, IPW and stratification, BPSA point estimates are lower than PSA point estimates in the context of NN matching and caliper matching. This may be attributed to the large pool of control observations available compared to treated observations. Both matching algorithms on average use only 50% of the available observations in their treatment effect estimation, and selection of controls into the matched set is sensitive to design uncertainty arising from both DEU (the PS model includes many possible confounders, some of which may be instrumental variables) as well as DDU (in the case of caliper matching). By averaging over multiple matched sets, BPSA was shown to produce more stable estimates of the treatment effect in Section 2.4 and it is reflected in this analysis via the contrast between PSA and BPSA-estimated treatment effects.

As witnessed in the simulation study, BPSA produces intervals similar in width to PSA when performed with stratification and DR. Only under IPW does BPSA create narrower intervals than PSA, a result which occurred in the simulation study when non-confounders were included in the PS model. The most unexpected result occurs under 1-1 NN and
caliper matching, where BPSA produces slightly wider intervals than PSA, a result not observed in the simulation study. This is due to the difference in proportion of treated observations, which was on average equal to the proportion of control observations in the simulated data but much lower than the proportion of control observations in the power plant data. The high number of control observations creates variability in the treatment effect estimate across draws from the PS distribution, as reflected in the high between-design variance and $PROP_{DU}$ displayed by both matching methods (Table 1.6.3). The design uncertainty captured by the between-design variability translates into a higher total variance than PSA.

Overall, the results of the analysis provide the consistent message that elevated exposure to coal-fired power plant emissions causally increases the overall annual concentration of ambient fine particulate matter, underscoring the care with which the secondary analysis in Henneman et al. [28] should be interpreted.

1.7 Discussion

This paper formalizes a sequential Bayesian procedure for marginalizing causal effect estimates over uncertainty associated with the design stage of a PSA. The ability to distinctly define design and analysis uncertainty aids understanding of the sensitivity of causal effect estimate performance to decisions made during the approximation of the design of a randomized study with propensity scores. The procedure outlined in this paper synthesizes a variety of related ideas that have recently appeared under the topic of Bayesian methods for propensity score analysis [14, 50, 53, 69, 77, 110].

The simulation study explores how design decisions such as covariate selection for the PS model and choice of PS implementation dictate not only the quantity of design uncertainty which results, but how it is propagated into estimates of causal effects. In particular, the simulation showed how a summary measure such as $PROP_{DU}$ quantifying the relative amount of “between” and “within” design uncertainty can be useful for understanding the impact of design uncertainty on posterior variability in estimates of $\Delta$. In general, higher $PROP_{DU}$ was associated with higher variability of $\Delta$ point estimates across replications, though specific impacts of this with regard to coverage and MSE varied between implementations. The primary goal of the simulation section was not to establish an superiority to BPSA relative to standard PSA. While BPSA resulted in lower MSE for the matching
implementations and stratification and comparable coverage to PSA under DR and stratification, neither method appeared uniformly superior. In order to focus on the nuances around between- and within-design uncertainty, the simulation study considered the implications of design choices within a single data generation, further limiting our ability to make general claims of the relative performance of BPSA vs. standard PSA approaches.

The work here shares conceptual similarities to that of Branson [19], who compares the analysis stage of a PSA to that of a RCT with randomization restricted to pre-determined bounds on covariate balance. However, while Branson describes a conditional distribution of treatment effects which accounts for certain limitations placed on PS implementations, the paper does not address uncertainty arising from models or implementations utilized in the design stage nor how to propagate it into estimation of the treatment effect. Furthermore, Branson primarily focuses on PS matching while this paper extends to any PS implementation. Our discussion also construes design uncertainty broadly and does not detail all the possible factors of design uncertainty. However, this subject matter may be related to topics such as low covariate overlap across treatments, low treatment prevalence and highly correlated confounders. For example, IPW is known to give erratic finite-sample performance in the presence of low covariate overlap, which we have replicated in our simulations [57]. Work exploring these connections in further detail would be a promising extension. Other extensions could incorporate non-parametric, machine-learning [56] or flexible [109] PS estimation methods which have become recently popular.

We framed BPSA as a special instance of modularization in Bayesian inference, where the mechanics of posterior updating were seperated in the “design” module and the “analysis” module. The overlap between posterior inference on modules and multiple imputation is discussed in Plummer et. al. [77], and this paper makes ample use of this connection to frame marginalizing over design uncertainty through the lens of between- and with-design uncertainty, as would be relevant to an analysis of missing data. Despite the usefulness of the parallel between BPSA and multiple imputation, better understanding of the limits of this commonality could further improve understanding. For instance, ν in the present framework is an entirely synthetic parameter more akin to a latent variable than a missing quantity with a probabilistic missing-data response mechanism, which is key to establishing both Bayesian and Frequentist validity for estimators based on multiple imputation. Further exploration of the role that the theory of multiple imputation can
play in this context is an interesting direction for future work.

Though propensity score analysis is widely used in various fields including epidemiology and econometrics, notions of design uncertainty are rarely considered explicitly. Bayesian propensity score analysis provides a versatile computational procedure which marginalizes over all components of design uncertainty in a clearly defined manner. Though the performance of BPSA varies across implementations, it has been regarded as an attractive alternative to conditioning inference on a single estimate of the propensity score [14]. Explication and exploration of the ideas presented here can hopefully ground future work investigating how Bayesian methods can add to the existing literature on propensity score analysis.
Figure 1.5.1: Relationship between between-design variability and \( \text{PROP}_{DU} \) under different PS models and PS implementations.
Figure 1.5.2: Relationship between within-design variability and $PROP_{DU}$ under different PS models and PS implementations
Figure 1.5.3: Relationship between bias and \textit{PROP}_{DU} under different PS models and PS implementations
Table 1.5.1: Simulation result table for BPSA

<table>
<thead>
<tr>
<th>Implementation</th>
<th>PS model</th>
<th>Empirical var</th>
<th>Avg total var</th>
<th>Avg between var</th>
<th>Avg within var</th>
<th>Bias</th>
<th>PRG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caliper matching (1-1)</td>
<td>confound</td>
<td>0.015</td>
<td>0.021</td>
<td>0.005</td>
<td>0.016</td>
<td>0.070</td>
<td>0</td>
</tr>
<tr>
<td>Caliper matching (1-1)</td>
<td>instru</td>
<td>0.035</td>
<td>0.027</td>
<td>0.008</td>
<td>0.020</td>
<td>0.091</td>
<td>0</td>
</tr>
<tr>
<td>Caliper matching (1-1)</td>
<td>instru_prog</td>
<td>0.023</td>
<td>0.035</td>
<td>0.015</td>
<td>0.019</td>
<td>0.092</td>
<td>0</td>
</tr>
<tr>
<td>Caliper matching (1-1)</td>
<td>instru_prog_noise</td>
<td>0.023</td>
<td>0.034</td>
<td>0.015</td>
<td>0.019</td>
<td>0.103</td>
<td>0</td>
</tr>
<tr>
<td>Caliper matching (1-5)</td>
<td>confound</td>
<td>0.015</td>
<td>0.017</td>
<td>0.001</td>
<td>0.011</td>
<td>0.070</td>
<td>0</td>
</tr>
<tr>
<td>Caliper matching (1-5)</td>
<td>instru</td>
<td>0.035</td>
<td>0.023</td>
<td>0.004</td>
<td>0.013</td>
<td>0.091</td>
<td>0</td>
</tr>
<tr>
<td>Caliper matching (1-5)</td>
<td>instru_prog</td>
<td>0.024</td>
<td>0.031</td>
<td>0.011</td>
<td>0.013</td>
<td>0.092</td>
<td>0</td>
</tr>
<tr>
<td>Caliper matching (1-5)</td>
<td>instru_prog_noise</td>
<td>0.023</td>
<td>0.030</td>
<td>0.010</td>
<td>0.013</td>
<td>0.103</td>
<td>0</td>
</tr>
<tr>
<td>DR</td>
<td>confound</td>
<td>0.014</td>
<td>0.020</td>
<td>0.000</td>
<td>0.019</td>
<td>0.001</td>
<td>0</td>
</tr>
<tr>
<td>DR</td>
<td>instru</td>
<td>0.066</td>
<td>0.091</td>
<td>0.009</td>
<td>0.076</td>
<td>0.010</td>
<td>0</td>
</tr>
<tr>
<td>DR</td>
<td>instru_prog</td>
<td>0.054</td>
<td>0.134</td>
<td>0.017</td>
<td>0.108</td>
<td>-0.012</td>
<td>0</td>
</tr>
<tr>
<td>DR</td>
<td>instru_prog_noise</td>
<td>0.053</td>
<td>0.143</td>
<td>0.016</td>
<td>0.115</td>
<td>-0.003</td>
<td>0</td>
</tr>
<tr>
<td>IPW</td>
<td>confound</td>
<td>0.016</td>
<td>0.015</td>
<td>0.004</td>
<td>0.011</td>
<td>0.092</td>
<td>0</td>
</tr>
<tr>
<td>IPW</td>
<td>instru</td>
<td>0.068</td>
<td>0.030</td>
<td>0.012</td>
<td>0.011</td>
<td>-0.001</td>
<td>0</td>
</tr>
<tr>
<td>IPW</td>
<td>instru_prog</td>
<td>0.079</td>
<td>0.036</td>
<td>0.026</td>
<td>0.011</td>
<td>-0.036</td>
<td>0</td>
</tr>
<tr>
<td>IPW</td>
<td>instru_prog_noise</td>
<td>0.084</td>
<td>0.039</td>
<td>0.028</td>
<td>0.011</td>
<td>-0.045</td>
<td>0</td>
</tr>
<tr>
<td>NN matching</td>
<td>confound</td>
<td>0.019</td>
<td>0.027</td>
<td>0.009</td>
<td>0.017</td>
<td>0.017</td>
<td>0</td>
</tr>
<tr>
<td>NN matching</td>
<td>instru</td>
<td>0.062</td>
<td>0.047</td>
<td>0.024</td>
<td>0.023</td>
<td>0.021</td>
<td>0</td>
</tr>
<tr>
<td>NN matching</td>
<td>instru_prog</td>
<td>0.042</td>
<td>0.058</td>
<td>0.036</td>
<td>0.023</td>
<td>0.026</td>
<td>0</td>
</tr>
<tr>
<td>NN matching</td>
<td>instru_prog_noise</td>
<td>0.041</td>
<td>0.056</td>
<td>0.033</td>
<td>0.023</td>
<td>0.037</td>
<td>0</td>
</tr>
<tr>
<td>Stratification</td>
<td>confound</td>
<td>0.013</td>
<td>0.013</td>
<td>0.001</td>
<td>0.012</td>
<td>0.078</td>
<td>0</td>
</tr>
<tr>
<td>Stratification</td>
<td>instru</td>
<td>0.029</td>
<td>0.034</td>
<td>0.005</td>
<td>0.029</td>
<td>0.100</td>
<td>0</td>
</tr>
<tr>
<td>Stratification</td>
<td>instru_prog</td>
<td>0.016</td>
<td>0.041</td>
<td>0.012</td>
<td>0.029</td>
<td>0.091</td>
<td>0</td>
</tr>
<tr>
<td>Stratification</td>
<td>instru_prog_noise</td>
<td>0.016</td>
<td>0.042</td>
<td>0.012</td>
<td>0.029</td>
<td>0.106</td>
<td>0</td>
</tr>
</tbody>
</table>
## Table 1.5.2: Simulation result table for PSA

<table>
<thead>
<tr>
<th>Implementation</th>
<th>PS model</th>
<th>Empirical var</th>
<th>Avg total var</th>
<th>Bias</th>
<th>Coverage</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caliper matching (1-1)</td>
<td>confound</td>
<td>0.021</td>
<td>0.023</td>
<td>0.060</td>
<td>0.930</td>
<td>0.024</td>
</tr>
<tr>
<td>Caliper matching (1-1) instru</td>
<td></td>
<td>0.042</td>
<td>0.043</td>
<td>0.084</td>
<td>0.930</td>
<td>0.049</td>
</tr>
<tr>
<td>Caliper matching (1-1) instru_prog</td>
<td></td>
<td>0.032</td>
<td>0.041</td>
<td>0.079</td>
<td>0.942</td>
<td>0.038</td>
</tr>
<tr>
<td>Caliper matching (1-1) instru_prog_noise</td>
<td></td>
<td>0.030</td>
<td>0.042</td>
<td>0.082</td>
<td>0.946</td>
<td>0.037</td>
</tr>
<tr>
<td>Caliper matching (1-5) confound</td>
<td></td>
<td>0.017</td>
<td>0.019</td>
<td>0.061</td>
<td>0.926</td>
<td>0.021</td>
</tr>
<tr>
<td>Caliper matching (1-5) instru</td>
<td></td>
<td>0.038</td>
<td>0.039</td>
<td>0.080</td>
<td>0.922</td>
<td>0.044</td>
</tr>
<tr>
<td>Caliper matching (1-5) instru_prog</td>
<td></td>
<td>0.027</td>
<td>0.038</td>
<td>0.077</td>
<td>0.940</td>
<td>0.033</td>
</tr>
<tr>
<td>Caliper matching (1-5) instru_prog_noise</td>
<td></td>
<td>0.028</td>
<td>0.038</td>
<td>0.085</td>
<td>0.932</td>
<td>0.036</td>
</tr>
<tr>
<td>DR</td>
<td>confound</td>
<td>0.014</td>
<td>0.018</td>
<td>0.001</td>
<td>0.954</td>
<td>0.014</td>
</tr>
<tr>
<td>DR</td>
<td>instru</td>
<td>0.053</td>
<td>0.059</td>
<td>0.009</td>
<td>0.924</td>
<td>0.053</td>
</tr>
<tr>
<td>DR</td>
<td>instru_prog</td>
<td>0.033</td>
<td>0.078</td>
<td>-0.008</td>
<td>0.990</td>
<td>0.033</td>
</tr>
<tr>
<td>DR</td>
<td>instru_prog_noise</td>
<td>0.033</td>
<td>0.078</td>
<td>-0.001</td>
<td>0.984</td>
<td>0.033</td>
</tr>
<tr>
<td>IPW</td>
<td>confound</td>
<td>0.016</td>
<td>0.018</td>
<td>0.012</td>
<td>0.954</td>
<td>0.016</td>
</tr>
<tr>
<td>IPW</td>
<td>instru</td>
<td>0.060</td>
<td>0.052</td>
<td>0.028</td>
<td>0.916</td>
<td>0.061</td>
</tr>
<tr>
<td>IPW</td>
<td>instru_prog</td>
<td>0.066</td>
<td>0.056</td>
<td>0.009</td>
<td>0.942</td>
<td>0.066</td>
</tr>
<tr>
<td>IPW</td>
<td>instru_prog_noise</td>
<td>0.069</td>
<td>0.060</td>
<td>0.008</td>
<td>0.938</td>
<td>0.069</td>
</tr>
<tr>
<td>NN matching</td>
<td>confound</td>
<td>0.025</td>
<td>0.031</td>
<td>0.012</td>
<td>0.964</td>
<td>0.025</td>
</tr>
<tr>
<td>NN matching</td>
<td>instru</td>
<td>0.088</td>
<td>0.080</td>
<td>0.011</td>
<td>0.928</td>
<td>0.088</td>
</tr>
<tr>
<td>NN matching</td>
<td>instru_prog</td>
<td>0.065</td>
<td>0.076</td>
<td>0.015</td>
<td>0.960</td>
<td>0.066</td>
</tr>
<tr>
<td>NN matching</td>
<td>instru_prog_noise</td>
<td>0.074</td>
<td>0.075</td>
<td>0.012</td>
<td>0.934</td>
<td>0.074</td>
</tr>
<tr>
<td>Stratification</td>
<td>confound</td>
<td>0.013</td>
<td>0.013</td>
<td>0.067</td>
<td>0.906</td>
<td>0.017</td>
</tr>
<tr>
<td>Stratification</td>
<td>instru</td>
<td>0.034</td>
<td>0.031</td>
<td>0.084</td>
<td>0.912</td>
<td>0.041</td>
</tr>
<tr>
<td>Stratification</td>
<td>instru_prog</td>
<td>0.021</td>
<td>0.031</td>
<td>0.071</td>
<td>0.980</td>
<td>0.026</td>
</tr>
<tr>
<td>Stratification</td>
<td>instru_prog_noise</td>
<td>0.021</td>
<td>0.032</td>
<td>0.084</td>
<td>0.962</td>
<td>0.028</td>
</tr>
</tbody>
</table>
Figure 1.6.1: Histogram of estimated propensity scores under treated and control observations
Figure 1.6.2: Estimated ATEs and constructed 95% confidence/credible intervals created via PSA and BPSA
Table 1.6.1: Application data: table of unadjusted averages of covariates across treatment

<table>
<thead>
<tr>
<th></th>
<th>Unexposed</th>
<th>Exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>15,512</td>
<td>7,211</td>
</tr>
<tr>
<td>Avg. estimated PS</td>
<td>0.25</td>
<td>0.46</td>
</tr>
<tr>
<td>Latitude</td>
<td>38.30</td>
<td>38.37</td>
</tr>
<tr>
<td>Longitude</td>
<td>-82.84</td>
<td>-79.46</td>
</tr>
<tr>
<td>County smoking rate</td>
<td>0.27</td>
<td>0.26</td>
</tr>
<tr>
<td>Total population</td>
<td>10,791</td>
<td>13,231</td>
</tr>
<tr>
<td>Percent residing in rural area</td>
<td>0.58</td>
<td>0.49</td>
</tr>
<tr>
<td>Percent of white residents</td>
<td>0.85</td>
<td>0.82</td>
</tr>
<tr>
<td>Percent of African-American residents</td>
<td>0.11</td>
<td>0.13</td>
</tr>
<tr>
<td>PctHighSchool</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>Median household income</td>
<td>39,385</td>
<td>43,594</td>
</tr>
<tr>
<td>Percent living below poverty threshold</td>
<td>0.13</td>
<td>0.12</td>
</tr>
<tr>
<td>Percent female</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>Percent of housing units occupied</td>
<td>0.87</td>
<td>0.90</td>
</tr>
<tr>
<td>Percent who have moved in past 5 years</td>
<td>0.42</td>
<td>0.42</td>
</tr>
<tr>
<td>Median house value</td>
<td>102,070</td>
<td>121,912</td>
</tr>
<tr>
<td>Population per Sq. Mile</td>
<td>1,167</td>
<td>2,455</td>
</tr>
<tr>
<td>Avg. temperature (2005)</td>
<td>286</td>
<td>286</td>
</tr>
<tr>
<td>Avg. relative humidity (2005)</td>
<td>0.0086</td>
<td>0.0085</td>
</tr>
</tbody>
</table>

Table 1.6.2: Estimated treatment effects of high power plant emissions on ambient pollution

<table>
<thead>
<tr>
<th></th>
<th>ATE [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNADJ regression</td>
<td>1.41 [1.35,1.47]</td>
</tr>
<tr>
<td>ADJ regression</td>
<td>1.46 [1.41,1.51]</td>
</tr>
<tr>
<td>BPSA PSA</td>
<td></td>
</tr>
<tr>
<td>NN matching</td>
<td>1.63 [1.52,1.74]</td>
</tr>
<tr>
<td>Caliper matching</td>
<td>1.59 [1.50,1.68]</td>
</tr>
<tr>
<td>Stratification</td>
<td>1.75 [1.67,1.84]</td>
</tr>
<tr>
<td>IPW</td>
<td>1.74 [1.64,1.85]</td>
</tr>
<tr>
<td>DR</td>
<td>1.95 [1.86,2.05]</td>
</tr>
</tbody>
</table>
Table 1.6.3: Between- and within-design variance under BPSA

<table>
<thead>
<tr>
<th>Method</th>
<th>Total variance</th>
<th>Between-design variance</th>
<th>Within-design variance</th>
<th>PROP&lt;sub&gt;DU&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN matching</td>
<td>0.0029</td>
<td>0.0018</td>
<td>0.0012</td>
<td>0.59</td>
</tr>
<tr>
<td>1-1 Caliper matching</td>
<td>0.0022</td>
<td>0.0011</td>
<td>0.0011</td>
<td>0.49</td>
</tr>
<tr>
<td>DR</td>
<td>0.0024</td>
<td>0.0003</td>
<td>0.0020</td>
<td>0.11</td>
</tr>
<tr>
<td>IPW</td>
<td>0.0012</td>
<td>0.0005</td>
<td>0.0006</td>
<td>0.46</td>
</tr>
<tr>
<td>Stratification</td>
<td>0.0019</td>
<td>0.0002</td>
<td>0.0016</td>
<td>0.13</td>
</tr>
</tbody>
</table>
Posterior Predictive Treatment Assignment Methods for Causal Inference in the Context of Time-Varying Treatments

2.1 INTRODUCTION

The effect of air pollution exposure on health outcomes often requires longitudinal analysis to account for seasonal patterns in pollution emissions and confounders such as temperature and humidity. However, in this context, time-varying covariates simultaneously act as confounders and mediators, resulting in biased effect estimates from standard longitudinal analysis methods whether or not they control for these covariates ([87]).
Developed by [83], marginal structural models (MSMs) are a class of causal models used to describe the marginal distribution of potential outcomes that would occur under different time-varying treatment patterns. Due to their interpretability and ease of use, MSMs have become the dominant method in the analysis of time-varying treatment effects. Typically, a MSM is fit with inverse probability weights (IPWs) which create an unconfounded pseudo-population where each observation represents multiple “copies” of itself. Anchoring IPWs to the probability of an individual experiencing treatment (propensity score) at each time point controls for time-varying confounding of treatment assignment across time points ([84]).

IPW has well-studied asymptotic properties ([43], [103], [102]). However, finite sample performance of IPW-estimated treatment effects are known to suffer when there is low “overlap” of covariate distributions, also referred to as a lack of common support, across different treatment patterns ([87], [23], [76]). Low covariate overlap can arise in cases of strong confounding, where some covariate profiles lead to very high (or low) probabilities of observing a particular treatment pattern, creating propensity score distributions across different treatment patterns with little “overlapping” density. Covariate overlap relates to the assumption of positivity - that all units have a probability bounded away from 0 or 1 of receiving either treatment and control. In extreme cases, the complete lack of covariate overlap violates the positivity assumption and estimands such as the ATE are not identifiable. Low overlap is especially common in the time-varying setting of pollution exposure, where spatial and temporal patterns of treatment lead to strong confounding, and IPW effect estimates have been shown to induce spurious effect estimates ([72]).

To mitigate the consequences of low covariate overlap, [87] has recommended usage of stabilized weights for the time-varying treatment setting, which often display only slightly more robustness compared to IPW and fail in finite samples as the number of time points increases. Other alternatives include weight truncation, the practice of replacing extreme weights with a more moderate truncation level, which could either be chosen in an *ad hoc* or data-driven ([34, 107]) manner, the pruning of observations with extreme propensity scores, including the efficient interval selection method from [27], and modeling of the full conditional distribution of the outcome via augmented weighting, G-computation ([83]), targeted maximum likelihood estimation (TMLE) ([104]) or the residual balancing method of [108].

Besides stabilized weighting, the aforementioned methods either rely on unverifiable
parametric assumptions and extrapolation or change the causal estimand, making interpretation of causal estimates ambiguous. Weight truncation and pruning both fall into the second category, yet are commonly utilized without consideration of what sub-population is targeted after observations are thrown out or weights are modified. For example, the Optimal Subpopulation Average Treatment Effect (OSATE) developed by [27] emerges as a consequence of variance minimization, which may not correspond to any sub-population of real-world relevance. [76] describes the problem with these methods as “trading off proximity to the initial target of inference for identifiability”. This paper seeks to fill the gap in methodology for time-varying treatments with an estimation method which targets a causal estimand that is policy relevant, identifiable, and able to be reliably estimated in settings with limited covariate overlap. We expand upon previous work describing inference for the “average treatment effect in the overlap population” (ATO), defined in the point treatment case by [57] as the average treatment effect over “units whose combination of characteristics could appear with substantial probability in either treatment group”.

In the point treatment case, [23] developed a Bayesian approach to estimate the ATO which utilizes stochastic pruning to iteratively select observations into a pruned subgroup which satisfies requirements for covariate overlap and confounding adjustment. The probability distribution governing membership in the subgroup is based on a quantity called the posterior-predictive treatment assignment (PPTA), which quantifies each unit’s probability to receive the treatment opposite of that which was observed, based on the estimated propensity score. The PPTA procedure falls under the umbrella of Bayesian propensity score analysis (BPSA), allowing it to not only account for design uncertainty arising from propensity score estimation ([59]), but also to average over each observation’s membership probability within the unconfounded subgroup. In the point treatment setting, [23] demonstrated PPTA’s ability to estimate treatment effects with much less variability across replications compared to IP weighted estimates, with or without truncation. [57] presented similar findings by weighting observations with their probability of receiving the opposite treatment, making it a weighting analogue to PPTA.

This paper extends both [23]’s PPTA procedure as well as [57]’s overlap weighting (OW) method to the time-varying treatment setting. Importantly, both methods are compatible with MSMs, retaining the benefits of this modelling framework. In a time-varying treatment setting with low covariate overlap, the ATO may be estimated in a more stable,
This paper begins with a description of standard methodology for the analysis of time-varying treatment data: marginal structural models with IPWs constructed conditional on estimated propensity scores (Section 2.2.2). Section 2.3 describes an extension of the stochastic pruning method developed by [23] utilized to estimate the ATO, while a weighting method by [57] which also estimates the ATO is extended in Section 2.3.4.

Section 2.4 demonstrates the superior performance of the extended PPTA and OW methods against IP and stabilized weights in a simulation study. Finally, we perform an analysis of power plant emission treatment on ischemic heart disease hospitalization rates across 18,488 zip codes in the US with high concentrations of coal-fired plants (Section 2.5). This paper concludes with a discussion in Section 2.6.

2.2 MSM AND COVARIATE OVERLAP WITH TIME-VARYING TREATMENTS

2.2.1 Notation

Assume \( n \) observations across \( d = 1, 2, \ldots, D \) time points, where each observation \( i = 1, 2, \ldots, n \) receives one of two treatments at each time point. Let \( T_{di} \) denote the treatment of the \( i^{th} \) unit at time point \( d \), with \( T_{di} = 1 \) denoting receipt of treatment and \( T_{di} = 0 \) otherwise. The \( n \)-dimensional vector of binary treatments received by all observations at time point \( d \) is \( T_d = [T_{d1} \ldots T_{dn}] \) while a single observation’s treatment history, the set of all treatments taken by observation \( i \) up to and including time point \( d \), may be denoted as \( \overline{T}_{di} = [T_{d1i} \ldots T_{dni}] \). We will refer to the complete treatment history \( \overline{T}_{Di} \) as observation \( i \)'s treatment pattern. The set of all treatment histories to time point \( d \) is defined as \( \overline{T}_d = [\overline{T}_{d1} \ldots \overline{T}_{dn}] \).

For each observation \( i \), \( p_X \) time-varying covariates are available at each time point \( d \), denoted with the vector \( X_{di} = [X_{d1i} \ldots X_{dpi}] \). Let \( X_d = [X_{d1} \ldots X_{dn}] \) represent the \( n \) by \( p_X \) matrix of covariates at time point \( d \), where it is assumed that each \( X_d \) closely proceeds each \( T_d \). All covariate histories at time point \( d \) are then \( \overline{X}_d = [X_{d1} \ldots X_{dn}] \). In addition to time-varying covariates, assume each observation has \( p_W \) fixed baseline covariate values, said to be observed at time 0 in order to imply that they are observed before \( X_d \) and \( T_d \). Let \( W_{oi} = [W_{o1} \ldots W_{opi}] \) be a vector of these covariate values for observation \( i \), and \( W_o = [W_{o1} \ldots W_{on}] \) be an \( n \) by \( p_W \) matrix containing baseline covariate information for all
observations. We assume a situation where the outcome of interest is measured at a single point in time at the end of the study, \( D + 1 \). We omit the time subscript for simplification and denote the outcome with \( Y = [Y_1, Y_2, \ldots, Y_n] \).

### 2.2.2 Marginal Structural Models and assumptions required to estimate a causal contrast

This paper is concerned with estimating a causal effect, which is defined with regards to the contrast in counterfactual outcomes, where \( \bar{t} \) and \( \bar{t}' \) represent differing treatment patterns across \( D \) time points:

\[
E(\bar{Y}_\bar{t}) - E(\bar{Y}_{\bar{t}'})
\]  

(2.1)

While a causal effect can be defined for any \( \bar{t}, \bar{t}' \), the number of possible such contrasts increases exponentially with the number of time points. To simplify, we follow convention and specify an unsaturated marginal structural model (MSM) representing a linear dose-response relationship between treatment and potential outcomes, where the time of dose has no impact on response ([36]):

\[
E(\bar{Y}_\bar{t}) = \alpha_0 + \Delta \text{sum}(\bar{t})
\]  

(2.2)

Commonly, the parameters of a MSM are estimated with inverse probability weighting. Contrasts in potential outcomes such as \( E(\bar{Y}_\bar{t}) - E(\bar{Y}_{\bar{t}'}) \) are estimated with observed-data contrasts such as \( E(Y|\bar{T}_D) - E(Y|\bar{T}'_D) \) by weighting observations to represent a pseudo-population where the assignment to different \( \bar{t} \) is unconfounded. Specifically, we model the marginal mean of counterfactual \( \bar{Y}_\bar{t} \) with observed-data model ([84]):

\[
E(Y|\bar{T}_D) = \alpha_0 + \beta \sum_{d=1}^{D} \bar{T}_d,
\]  

(2.3)

Where \( \beta = \Delta \) when the parameters of (17) are estimated with specific weighting schemes and assumptions, such as the inverse probability of treatment weights (IPW), the assump-
tions of which will be specified. Towards this specification, we follow the common for-
mulation of estimating causal effects for time-varying treatments by conceptualizing the
treatment assignment mechanism as arising from a sequentially randomized trial, where
current treatment is assigned with some probability conditional on past observed treat-
ments and covariate history.

This conceptualization implies that the probability of observing any treatment pattern
\((T_d; t_{d-1}...t_1)\) is equal to the product of the sequential probabilities of treatment assign-
ments, conditional on covariate history, where the term \(T_{d-1}\) disappears when \(d = 1\):

\[ P(\overline{T}_{di} = [t_d; t_{d-1}...t_1]|\overline{X}_{di}, W_o) = \prod_{d=1}^{D} P(T_{di} = t_d|\overline{T}_{(d-1)i}, \overline{X}_{di}, W_o) \]  

(2.4)

Causal inference with observational data in a sequentially-assigned treatment setting
relies on the following three assumptions: that the treatment is “sequentially randomized”,
also known as conditional sequential exchangeability (\(Y^i \perp T_d|\overline{T}_{d-1}, \overline{X}_d, W_o\) for all \(d = 1...D\) when \(\overline{T}_D = \overline{t}\), sequential consistency (\(Y^i = Y\) for any subject with \(\overline{T}_D = \overline{t}\)), and
sequential positivity (\(P(T_d = 1|\overline{T}_{d-1}, \overline{X}_d, W_o)\) is bounded away from 0 and 1, given that
\(P(\overline{T}_{d-1}, \overline{X}_d, W_o) > 0\), for all \(d\) ) ([87]).

We model the unknown “sequentially randomized” treatment assignment assumed in
Expression 2.4 with propensity score (PS) models defined such that treatment at each time
\(d\) is modelled conditionally on past treatments \(\overline{T}_{(d-1)}\), past values of time-varying covari-
ates, \(\overline{X}_d\), baseline covariates \(W_o\), and a vector of parameters, \(a_d\). Let \(f(.)\) represent a generic
model (which we assume throughout the rest of the paper is correctly specified):

\[ P(T_d|\overline{T}_{(d-1)}; \overline{X}_d, W_o) = f(\overline{T}_{(d-1)}, \overline{X}_d, W_o, a_d) \]  

(2.5)

These PS models allow estimation of the propensity score \((\hat{e}_{di})\) or probability of ob-
servation \(i\) receiving treatment rather than control at time \(d\) \((P(T_{di} = 1|\overline{T}_{(d-1)i}, \overline{X}_d, W_o))\),
a term of the product in Expression 2.4.

If the three assumptions stated earlier hold, inverse probability weighting may be uti-
lized to create a pseudo-population where randomization probabilities depend at most on
past treatment history, thereby removing time-varying confounding ([36]). In effect, each
observation is weighted by the inverse of their probability of experiencing the treatment pattern they were observed to experience. The sequential assumptions in Expression 2.4 imply that IPWs can be calculated by taking the product of the inverse conditional probability of receiving the treatment observed at each time point over all time points:

\[ w_i = \prod_{d=1}^{D} \left[ \frac{T_{di}}{P(T_{di} = 1 | \hat{T}_{(d-1)i}, \bar{X}_d, W_{oi})} + \frac{1 - T_{di}}{P(T_{di} = 0 | \hat{T}_{(d-1)i}, \bar{X}_d, W_{oi})} \right] \] (2.6)

\[ = \prod_{d=1}^{D} \left[ \frac{T_{di}}{\hat{e}_{di}} + \frac{1 - T_{di}}{1 - \hat{e}_{di}} \right] \] (2.7)

IPW creates a pseudo-population of combined treated and control populations so that parameter \( \Delta \) in the MSM represents the ATE ([57]).

The drawbacks of utilizing MSM with IPW in settings with low covariate overlap is well-known, especially in the context of time-varying treatments. Large variability in estimated weights results in a few observations which are weighted so highly that they dominate the sample ([84], [26], [23]). In this case, [84] recommends utilizing stabilized weights (SW), which reduces the variability in weights compared to IPW by fixing the mean of weights at one when the PS model is correctly specified ([26]). However, simulations (Section 2.4) demonstrate how variability in SWs is still too high to estimate the ATE in a stable manner.

In attempting to re-create a pseudo-population of treated and control individuals with balanced covariate profiles, IPW prioritizes observations which fall into an area of the PS distribution dominated by observations which experienced the opposite treatment. With IPW in the time-varying treatment context, observations which exhibit low overlap at just one time point may be assigned an extreme weight and dominate the sample (this is discussed in more detail in Appendix 2).

In the time-varying treatment case low covariate overlap is of special concern since multiple time points increase the probability of chance imbalance and the unique formulation of PS models in the time-varying treatment context includes an increasing number of past treatments and covariates at each time point, risking historical predictors acting as instrumental variables if after a certain “lag time” they no longer affect the outcome. This is why a common assumption in the formulation of PS models for time-varying treatments
is to limit the dependence of treatment at time \( d \) to past treatments and covariates from a smaller number of previous time points than the complete conditional probabilities in Expression 2.5 \( (T_{(d-1)}, \tilde{X}_d, W_o) \) ([36, 84]).

This paper limits this dependence to two time points, effectively assuming conditional independence of current treatment to all treatments or time-varying covariate values more than two time point previous:

\[
T_d \perp \bar{T}_{(d-3)}, \tilde{X}_{(d-2)} | T_{(d-1)}, T_{(d-2)}, X_d, X_{d-1}
\]  

(2.8)

The advantage of this assumption is that it allows pooling of information across treatment patterns with different treatment and time-varying covariate values beyond two time points. For time \( d \), this calculates PS distributions with higher overlap between treated and control observations by removing potential instrumental variables and relaxes the sequential positivity assumption. The risk of this assumption is misspecification of the PS model and biased treatment effect estimates if conditional independence does not hold.

This paper formulates \( f(.) \) as a logit link model and lets \( a_d = [\omega_0, \psi_1, \psi_2, \gamma_0, \gamma_1, \gamma'_1] \) where each \( \gamma_d \) is a vector length \( p_X \) and \( \gamma' \) is a vector length \( p_W \). The PS model at time point \( d \) is defined as:

\[
\text{logit}(P(T_{di} = 1)) = \omega_o + \psi_1 T_{(d-1)} + \psi_2 T_{(d-2)} + \gamma_1 X_{di} + \gamma_2 X_{(d-1)i} + \gamma' W_{oi}
\]  

(2.9)

Note that \( a_d \) may be a different length at each time point as treatment assignment models at later time points condition on earlier observed treatments and time-varying covariates. Let \( \tilde{a}_d \) represent the set \( [a_1, ..., a_d] \), where \( \tilde{a} = \tilde{a}_D \).

While the ATE is the most common contrast of interest, in Section 2.3 we describe how the parameter \( \Delta \) in a MSM may represent contrasts over alternate sub-populations depending on the target population which a particular weighting scheme attempts to recreate.

All methodological development in the following sections are made under the MSM specification outlined in Section 2.2.2 and the PS model in Expression 2.9.
2.3 Estimating the ATO in the time-varying treatment setting

When violation of positivity is of concern, we are motivated to target our effect estimate over a subgroup of individuals which exist in a region of substantial covariate overlap at all time points. Estimation of the treatment effect within this sub-population (defined as the average treatment effect on the overlap population or ATO) may also be motivated by this sub-group being of special interest in the policy setting.

2.3.1 Estimating the ATO in the Point Treatment Setting with PPTA

For the point treatment case, [23] developed a stochastic pruning approach for estimating causal effects in settings with limited overlap. Using the mechanics of Bayesian inference, estimation of causal effects follows marginalization over the distribution of a latent variable governing the membership of each observation within a pruned subset of data exhibiting non-confoundedness and high covariate overlap, thereby calculating estimates of the ATO that offer advantages over standard IPW estimation of the ATE in settings with limited overlap.

[23] first conceptualizes an unconfounded subset of the data comprised of observations which display high covariate overlap. Membership status of observations in this unconfounded subset is regarded as unknown and anchored to a quantity called the posterior-predictive treatment assignment (PPTA). The PPTA is distributed according to a Bernoulli distribution with probability of success equal to the estimated propensity score. Denote the PPTA of observation $i$ as $\tilde{T}_i$ and $\tilde{T} = [\tilde{T}_1, ..., \tilde{T}_n]$. For the point treatment setting, the PPTA is linked to membership in the unconfounded subset through a latent variable, $S_i = \mathbb{1}(\tilde{T}_i = 1 - t|T_i = t)$ for $i = 1, 2, \ldots, n$, where $\mathbb{1}(\cdot)$ is the indicator function and $\{i; S_i = 1\}$ represents an unconfounded “overlap subset” of observations for which propensity score distributions for treatment and control observations display substantial overlap and a causal effect can be estimated. Iteratively sampling values of $\tilde{T}$ and $\{i; S_i = 1\}$ amounts to stochastically pruning observations according to their likelihood of having received the treatment opposite of that observed. Averaging causal estimates across many iterations of the unknown “overlap subset” corresponds to estimating what was defined in [57] as the ATO: a causal treatment effect defined in the
observations “whose combination of characteristics could appear with substantial probability in either treatment group”, where “substantial” is defined relative to the distributions of PS under both treatments. Details of this approach are clarified in Section 2.3.2 where we extend to these ideas to the setting of time-varying treatments.

2.3.2 Extending PPTA to the Time-Varying Treatment Setting

For the time-varying setting, we refine the notion of the overlap population in the point treatment case to define the “consistent overlap population” (COP) as the sub-population of individuals whose combination of characteristics may appear with substantial probability in either treatment group at every time point of possible treatment. We extend the PPTA procedure to estimate the average treatment effect in the COP, continuing to refer to this quantity as the ATO.

We extend existing notation to account for time: let \( \tilde{T}_{di} \) represent the PPTA for observation \( i \) at time \( d \), where \( \tilde{T}_{di} \) is governed by the estimated propensity score at time \( d \), estimated conditional on past treatments and covariates to adjust for time-varying confounding. Let the latent quantity \( S_{di} = 1(\tilde{T}_{di} = 1 - t_d|T_{id} = t) \) denote the “overlap state” of the \( i \)th observation at time \( d \), where \( \tilde{T}_{di} \sim \text{Bernoulli}(\epsilon_{di}) \), connecting the PPTA to the propensity score of observation \( i \) at time \( d \). Let \( \tilde{S}_i = [\tilde{S}_{i1}; \ldots; \tilde{S}_{iD}] \) represent the pattern of overlap states for observation \( i \), \( S_d = [S_{d1}; \ldots; S_{dn}] \) represent the overlap states for all observations at time point \( d \) and \( \tilde{S} = [\tilde{S}_1; \ldots; \tilde{S}_n] \) represent the pattern of overlap states for all observations in the data.

While the point treatment case considers \( \{i; S_i = 1\} \) to be the “overlap subset,” the time-varying case uses the PPTA vector \( \tilde{T}_i \) to define a “consistent overlap subset” (COS) of observations estimated to exhibit overlap across all \( D \) time points, that is, the COS is defined for one draw from the PPTA distribution as \( \{i; \prod_{d=1}^{D} S_{di} = 1\} \). This subset of observations represents those for which time-varying treatment assignment is not confounded and for which propensity score distributions overlap between treated and untreated at all time points.

Draws from the posterior predictive distributions of \( \tilde{T}_{di} \) (and \( S_{di} \)) follows from: 1) specifying a propensity score model and corresponding likelihood as in Expression 2.9; 2) specifying a prior distribution for the unknown parameters of the propensity score model \( \pi(a_d) \); and 3) evaluating the posterior distribution of the propensity score parameters.
Draws from this posterior distribution of \( \bar{\alpha} \) (and, by extension, the propensity scores) permits iterative sampling of the COS that marginalizes over estimation uncertainty in the propensity scores in order to obtain a probability distribution for each observation’s unknown membership in the COS, denoted with \( f(\bar{S} | \bar{W}_0, \bar{X}_d, \bar{T}_d) \):

\[
f(\bar{S} | \bar{W}_0, \bar{X}_d, \bar{T}_d) = \prod_{d=1}^{D} f(S_d | \bar{W}_0, \bar{X}_d, \bar{T}_d) \quad (2.10)
\]

\[
= \prod_{d=1}^{D} \int_{a_d} f(S_d | \bar{W}_0, \bar{X}_d, \bar{T}_d, a_d) f(a_d | \bar{W}_0, \bar{X}_d, \bar{T}_d) da_d \quad (2.11)
\]

\[
= \prod_{d=1}^{D} \int_{a_d} f(S_d | \bar{W}_0, \bar{X}_d, \bar{T}_d, a_d) \pi(a_d) \prod_{i=1}^{n} L(T_{di} | \bar{W}_0, \bar{T}_{(d-1)i}, \bar{X}_{di}) da_d \quad (2.12)
\]

This expression relies on the assumption of conditional independence over time, that \( S_d \perp S_{(d-1)} | \bar{W}_0, \bar{X}_d, \bar{T}_d \). This assumption follows the construction of \( S_d \) from propensity scores, which are estimated independently at each time point.

With the above probability distribution for \( f(\bar{S} | \bar{W}_0, \bar{X}_d, \bar{T}_d) \), estimation of causal effects follows from evaluating a causal contrast conditional on each draw of \( \bar{S} \), and averaging these contrasts over the posterior-predictive distribution of this quantity. The result is estimates of the ATO, \( \Delta \), a causal contrast that weights observations in accordance with their posterior probability of membership in the COP:

\[
f(\Delta | \bar{W}_0, \bar{X}, \bar{T}, \bar{Y}) \quad (2.13)
\]

\[
= \int_{\bar{S}} f(\Delta | \bar{W}_0, \bar{X}, \bar{T}, \bar{Y}, \bar{S}) f(\bar{S} | \bar{W}_0, \bar{X}_d, \bar{T}_d) d\bar{S} \quad (2.14)
\]

\[
= \int_{\bar{S}} f(\Delta | \bar{W}_0, \bar{X}, \bar{T}, \bar{Y}, \bar{S}) f(\bar{S} | \bar{W}_0, \bar{X}_d, \bar{T}_d) d\bar{S} \quad (2.15)
\]

Where \( f(\Delta | \bar{W}_0, \bar{X}, \bar{T}, \bar{Y}, \bar{S}) \) is a normal distribution as modeled by the MSM, with parameters estimated conditional on the COP as defined by \( \bar{S} \). The computational procedure which follows describes how to conduct such marginalization in order to make draws from the posterior marginalized distribution of \( \Delta \) and estimate the ATO. This is a special case of marginalizing over “design uncertainty” as described in [59].
2.3.3 PPTA Computation

This section outlines the computational procedure of PPTA. In the first stage, multiple draws are taken from \( f(\bar{S}|\bar{W}_o, \bar{X}_d, \bar{T}_d) \) using the following procedure:

1. For each time point \( d \) in \( 1 \ldots D \):
   (a) Obtain a sample of \( K \) draws from the posterior distribution of the parameters \( (a_d) \) of the propensity score model. In this paper, we utilize the PS model from Expression 2.9. Samples \( a_{dk}, k = 1 \ldots K \) are taken with standard MCMC routines, like the R package MCMCpack.
   (b) For each of the \( K \) draws from posterior distribution of \( a_d \):
      i. Calculate a set of \( n \) propensity scores \( (e_{dk}) \) with an expit transform.
      ii. Draw the posterior-predictive treatment assignment for all observations from a Bernoulli distribution as in Expression 7.

The entirety of step 1 may be conceived as taking \( K \) draws from the posterior predictive distribution of \( S_d \). Estimation of the treatment effect is conducted after PPTA has been identified for all observations in all time points.

2. For each \( k = 1 \ldots K \):
   (a) Prune all observations except ones for which \( \prod_{d=1}^{D} S_{dk} = 1 \). Observations which remain are in the estimated COS for iteration \( k \).
   (b) Conditional upon this sub-population, calculate the ML estimate of \( \hat{\Delta}_k \) with the MSM in Expression 17.

Each \( \hat{\Delta}_k \) may be thought of as a sample from the marginal posterior distribution \( f(\Delta|\bar{W}_o, \bar{X}, \bar{T}, \bar{Y}) \). If desired, a point estimate may be calculated by taking an average over samples. A standard error which produces expected Frequentist coverage may be estimated with a bootstrapping procedure which draws multiple samples with replacement from the original dataset and performs the entire PPTA procedure on each bootstrapped sample.
2.3.4 A weighting analogue to PPTA

Overlap weighting (OW) was developed in the point-treatment setting by [57] and extended to the time-varying treatment setting in this paper. The formula for overlap weights at each time point are exactly equal to the posterior-predictive treatment probability as described in Expression 2.8, making it a weighting analogue to the PPTA procedure. Construction of the overlap weight allows it to target the ATO, also similar to PPTA. However, while the PPTA procedure is probabilistic and a natural fit for implementation with Bayesian computation in order to marginalize over the posterior distribution of PS, OW is calculated conditional on a maximum likelihood estimate of the PS at each time point. The final OW for each individual is calculated as a product of all time-specific weights:

\[ w_{i,OW} = \prod_{d=1}^{D} T_{di}(1 - \hat{e}_{di}) + (1 - T_{di})\hat{e}_{di} \]

After the OW are calculated, they may be utilized in the same manner as IPW to create an unconfounded pseudo-population upon which parameters of the MSM (representing the ATO) may be estimated.

2.4 Simulation

The following simulation study compares PPTA performance against that of IPW, SW and OW. We investigate settings with a homogeneous treatment effect across the entire covariate distribution, as well as settings where the treatment effect is heterogeneous with regard to observations’ likelihood of membership in the COP. We also demonstrate how PPTA may be utilized post-hoc to quantify the probability of an observation’s membership in the COP in an intuitive and data-driven manner.

2.4.1 Data generation

All simulated data assume that the PS model in Expression 2.9 and the MSM in Equation 17 are correctly specified. Let \( n = 5000 \) be the sample size of each dataset, and \( p_X = p_W = 3 \) be the number of time-varying covariates and baseline covariates (respectively) included in the data generation.
Covariate and treatment data are simulated as follows:

1. Simulate $p_W$ baseline covariates from independent normal distributions centered at 0 with variance 1.

2. For all time points $d$ in $[1, D]$:
   (a) Simulate $p_X$ time-varying covariates from a $MVN(\mu_{d}, \Sigma_{X,d})$ where $\Sigma_{X,d}$ has 1s along the diagonal and no correlation between covariates and $\mu_{d} = [\mu_{X1,d}, \ldots, \mu_{Xp,d}]$.
   Time-varying covariates depend on previously observed treatments and past covariate values from up to two time points previous. Let $\mu_{i,r,d} = \tau_{T1}T_{i,(d-1)} + \tau_{T2}T_{i,(d-2)} + \tau_{X1}X_{i,r,(d-1)} + \tau_{X2}X_{i,r,(d-2)}$ where $r = 1\ldots p_{X1}$, $\tau_{T1} = \tau_{X1} = 0.2$ and $\tau_{T2} = \tau_{X2} = 0.1$.
   (b) Calculate “true” propensity scores with $e_d = \text{expit}(a_0 + a_W W_0 + a_X X_{d-1} + a_{X1} T_{d-1} + a_{T1} T_{d-2} + a_{X2} X_{d-1} + a_{T2} T_{d-2})$ where $a_0 = 0$, $a_W = [0.3, 0.3, 0.3]$, $a_X = [1, 1, 1]$, $a_{X1} = [0.5, 0.5, 0.5]$, $a_{T1} = [0.5, 0.5, 0.5]$, and $a_{T2} = [0.3, 0.3, 0.3]$.
   (c) Simulate treatment assignment from $T_d = \text{Bernoulli}(e_d)$.

**Homogeneous treatment effect**

For the setting with a homogeneous treatment assignment, outcomes are simulated as follows:

3. Simulate outcome with the model $Y = \beta_o + \Delta \sum_{d=1}^{D} T_d + \beta_W W_0 + \sum_{d=1}^{D} \beta_{X,d} X_{d} + \epsilon$ where $\beta_o = -1$, $\Delta = 0.5$, $\beta_W = [0.3, 0.3, 0.3]$ and $\epsilon \sim N(0, 1)$. $\beta_{X,d} = \frac{0.1}{D-d+1}$ represents a decreasing association between outcome and time-varying covariate $X_{d}$ over time.

We simulate 240 datasets each for $D = 3$ and $D = 5$. In this setting, the ATE and ATO are equivalent, and both equal to 0.5.

**Heterogeneous treatment effect**

To isolate the implications of heterogeneity in treatment effects across membership in the overlap population, we provide an additional data generation that uses a proxy for COP membership directly in the data-generating outcome model. This generation exaggerates
the relationship between overlap and treatment effects in order to illustrate PPTA and OW performance when the ATO does not equal the ATE, without intention to reflect realistic data generation.

Covariates and treatments are simulated as before, but with outcomes simulated as follows:

3. Simulate outcome with the model $Y = \beta_0 + \Delta \sum_{d=1}^{D} T_d O_d^* + \beta_W W_0 + \sum_{d=1}^{D} \beta_X X_d + \varepsilon^*$. Where $O_d^*$ represents membership in something akin to the overlap population.

Let $O_{di} = 1$ if observation $i$ falls within a quantile (of 20) of the the distribution of the true propensity score ($e_{di}$) where more than 10% of observations are either treated or control at time $d$. From this point forward, we will refer to observations with $O_d^* = \prod_{d=1}^{D} O_{di} = 1$ as members of the data-generated COP (DGCOP).

This simulates data where only members in the DGCOP experience a treatment effect of 0.5, while all other observations experience a null effect. Note that the manner in which overlap state is determined in the data-generating mechanism is different than the manner in which the model for overlap state in PPTA or OW.

We simulate 240 datasets each for $D = 3$ and $D = 5$.

2.4.2 ANALYSIS METHODOLOGY

IPW, SW, OW and PPTA are utilized in conjunction with the same marginal structural model. IP weights are calculated as specified in Expression 2.6 conditional on the ML-estimated $\hat{e}_d$. SW are calculated on the same set of ML-estimated PS following the formula described in [84]. PPTA was implemented as described in Section 2.3.3 with $K = 1500$. Unlike PPTA, the weighting methods do not draw multiple sets of PS from its posterior predictive distribution (thereby marginalizing over the uncertainty in PS estimates ([59])), instead conditioning treatment effect estimation upon one ML-estimated set of PS at each time point. Bootstrap SEs are calculated for all methods, as they have been found to produce approximately correct SEs for IPW in the longitudinal context ([16]) and was found in simulations to produce approximately equal SE estimates as the robust sandwich estimator.
Table 2.4.1: Simulation results for homogeneous treatment effect data generation

<table>
<thead>
<tr>
<th>Method</th>
<th>Time points</th>
<th>True Delta</th>
<th>Bias</th>
<th>Emp SE</th>
<th>Boot SE</th>
<th>Boot Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPW</td>
<td>3</td>
<td>0.5</td>
<td>0.034</td>
<td>0.132</td>
<td>0.078</td>
<td>0.763</td>
</tr>
<tr>
<td>OW</td>
<td>3</td>
<td>0.5</td>
<td>-0.019</td>
<td>0.042</td>
<td>0.041</td>
<td>0.907</td>
</tr>
<tr>
<td>PPTA</td>
<td>3</td>
<td>0.5</td>
<td>-0.019</td>
<td>0.042</td>
<td>0.041</td>
<td>0.899</td>
</tr>
<tr>
<td>SW</td>
<td>3</td>
<td>0.5</td>
<td>0.036</td>
<td>0.113</td>
<td>0.066</td>
<td>0.747</td>
</tr>
<tr>
<td>IPW</td>
<td>5</td>
<td>0.5</td>
<td>0.062</td>
<td>0.177</td>
<td>0.106</td>
<td>0.700</td>
</tr>
<tr>
<td>OW</td>
<td>5</td>
<td>0.5</td>
<td>-0.020</td>
<td>0.064</td>
<td>0.059</td>
<td>0.921</td>
</tr>
<tr>
<td>PPTA</td>
<td>5</td>
<td>0.5</td>
<td>-0.019</td>
<td>0.066</td>
<td>0.062</td>
<td>0.907</td>
</tr>
<tr>
<td>SW</td>
<td>5</td>
<td>0.5</td>
<td>0.065</td>
<td>0.116</td>
<td>0.063</td>
<td>0.490</td>
</tr>
</tbody>
</table>

2.4.3 Results

Table 2.4.2: Simulation results for heterogeneous treatment effect data generation

<table>
<thead>
<tr>
<th>Method</th>
<th>Time points</th>
<th>True Delta</th>
<th>Bias</th>
<th>Empirical SE</th>
<th>Bootstrap SE</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPW</td>
<td>3</td>
<td>0.16</td>
<td>0.035</td>
<td>0.118</td>
<td>0.081</td>
<td>0.747</td>
</tr>
<tr>
<td>OW</td>
<td>3</td>
<td>0.37</td>
<td>-0.001</td>
<td>0.048</td>
<td>0.043</td>
<td>0.898</td>
</tr>
<tr>
<td>PPTA</td>
<td>3</td>
<td>0.37</td>
<td>-0.001</td>
<td>0.048</td>
<td>0.043</td>
<td>0.906</td>
</tr>
<tr>
<td>SW</td>
<td>3</td>
<td>0.16</td>
<td>0.045</td>
<td>0.108</td>
<td>0.069</td>
<td>0.664</td>
</tr>
<tr>
<td>IPW</td>
<td>5</td>
<td>0.10</td>
<td>0.081</td>
<td>0.138</td>
<td>0.095</td>
<td>0.616</td>
</tr>
<tr>
<td>OW</td>
<td>5</td>
<td>0.33</td>
<td>-0.014</td>
<td>0.071</td>
<td>0.070</td>
<td>0.938</td>
</tr>
<tr>
<td>PPTA</td>
<td>5</td>
<td>0.33</td>
<td>-0.011</td>
<td>0.072</td>
<td>0.073</td>
<td>0.938</td>
</tr>
<tr>
<td>SW</td>
<td>5</td>
<td>0.10</td>
<td>0.082</td>
<td>0.094</td>
<td>0.060</td>
<td>0.463</td>
</tr>
</tbody>
</table>

Homogeneous treatment effects

Table 2.4.1 contains measures of average bias, empirical standard deviation of point estimates across replications, bootstrapped standard error estimates and coverage under intervals created with bootstrapping. Under the homogeneous treatment effect data generation mechanism, the ATE is synonymous with the ATO, and the rest of this section will refer to these estimands interchangeably.
As seen in Table 2.4.1, the empirical variability of point estimates across replications increases with the number of time points for all analysis methods, an expected result considering that increasing the number of treatment patterns decreases covariate overlap on average when the sample size is held constant. Within each set of time points, estimates of ATE obtained with IPW and SW display larger variability (0.118 and 0.108 for 3 time points, respectively and 0.138 and 0.094 for 5 time points, respectively) over replications compared to OW and PPTA (0.048 and 0.048 for 3 time points, respectively and 0.071 and 0.072 for 5 time points, respectively). This result is consistent with previous findings in the point treatment case by [23].

Since PPTA is a stochastic analogue to OW, its posterior mean for the distribution of \( \Delta \) approach the point estimates of \( \Delta \) obtained by OW. However, as the number of time points increases, these two quantities differ due to decrease in covariate overlap leading to additional stochasticity in the propensity score distribution and the MCMC. These factors also effect empirical variability, average bias and bootstrap SEs. A reduced number of simulations with \( K = 10,000 \) resulted in closer agreement between the PPTA and OW approaches, indicating the likelihood that the discrepancy is due at least in part to monte carlo error in the MCMC.

Bootstrap SEs for PPTA and OW are closer to the empirical SD of treatment effect estimates across replications, and coverage for PPTA and OW are closer to nominal levels than for IPW and SW. Furthermore, coverage decreases and average bias increases for IPW and SW from 3 to 5 time points, while neither is observed for OW and PPTA.

**Heterogeneous treatment effects**

Table 2.4.2 presents simulation results under the heterogeneous treatment effect data generation where the ATO is not equal to the ATE. Recall that with heterogeneous treatment effects the PPTA and OW approaches are not designed to estimate the same quantity.
as the IPW and SW approaches. Bias and coverage measures are calculated with respect to the ATE for IPW and SW methods and to the ATO for PPTA and OW methods, where the true ATE is calculated by multiplying $\Delta = 0.5$ with the proportion of observations in the DGCOP as estimated on a dataset of $n = 9 \times 10^6$ (31% for 3 time points and 12% for 5 time points), and the ATO was estimated by calculating overlap weights on a large dataset ($n = 9 \times 10^9$) with true propensity scores.

Similar to observed performance in Section 2.4.3, treatment effects estimated by IPW and SW are much more variable across replications compared to OW and PPTA estimated treatment effects. Mirroring findings from Section 2.4.3, OW/PPTA display lower bias for estimating the ATO than IPW/SW does for estimating the ATE. Coverage rates with bootstrapped intervals are closer to nominal for OW/PPTA while IPW/SW intervals result in under-coverage, particularly in the 5 time point case.

2.4.4 Post-Hoc Examination of the COP

It is complex to identify members of the COP via examination of the overlap weights themselves. On the other hand, PPTA does provide an intuitive metric for identifying members of the target subgroup of interest as well as providing an empirical measure of how much overlap exists in the sample.

Table 2.4.3 includes the average size and SD of COS across iterations as well as the proportion of observations utilized by the PPTA procedure in at least one COS. Over all iterations, PPTA utilizes on average 67% of the 3 time point data with an average COS size of 125, but only only utilizes on average 18% of the 5 time point data with an average COS size of just 9. This is an indication that poor covariate overlap exists in the 5 time point case and would not support a causal estimate if the number of time points were to increase.

2.5 Causal effect of treatment to power plant emission on ischemic heart disease hospitalization rate at the zip-code level

Ischemic heart disease (IHD), also known as coronary artery disease, is the most common of all coronary diseases, affecting over 3 million Americans each year ([73]). Globally, IHD and associated cardiovascular diseases cause one-third of all deaths in people
over 35 ([94]).

The link between cardiac mortality and morbidity and air pollutants has been well-studied ([62, 72, 78, 79]). Recently, advanced air pollution models have been used to quantify pollutant exposure from specific point sources, such as coal-fired power plants ([42]), and establish a significant relationship to IHD hospitalizations in the Northeast, Industrial Midwest and Southeast regions of the United States ([28]).

Little literature ([101]) exists on estimating a causal effect of pollution exposure accounting for the fluctuation of exposure levels across seasons, as well as its complicated longitudinal associations with temperature and humidity. This analysis seeks to fill that gap by combining a recently-developed reduced complexity air quality transport model with causal analysis of time-varying treatments in order to estimate the causal relationship between power plant emission treatment and IHD hospitalization rate at the zip code level, controlling for seasonal effects. In this analysis, we are motivated to target the ATO due to geographical characteristics which result in many areas of the US not experiencing drastic fluctuations in pollution exposure (as seen in Figure 2.5.1). This results in strong confounding and low covariate overlap, since many areas exhibit characteristics inconsistent with experiencing more than one treatment level and thus contribute no information to estimating treatment contrasts (see Section 2.5.2 and Table 2.5.2). By performing this analysis with PPTA we are able to target the ATO, an estimand which is identifiable and valuable from a policy standpoint, that of the treatment effect on the sub-population most likely to receive a variety of treatment patterns.

2.5.1 Data description

The dataset comprises of 18,480 zip code-level observations from the Southwest, Northwest and Industrial Midwest regions of the continental United States. The outcome of IHD emission rates per 10,000 person-years came from 2012 Medicare records.
Table 2.5.1: Table of baseline covariates.

<table>
<thead>
<tr>
<th>Confounder name</th>
<th>Mean for 0 times of treatment</th>
<th>Mean for 1 times of treatment</th>
<th>Mean for 2 times of treatment</th>
<th>Mean for 3 times of treatment</th>
<th>Mean for 4 times of treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area (sq mi)</td>
<td>59.8</td>
<td>49.3</td>
<td>46.9</td>
<td>40.1</td>
<td></td>
</tr>
<tr>
<td>Total population</td>
<td>10,571</td>
<td>9,932</td>
<td>8,136</td>
<td>9,516</td>
<td></td>
</tr>
<tr>
<td>Proportion of women</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Proportion of Hispanic Americans</td>
<td>0.06</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>Proportion of American Indians</td>
<td>0.007</td>
<td>0.003</td>
<td>0.002</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>Proportion of Asian Americans</td>
<td>0.02</td>
<td>0.01</td>
<td>0.008</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td>Proportion of African Americans</td>
<td>0.13</td>
<td>0.10</td>
<td>0.04</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>Proportion of Pacific Islanders</td>
<td>0.0004</td>
<td>0.0003</td>
<td>0.0002</td>
<td>0.0003</td>
<td></td>
</tr>
<tr>
<td>Proportion of White Americans</td>
<td>0.81</td>
<td>0.86</td>
<td>0.93</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Proportion of “other” races</td>
<td>0.02</td>
<td>0.01</td>
<td>0.006</td>
<td>0.008</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.5.1: Dichotomized treatment maps for winter (top left), spring (top right), summer (bottom left) and winter (bottom right) 2012

Figure 2.5.2: COP as identified by PPTA

Observations with a non-zero probability of being in the COP

Consistent overlap population as identified by PPTA
Table 2.5.2: Treatment pattern prevalence after dichotomization

<table>
<thead>
<tr>
<th>Winter treatment</th>
<th>Spring treatment</th>
<th>Summer treatment</th>
<th>Fall treatment</th>
<th>Number of observations</th>
<th>Number of observations contributing to PPTA estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9695</td>
<td>875</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>200</td>
<td>92</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4001</td>
<td>1785</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>773</td>
<td>509</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>670</td>
<td>26</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>663</td>
<td>36</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>108</td>
<td>108</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>212</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2065</td>
<td>30</td>
</tr>
</tbody>
</table>
Figure 2.5.3: Distribution of overlap weights

Overlap Weights - regional distribution

Figure 2.5.4: Distribution of IP weights

IP Weights - regional distribution
Table 2.5.3: Estimates of the causal effect of an additional season of high exposure to air pollution on IHD admissions (per 10,000 person-years) at the zip-code level

<table>
<thead>
<tr>
<th>Method</th>
<th>TE</th>
<th>Bootstrap SE</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPW</td>
<td>37.9</td>
<td>17.2</td>
<td>[4.2, 71.7]</td>
</tr>
<tr>
<td>PPTA</td>
<td>11.8</td>
<td>6.8</td>
<td>[-1.5, 25.2]</td>
</tr>
<tr>
<td>OW</td>
<td>11.0</td>
<td>6.7</td>
<td>[-2.2, 24.1]</td>
</tr>
<tr>
<td>SW</td>
<td>37.8</td>
<td>16.9</td>
<td>[4.7, 71.0]</td>
</tr>
</tbody>
</table>

Table 2.5.4: Distribution of weights for IPW, SW and OW and posterior probability of inclusion for PPTA calculated on application data

<table>
<thead>
<tr>
<th></th>
<th>IPW</th>
<th>SW</th>
<th>OW</th>
<th>PPTA (posterior probability of inclusion)</th>
</tr>
</thead>
<tbody>
<tr>
<td>75% percentile</td>
<td>6.10</td>
<td>0.94</td>
<td>0.0002</td>
<td>0.000</td>
</tr>
<tr>
<td>95% percentile</td>
<td>22.6</td>
<td>2.83</td>
<td>0.0064</td>
<td>0.0067</td>
</tr>
<tr>
<td>99% percentile</td>
<td>95.0</td>
<td>6.67</td>
<td>0.022</td>
<td>0.023</td>
</tr>
<tr>
<td>Maximum</td>
<td>$5.2 \times 10^8$</td>
<td>$1.7 \times 10^8$</td>
<td>0.33</td>
<td>0.35</td>
</tr>
<tr>
<td>% data used</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>20.5%</td>
</tr>
</tbody>
</table>

The continuous treatment metric characterizes exposure to $SO_2$ emissions from power plants and atmospheric byproducts of $SO_2$ reaction in the atmosphere. $SO_2$ is a gas emitted as a byproduct of coal burning. In the atmosphere, $SO_2$ readily converts to sulfate, which contributes to elevated particulate matter (e.g., $PM_{2.5}$) concentrations. Both gaseous $SO_2$ and particulate matter have been linked to adverse health effects.

Individual power plant contributions were mapped with HYSPLIT, an air parcel trajectory and dispersion model which simulates dispersions of parcels emitted every six hours from all coal-fired electricity generating units in the continental USA. Parcel dispersion was informed by wind speeds and directions in 2012 from the National Center for Environmental Prediction/National Center for Atmospheric Research reanalysis project ([51]) and followed for up to 10 days. Finally, parcel concentration was calculated within a fine grid, spacially averaged across grids within the boundaries of each zip code and summed across all contributing power plants in order to obtain total monthly treatment concentrations at the zip code level. Further details of exposure modelling and the evaluation used
Figure 2.5.5: Distribution of stabilized weights

Stabilized Weights - regional distribution

Baseline covariates (a list of which can be seen on Table 2.5.1) come from 2012 Business Analyst Census data. Table 2.5.1 shows that zip codes with high exposure for more seasons are smaller in area and population, and have a less diverse population compared to zip codes with high exposure for fewer seasons. Time-varying covariates included are average temperature and humidity for each season, taken from [51].

2.5.2 Seasonal treatment patterns

Visualizing regional patterns of dichotomized treatment for winter, spring, summer and fall (Figure 2.5.1) in 2012 reveals high treatment around the Ohio river valley during
winter/fall months and a larger footprint of high treatment ratiating outward in warmer months as both energy consumption increases and atmospheric conditions contribute to increased emissions transport ([13, 95]). Table 2.5.2 details the prevalence of treatment patterns which were observed in the data, as well as how many observations within each treatment pattern contributed to the PPTA estimate.

Regional differences in exposure levels across time motivates targeting the ATO rather than the ATE. In this data context, zip codes which are more likely to fall within COP are those which experience seasonally high treatment rather than pervasively high treatment to emissions since they are the zip codes most likely to have similar covariate profiles when exposed and unexposed.

### 2.5.3 Analysis results

Estimated effects of coal emissions exposure on IHD hospitalization rates as well as 95% intervals are presented in Table 2.5.3. The average effect across all zip codes of experiencing an additional season of high exposure to power plant particle emissions is a 37.9 (95% CI: [4.2, 71.7]) increase in rate of IHD emissions per 10,000 person-years as estimated via IPW and a 37.8 (95% CI: [4.7, 71.0]) increase as estimated via SW. The average effect among zip codes in the COP of experiencing an additional season of high exposure to power plant emissions is an increase of 11.8 (95% CI: [-1.5, 25.2]) in rate of IHD emissions per 10,000 person-years as estimated by PPTA and an increase of 10.0 (95% CI: [-2.2, 24.1]) as estimated by OW.

We conclude from this analysis that the treatment effect is lower among the COP compared to the general population. Furthermore, 95% confidence intervals constructed for PPTA and OW from bootstrapped SE estimators include 0 while CI for IPW and SW do not, suggesting a lack of evidence for a significant causal effect of power plant emissions on IHD hospitalizations within the COP, while there is evidence for a significant causal effect among the general population.

To interrogate why the results of PPTA/OW analyses and those of IPW/SW analyses differ, Table 2.5.4 contains information on the spread of weights for IPW, SW and OW as well as probability of inclusion for PPTA. Both IPW and SW display very extreme weights, with a maximum of $5 \times 10^8$ for IPW and $2 \times 10^8$ for SW. These maximum weights carry extreme influence in estimation of the ATE under these methods considering that 99% of
IPW are less than 95% and 99% of SW are less than 6.7.

Overlap weights and the PPTA probability of inclusion have nearly identical percentile spreads, with OW assigning the majority of the sample a negligible weight while PPTA does not include 79.5% of observations in the analysis at all (Table 2.5.4).

PPTA’s inclusion status of observations in the COS makes it easy to visualize observations likely to be in the COP. Figure 2.5.2 colors observations based on whether they ever appear in the COS with zip code i colored in yellow if \( \prod_{d=1}^{D} S_{dik} = 1 \) for any iteration \( k \) and colored in blue otherwise. ZIP codes which have substantial probability of receiving a variety of treatment patterns tend to border the Ohio river valley. Comparing this to Figure 2.5.1 provides valuable information about regions which receive consistent exposure, regions which are not likely to receive any exposure, and the regions are most likely to alternate between exposed and unexposed over time. Figure 2.5.3, which maps the overlap weights to zip codes, also displays a similar pattern, with the highest weighted zip codes appearing in Wisconsin, Michigan, Illinois and the border of the Ohio river valley. Note that Figure 2.5.2 is a binary classification of the same continuous illustration in Figure 2.5.3.

For comparison, maps for the weights from the IPW and SW analyses across zip codes appear in Figures 2.5.4 and 2.5.5. Unlike maps for PPTA and OW, highly weighted zip codes under IPW and SW do not seem to display a consistent geographic pattern, and the two maps do not seem to resemble each other even though IPW and SW prioritize the same characteristics in observations they up-weight. Patterns in SW more closely resemble that of overlap weighting, while IPW highly weights zip codes in Kentucky, Tennessee, Alabama, Georgia, Pennsylvania and Michigan. See Appendix 2 for further discussion on how weight assignments differ between IPW and OW.

2.6 DISCUSSION

This paper extends an estimand for the effect of time-varying treatments which is identifiable and able to be reliably estimated in finite samples when low covariate overlap is present as well as describing the treatment effect over an interpretable, policy-relevant subgroup, the COP. We extend two methods for estimating the ATO to the time-varying treatment setting within a MSM framework - the Bayesian stochastic pruning method PPTA as well as a weighting analogue OW.

Estimating the ATO over the COP retains all the advantages of estimating the ATO over
the overlap population in the point treatment setting. Since the “true propensity score” for all observations in the COP is bounded away from 0 and 1 at all time points, the COP sub-population naturally describes observations which fulfill the sequential positivity assumption. Furthermore, policy relevance of the estimation of treatment effect over the COP is clear, as this sub-set of the population may be the most susceptible to future policy interventions which affect the exposure.

Simulations in Section 2.4 demonstrate how the ATO may be estimated via PPTA or OW with less bias and less variability across replications than the ATE via IPW or SW, and that bootstrapped standard errors for ATO result in uncertainty intervals exhibiting nearer to nominal coverage. These trends persist both when the treatment effect is homogeneous or heterogeneous across the covariate space.

When the treatment effect is heterogeneous across the covariate distribution, the ATO may not be substituted as an estimate of the ATE. However, in many cases, particularly with time-varying treatment data, the ATE may not be identifiable if low overlap which results in a violation of the positivity assumption. In this case, the ATO may be an attractive option over standard pruning or weight truncation methods which change the estimand in a less intuitive manner.

Since all four methods up-weight observations with a lower probability of receiving their own treatment pattern, it may be difficult to understand how these weighting schemes display such different performance. As discussed in further detail in Appendix 2, observations are only assigned a large OW/PPTA probability if they have a large to moderate probability of receiving the opposite treatment than observed at all time points, resulting in large weights/inclusion probabilities being assigned to observations which display consistently high overlap. However, observations may be assigned an extreme IPW/SW based on their PS at a single time point, and observations exhibiting low covariate overlap are commonly assigned extreme IPW.

While IPW, SW and OW use 100% of the sample, in the context of low covariate overlap most of the sample is so down-weighted that their effect on the final estimate is negligible. PPTA provides an intuitive manner to quantify this through the proportion of the entire sample which falls in the COP. By examining the sub-set of observations who contribute to the PPTA estimate, the investigator may be able to identify characteristics of these observations which inform their status in the consistent overlap population. The PPTA procedure is also advantageous for marginalizing over the posterior distribution of PS, thus
accounting for an additional level of uncertainty stemming from the PS estimation procedure ([59]).

There are two important features to estimation of the ATO via PPTA in the time-varying treatment setting. First, PPTA is not a wholly Bayesian procedure and does not recover the posterior distribution of treatment effects with expected Bayesian validity (more details of this may be found in [59]). Boostrapped standard errors are utilized for PPTA, but as in the case of OW they do not create confidence intervals which reach nominal Frequentist coverage in simulations. The asymptotic properties of bootstrapped standard errors in the context of time-varying treatments is not well understood, but has been found in survival literature to perform better in simulations than robust sandwich estimators ([16]), which are available for all weighing methods but not PPTA.

Second, PPTA relies on MCMC draws from the distribution of overlap statuses at each time point. In cases of low covariate overlap or a high number of time points, pruned subsamples within a single iteration may be too small to support treatment effect estimation. In Section 2.4.3, Table 2.4.3 reports the average size of pruned samples under 5 time points to be 9.2. We found that performing PPTA on data simulated under the same procedures detailed in Section 2.4.1 for 7 time points resulted in many COS with no observations. It can be argued that this behavior of PPTA is responding to the lack of covariate overlap in the data, and even though weighted estimates for ATO and ATE may be obtained under these circumstances, they should not be trusted.

Applying PPTA, OW, IPW and SW to an causal analysis of seasonal air pollution exposure on IHD hospitalization rates provided evidence that an increased season of high pollution exposure resulted in higher hospitalization rates over the entire population, but not the COP. By mapping membership in the COS (estimated via PPTA) to zip code locations, we are able to visualize that regions of the US likely to be in the COP border the Ohio river valley. One limitation of the analysis is the loss of information resulting from dichotomization of the continuous pollution exposure, which was necessary to use the method presented in this paper. A promising avenue for future research is extending PPTA for use with continuous time-varying exposure levels with generalized propensity scores.
Flexible modeling of the causal effect of ozone pollution on violent crime with BART

3.1 Introduction

Many epidemiological studies have found evidence associating air pollution exposure with psychological disorders such as suicide \([17, 55, 60, 99]\), psychiatric emergencies \([89]\), anxiety \([35, 80]\) and depression \([98, 105]\). The social costs of air pollution may already be substantial, and are likely to grow in the face of climate change \([66]\).

The US Environmental Protection Agency (EPA) sets National Ambient Air Quality Standards (NAAQS) for air pollution based on six key pollutants, one of which is ozone \([7]\). Ozone, or \(O_3\), is a naturally occurring element in the earth’s atmosphere but is harmful to human health at the ground-level \([18, 21, 33, 81]\).
This study investigates the causal effect of ozone exposure on violent crime. While climatic and environmental variables have been considered as factors driving the seasonal trends in crime as early as 1904 [31], and the relationship between air pollution and rates of homicide and suicide have been written about as early as 1948 [29], academic consensus on this subject is far from over.

Confirming a relationship between ozone and crime is complicated by the fact that violent crime displays a well-studied seasonal pattern and a complicated, non-linear relationship with weather variables, especially temperature [25, 82]. Meanwhile, ozone at the ground level is formed by chemical reactions between nitrogen oxides and volatile organic compounds emitted from cars, power plants and other pollution emission sources [9]. Ozone is typically highest in urban areas on hot days, but may reach high levels in colder months and easily transported by wind [9]. There is a gap for analyses of ozone’s effect on violent crime which utilize flexible non-parametric modeling techniques to account for confounding from these complex relationships.

3.1.1 Neurobiological effects of ozone

Ozone is highly reactive and triggers oxidative stress in the airways and impairs pulmonary function. Due to its reactivity, the gas does not penetrate beyond the membranes lining the respiratory tract and lungs, so brain effects likely occur via generation of free radicals from lipid peroxidation.

The literature on effects of ozone exposure on aggressive behaviors is limited and inconsistent: While Musie et al reported that ozone exposure led to increased aggression, Petruzzi et al (NO PMID 1995) found that continuous 1200 ppb ozone for 20 days induced a decrease in time spent attacking, and an increase in freezing behavior in a paired aggression test.

Air pollutants, particularly ozone, are related to 8-Hydroxy-2’-Deoxyguanosine (8-OH-DG), a marker of oxidative DNA Damage (20980452) and to 8-isoprostane (8-ISO), a measure of lipid oxidation 18087591. In human experiments, 4 hours of 0.2 ppm ozone exposure led to a 79% increase in 8-ISO levels 18 hours after exposure, while after ambient ozone exposure, 8-OH-DG levels were elevated with a 7 to 15 day lag. Notably, both 8-OH-DG and 8-ISO levels are elevated in intermittent explosive disorder (but not axis I or axis II psychiatric disorders), and further correlated to measures of actual aggressive
3.1.2 Existing epidemiological evidence for an ozone-crime link

Previous studies have produced mixed results in accessing the relationship between ozone exposure and violent crime. A study by Rotton and Frey [90] found a significant association between ozone levels and calls for family disturbances but was conducted in 1975-1976 and utilized just two years of data from a single city (Dayton, Ohio). Changes in demographic variables as well as statistical methodology since that time are arguments against the contemporary generalizability of this study.

A more recent panel study performed by Lu et. al. [66] did not differentiate between individual air pollutants and standardized them into a "composite score", finding a significant association between this score and 5 categories of violent and non-violent crime. However, they based their effect estimates on annual measures of both pollution and crime (over 9 years from 9,360 US cities) which disregards the important seasonal trends in both measures and did not account for weather or climate variables in their model.

Finally, while both of Herrnstadt et. al.'s [45] studies on LA and Chicago found significant causal effects of being downwind to air pollution on daily violent crime rates, their instrumental variable (IV) methodology required daily comparisons of upwind regions to downwind regions within the cities which 1) did not separate or identify effects of individual pollutants, 2) did not quantify the magnitude of pollution exposure which differentiated “upwind” sites to “downwind” sites on any given day, 3) relied on unproven assumptions about the exchangeability of upwind and downwind sites with regards to socio-economic and demographic variables and 4) cannot easily be replicated on data from other locations where wind patterns are not dictated by unique geographical features (the LA basin, or Chicago highways which bisect communities of equal socio-economic value).

Due to the limitations of these previous studies, the association between ozone exposure and violent crime is still uncertain. This paper aims to fill this gap in literature with an analysis of daily measures of ozone concentration and rates of violent crime across six US cities from 2009 to 2018. This daily time-series data is able to capture and control for seasonal trends in both crime and ozone concentration. Two categories of analyses are performed: the first compares across “high” versus “low” ozone exposure as determined by NAAQ 8-hour exposure standards, and the second estimates the dose-response rela-
Table 3.2.1: Summary statistics for exposure, outcome and a subset of confounder variables provided for all six cities

<table>
<thead>
<tr>
<th>Type of variable</th>
<th>Exposure</th>
<th>Outcome</th>
<th>Confounders</th>
<th>Sensitivity analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max Ozone (continuous) in ppm (log)</td>
<td>Max Ozone (dichotomized)</td>
<td># aggravated assault reports</td>
<td>Avg PM$_{2.5}$ (μg/m$^3$)</td>
</tr>
<tr>
<td>Chicago</td>
<td>Min</td>
<td>0.13</td>
<td>1/1/01</td>
<td>1.70</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>3.37</td>
<td>0.015</td>
<td>44.07</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>3.23</td>
<td>41</td>
<td>11.04</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>9.27</td>
<td>131</td>
<td>12/30/17</td>
</tr>
<tr>
<td>NYC</td>
<td>Min</td>
<td>0.10</td>
<td>1/2/06</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>3.21</td>
<td>0.015</td>
<td>51.33</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>3</td>
<td>50</td>
<td>8.44</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>9.17</td>
<td>132</td>
<td>12/30/17</td>
</tr>
<tr>
<td>LA</td>
<td>Min</td>
<td>1.04</td>
<td>5</td>
<td>1/2/10</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>4.37</td>
<td>0.014</td>
<td>23.5</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>4.45</td>
<td>23</td>
<td>11.07</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>8.43</td>
<td>45</td>
<td>12/30/17</td>
</tr>
<tr>
<td>Philadelphia</td>
<td>Min</td>
<td>0.10</td>
<td>6</td>
<td>1/2/06</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>3.54</td>
<td>0.020</td>
<td>23.46</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>3.4</td>
<td>23</td>
<td>9.39</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>9.78</td>
<td>71</td>
<td>12/30/17</td>
</tr>
<tr>
<td>Phoenix</td>
<td>Min</td>
<td>0.77</td>
<td>2</td>
<td>11/1/15</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>4.57</td>
<td>0.014</td>
<td>12.52</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>4.69</td>
<td>12</td>
<td>7.09</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>7.81</td>
<td>29</td>
<td>12/30/17</td>
</tr>
<tr>
<td>Atlanta</td>
<td>Min</td>
<td>0.20</td>
<td>0</td>
<td>3/1/09</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>3.92</td>
<td>0.024</td>
<td>6.33</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>3.80</td>
<td>6</td>
<td>10.18</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>10</td>
<td>21</td>
<td>10/13/18</td>
</tr>
</tbody>
</table>

The relationship between ozone concentration and violent crime by keeping ozone a continuous variable. These two analyses are performed on an “all city” dataset which aggregates data from all six cities, as well as separate analyses for data within each city to control for time-invariant socio-economic and demographic differences between cities.

All analyses utilize Bayesian additive regression tree (BART) models in order to flexibly control for weather, seasonal and other daily time-varying covariates. BART is one of several powerful machine learning procedures emerging at the forefront of flexible statistical modelling and this paper is among the first to utilize this tool in an analysis of air pollution exposure effects.

### 3.2 Methodology

#### 3.2.1 Data collection

Daily counts for reported aggravated assault (AA) was chosen the outcome since reports were publicly available, and it was a category of violent crime a substantial daily count for all cities, and was the most reliably reported between cities. Police reports for AA were sourced from each individual city’s police department websites [1–6]. The range in
dates publicly available varied by city, with the Chicago providing the longest time series of crime data (5987 days, 2001 to 2017) and Phoenix providing the shortest (786 days, 2015 to 2017).

Daily maximum ozone concentration in \( ppm/m^3 \) and \( PM_{2.5} \) in \( \mu g/m^3 \) were sourced from air quality monitors within city limits of Chicago, NYC, LA, Philadelphia, Phoenix and Atlanta operated by the EPA [10]. Monitor readings taken every 8 hours were averaged for a daily reading. Daily UV index within each city was sourced from the National Weather Service Climate Prediction Center [11]. Precipitation (in.) and maximum daily temperature (Fahrenheit) was sourced from the National Oceanographic and Atmospheric Administration’s Climatology Network [8].

661 days with missing values for average ozone exposure, UV index or average \( PM_{2.5} \) were removed from the analysis, as the source of missingness was due to monitoring equipment failure and thus very unlikely to be associated with the outcome in any way (missing completely at random). Missingness was not associated with city identifiers or any other covariate. Sensitivity analyses (not included) performed with missing data procedures standard to the bartMachine R package [52] found no substantial difference in analysis results.

Summary statistics for covariates separated by city (Chicago, NYC, LA, Philadelphia, Phoenix and Atlanta) are provided in Table 3.2.1.

3.2.2 BART for causal inference

Bayesian additive regression trees (BART) fall under the umbrella of decision trees, machine learning procedures which construct prediction models by recursively partitioning the covariate space [24]. Due to its semi-parametric nature, BART may easily incorporate non-linearity and higher-order interactions into its predictions, and may be utilized to impute estimates of counterfactual outcomes in a classical causal estimation framework [39, 40, 46, 75].

3.2.3 Binary treatment

BART models an observed outcome \( (Y) \) as a function of covariates \( (X \in \mathbb{R}^p) \) and binary treatment \( (Z) \):
\[ Y = f(Z, X) + \epsilon \]  
\[ f(Z, X) = \sum_{t=1}^{T} \mathbb{T}_t(Z, X) \]  
\[ \epsilon \sim N(0, \sigma^2) \]

Where \( \mathbb{T}_t(X) \) is binary regression tree \( t \) out of \( T \) total trees and \( \epsilon \) is a normally-distributed error term. We make the standard causal assumptions of consistency \( (Y_\circ = Y|Z = z) \), positivity \( (0 < P(Z = 1) < 1 \text{ for all observations}) \) and conditional exchangeability \( (Y_\circ, Y, Z|X = x) \), and model the conditional expectations of the distribution of counterfactuals with:

\[ E(Y_1|X = x) = E(Y|Z = 1, X = x) = f(1, x) \]  
\[ E(Y_0|X = x) = E(Y|Z = 0, X = x) = f(0, x) \]

Individual-level effects may be estimated by taking the difference between the observed outcome under observation \( Z \) and its imputed predicted outcome under observation \( 1 - Z \). Let \( \Delta_i \) represent the treatment effect for individual \( i \):

\[ \Delta_i = \begin{cases} 
Y_i - f(0, X_i), & \text{if } Z_i = 1 \\
 f(1, X_i) - Y_i, & \text{if } Z_i = 0
\end{cases} \]

Average treatment effects may then be estimated by averaging over estimated individual causal effects \( \bar{\Delta} = \frac{\sum_{i=1}^{n} \Delta_i}{n} \).

Computationally, this amounts to training separate BART models on treated and control observations, including all relevant confounders in both BARTs. Individual predicted effects may be calculated by first imputing counterfactual outcomes of treated observations were they in the control group with the BART models trained on the control observations and visa versa. Averaging over individual-level differences produces a single estimate for the difference in outcome associated with high or low ozone exposure status.
BART estimates its posterior samples for its parameters through use of a Bayesian backfitting algorithm. A description of the algorithm may be found in Appendix ?? as well as several other papers [24, 52, 75]. One draw of a predicted outcome for each individual is calculated by summing over predicted outcomes from all trees. One advantage of BART over similar non-parametric machine learning procedures is its robust performance utilizing the default prior [24] (forestalling the need for computationally expensive cross-validation procedures), as well as easily accessible Bayesian posterior credible intervals for quantifying uncertainty.

3.2.4 Continuous treatment

Hill [46] provides an extension of BART for estimating the causal effect of a continuous treatment. Let $W$ represent a vector of continuous treatments. Instead of fitting separate BARTS on control and treated groups, a single BART is fitted with $W$ as a continuous predictor and used to predict the counterfactual $Y_{w+\delta}$, the outcome of each individual had they received treatment level $w + \delta$. Where $\delta$ is any “step size” of choice. Then,

$$\Delta_i = f(W_i + \delta, X_i) - Y_i$$  \hspace{1cm} (3.7)

And posterior inference may proceed as in the binary treatment setting.

3.2.5 Statistical analysis

In our statistical analysis, each day is treated as a single observation. The relationship between ozone exposure and AA reports within each city is examined first by dichotomizing ozone exposure into “high” and “low” at the NAAQS (0.07 ppm/m$^3$). Then we examine a dose-response relationship between continuous ozone exposure and daily counts of AA.

NAAQS dichotomized ozone exposure effect estimation

Days where maximum ozone concentration exceeded 0.07 ppm/m$^3$ were labeled “high exposure” days, while days which did not exceed this standard were labeled “low exposure days.”
For each city, estimates of the causal effect and posterior credible intervals were obtained from BART utilizing the procedure described in 3.2.3, controlling for date, average $PM_{2.5}$, precipitation, and maximum temperature and continuous variables, as well as indicators for whether the observation falls on a friday/saturday/sunday or a national holiday. Estimates of the causal effect and posterior credible intervals were also obtained for the all-city data using the same procedure and predictors, and adding indicator variables for each city.

For comparison, a linear regressions (OLS) were fitted with AA counts as the outcome and an indicator for high ozone exposure included as a predictor, controlling all the same confounders as included in the BART analyses for the all-city and six city-specific datasets. Reported 95% confidence intervals were obtained from the regression output.

### Continuous ozone exposure effect estimation

Causal effects and posterior predictive intervals describing the dose-response relationship of increasing ozone concentration by $1/100th$ of a ppm/m³ were obtained by the procedure described in Section 3.2.4, with $\delta = 1/100$. Comparative estimates of association from a linear regression were obtained by including continuous ozone concentration as a predictor in the model alongside previously discussed confounders. Confounders for BART and OLS were kept consistent to the analyses performed for dichotomized ozone exposure on the all-city and six city-specific datasets.

All statistical analyses were performed in R, utilizing the bartMachine package with 50 trees and 1000 iterations run. A brief description of the BART algorithm, as well as the priors utilized in this analysis may be found in Appendix ??.

## 3.3 Results

Table 3.3.1 contains BART and OLS estimates of the causal relationship between ozone concentration (dichotomized and continuous) and AA counts for the all-city data as well as for the six cities separately. Table 3.3.2 contains measures of fit (Pseudo/Adjusted $R^2$ and RMSE) for BART and OLS under all analyses discussed in this section. Fit measures obtained for BART are calculated with respect to the training data, and though they
Table 3.3.1: Effect estimates and 95% CI for dichotomized and continuous ozone exposure analyses

<table>
<thead>
<tr>
<th></th>
<th>Binary analysis</th>
<th>Continuous analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>BART</td>
</tr>
<tr>
<td>All cities</td>
<td>-0.32 [-1.36, 0.72]</td>
<td>7.53 [3.32, 11.63]</td>
</tr>
<tr>
<td>Chicago</td>
<td>3.86 [1.66, 6.05]</td>
<td>14.63 [3.72, 25.22]</td>
</tr>
<tr>
<td>NYC</td>
<td>-2.84 [-5.85, 0.17]</td>
<td>6.85 [-1.85, 15.41]</td>
</tr>
<tr>
<td>LA</td>
<td>0.72 [-1.08, 2.52]</td>
<td>3.81 [0.29, 7.30]</td>
</tr>
<tr>
<td>Philadelphia</td>
<td>-2.04 [-3.40, -0.68]</td>
<td>3.33 [-1.45, 8.14]</td>
</tr>
<tr>
<td>Phoenix</td>
<td>-2.16 [-4.40, 0.01]</td>
<td>-1.03 [-3.22, 0.97]</td>
</tr>
<tr>
<td>Atlanta</td>
<td>-0.17 [-0.78, 0.44]</td>
<td>1.46 [-0.17, 3.24]</td>
</tr>
</tbody>
</table>

Table 3.3.2: Measures of fit for BART and OLS: Adjusted/Pseudo $R^2$ and RMSE

<table>
<thead>
<tr>
<th></th>
<th>All cities</th>
<th>Chicago</th>
<th>NYC</th>
<th>LA</th>
<th>Philadelphia</th>
<th>Phoenix</th>
<th>Atlanta</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Binary analysis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>0.76</td>
<td>0.63</td>
<td>0.46</td>
<td>0.29</td>
<td>0.24</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>5.76</td>
<td>10.14</td>
<td>11.49</td>
<td>5.05</td>
<td>3.03</td>
<td>2.67</td>
<td></td>
</tr>
<tr>
<td>High exposure BART</td>
<td>Pseudo $R^2$</td>
<td>0.96</td>
<td>0.94</td>
<td>0.81</td>
<td>0.46</td>
<td>0.71</td>
<td>0.42</td>
</tr>
<tr>
<td>RMSE</td>
<td>5.38</td>
<td>5.30</td>
<td>7.11</td>
<td>4.07</td>
<td>3.04</td>
<td>2.08</td>
<td></td>
</tr>
<tr>
<td>Low exposure BART</td>
<td>Pseudo $R^2$</td>
<td>0.89</td>
<td>0.88</td>
<td>0.66</td>
<td>0.32</td>
<td>0.34</td>
<td>0.18</td>
</tr>
<tr>
<td>RMSE</td>
<td>5.77</td>
<td>7.66</td>
<td>8.50</td>
<td>5.18</td>
<td>3.65</td>
<td>2.39</td>
<td></td>
</tr>
<tr>
<td><strong>Continuous analysis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>0.76</td>
<td>0.64</td>
<td>0.48</td>
<td>0.29</td>
<td>0.23</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>7.74</td>
<td>10.11</td>
<td>11.38</td>
<td>5.04</td>
<td>3.04</td>
<td>2.67</td>
<td></td>
</tr>
<tr>
<td>BART</td>
<td>Pseudo $R^2$</td>
<td>0.80</td>
<td>0.78</td>
<td>0.55</td>
<td>0.24</td>
<td>0.44</td>
<td>0.18</td>
</tr>
<tr>
<td>RMSE</td>
<td>5.90</td>
<td>7.71</td>
<td>8.41</td>
<td>5.19</td>
<td>3.67</td>
<td>2.38</td>
<td>2.29</td>
</tr>
</tbody>
</table>
do not access the predictive capabilities of BART, they are presented as a direct comparison to OLS measures of fit. In all analyses, the BART model achieves a higher Pseudo-$R^2$ and lower RMSE than OLS, suggesting that the flexibility of BART is capturing non-linearity and/or higher-order interactions that OLS is not.

3.3.1 **Analyses with NAAQS-dichotomized ozone exposure**

In the analysis of data aggregated across all cities, BART estimates a causal effect of $7.53 ([3.32, 11.63])$ more AA reports on high ozone exposure days compared to low ozone exposure days after flexibly controlling for time-varying confounders. For all cities except Phoenix, an ozone concentration above NAAQS results in a higher number of reported AAs. Chicago exhibits the highest estimated effect of high ozone exposure on AAs, with high ozone exposure resulting in $14.63 ([3.72, 25.22])$ more reported assaults than low ozone exposure days. The credible intervals for this positive effect do not include zero for Chicago and LA analyses.

After dichotomization at NAAQS for ozone, all cities exhibited an extremely low proportion of days which were labeled “high exposed” (Table 3.2.1), with Atlanta exhibiting the highest proportion of “high exposed” days at $2.4\%$ (82 days) and LA and Phoenix exhibiting the lowest proportion of “high exposed” days at $1.4\%$ (39 days for LA and 11 days for Phoenix).

The small number of high exposed data points within each city explains the wide credible intervals around BART estimates for all cities and the especially small number of high exposed days for Phoenix helps explain why BART estimates for this city do not have the same direction as for other cities.

Unlike BART, OLS estimates a negative causal effect between high ozone exposure and number of reports for AA in most cities (NYC, Philadelphia, Phoenix and Atlanta) as well as in the aggregated data. Only in Chicago and Philadelphia are these differences significant. The contradiction between OLS and BART results in this analysis suggests that the high correlation between dichotomized ozone exposure status and weather variables may be nullifying and even reversing the direction of estimated effect.
Table 3.4.1: Effect estimates for continuous and dichotomous ozone exposure on non-violent crimes, larceny and auto theft, by city with bolded significant effects.

<table>
<thead>
<tr>
<th>City</th>
<th>Continuous analysis</th>
<th></th>
<th></th>
<th>Auto theft</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>BART</td>
<td></td>
<td>OLS</td>
<td>BART</td>
<td></td>
</tr>
<tr>
<td>Chicago</td>
<td>-0.96</td>
<td>-0.06</td>
<td></td>
<td>0.03</td>
<td>0.89</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>[-2.2, 1.3]</td>
<td>[-4.6, -10.1]</td>
<td></td>
<td>[-2.15, 2.22]</td>
<td>[-4.8, 10.8]</td>
<td></td>
</tr>
<tr>
<td>NYC</td>
<td>-0.66</td>
<td>-0.78</td>
<td></td>
<td>-0.14</td>
<td>-0.48</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>[-1.72, 0.39]</td>
<td>[-5.62, 4.79]</td>
<td></td>
<td>[-2.15, 1.93]</td>
<td>[-0.78, -0.19]</td>
<td></td>
</tr>
<tr>
<td>LA</td>
<td>-1.46</td>
<td>-2.98</td>
<td></td>
<td>-0.11</td>
<td>-0.38</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>[-3.56, 0.64]</td>
<td>[-1.69, 0.31]</td>
<td></td>
<td>[-2.15, 1.93]</td>
<td>[-0.31, 0.19]</td>
<td></td>
</tr>
<tr>
<td>Philadelphia</td>
<td>-4.65</td>
<td>-4.17</td>
<td></td>
<td>-0.14</td>
<td>-0.58</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>[-11.2, 2.0]</td>
<td>[-4.79, -3.5]</td>
<td></td>
<td>[-2.12, 1.51]</td>
<td>[-1.01, -0.12]</td>
<td></td>
</tr>
<tr>
<td>Phoenix</td>
<td>-0.99</td>
<td>-0.54</td>
<td></td>
<td>-0.41</td>
<td>-0.54</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>[-1.33, 1.32]</td>
<td>[-0.73, -0.72]</td>
<td></td>
<td>[-2.12, 1.51]</td>
<td>[-1.01, -0.12]</td>
<td></td>
</tr>
<tr>
<td>Atlanta</td>
<td>-0.35</td>
<td>-0.28</td>
<td></td>
<td>-0.68</td>
<td>-0.56</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>[-2.36, 2.0]</td>
<td>[-1.38, 2.4]</td>
<td></td>
<td>[-1.93, 1.09]</td>
<td>[-1.01, -0.12]</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.4.1: Histograms of daily counts of AA

3.3.2 Analyses with continuous ozone exposure

In analyzing the all-city dataset, BART estimates a 0.30 ([0.15, 0.44]) increase in AA reports in response to an increase of 1/100 of a ppm/m³ in maximum daily ozone exposure. For Chicago, BART estimates a positive causal relationship between ozone concentration and number of AA reports, with a posterior credible interval of effect that does not include 0. For all other cities, BART estimates a smaller (or, in the case of Phoenix, negative) effect of increasing ozone exposure on AAs with credible intervals which do include 0.

Compared to BART, OLS estimates larger effects in the aggregated all-city analysis as well as within all cities except Philadelphia and Atlanta. Confidence intervals estimated in all cities except Philadelphia and Atlanta do not include 0, indicating a significant positive causal effect. Estimation of larger effects via OLS may be occurring due to its parametric assumptions lending power to the estimate. However, since BART still displays superior fit metrics (Table 3.3.2) to OLS for every analysis, these parametric assumptions may not be holding.
Table 3.4.2: Rate ratio estimates from Poisson regressions

<table>
<thead>
<tr>
<th></th>
<th>Dicho ozone</th>
<th>Cont ozone</th>
</tr>
</thead>
<tbody>
<tr>
<td>All cities</td>
<td>0.98 [0.97, 1.00]</td>
<td>1.01 [1.01, 1.02]</td>
</tr>
<tr>
<td>Chicago</td>
<td>1.01 [0.98, 1.04]</td>
<td>1.01 [1.00, 1.01]</td>
</tr>
<tr>
<td>NYC</td>
<td>0.95 [0.91, 0.98]</td>
<td>1.01 [1.01, 1.01]</td>
</tr>
<tr>
<td>LA</td>
<td>1.02 [0.96, 1.09]</td>
<td>1.02 [1.01, 1.03]</td>
</tr>
<tr>
<td>Philadelphia</td>
<td>0.92 [0.88, 0.96]</td>
<td>1.00 [0.99, 1.01]</td>
</tr>
<tr>
<td>Phoenix</td>
<td>0.84 [0.70, 1.00]</td>
<td>1.04 [1.01, 1.07]</td>
</tr>
<tr>
<td>Atlanta</td>
<td>0.97 [0.89, 1.05]</td>
<td>1.01 [0.99, 1.02]</td>
</tr>
</tbody>
</table>

3.4 Discussion

This study analyzed the association between ozone exposure and reports of AA using time-series data from six cities across the US, flexibly controlling for non-linear confounding and higher-order interactions from daily weather confounders with BART. Our first all-city analysis presented evidence that ozone exposure above NAAQS result in high numbers of reported AA compared to ozone concentrations which stay below the standard. Our second all-city analysis suggests that a causal dose-response relationship exists between continuous ozone exposure and number of reported AAs. Evidence is weaker for both conclusions when looking at analyses performed with city-specific data, which is understandable since there is a loss of power when the sample size is divided. This is particularly evident for analyses with data from Phoenix, which has the lowest sample size of all cities, and Atlanta, which has the lowest average number of AA reports of all cities. It is promising, however, that all causal effects estimated by BART (except in Phoenix) suggests that ozone concentration has a positive relationship with violent crime.

Discrepancies exist between estimates of causal effect obtained with OLS and BART for both analyses. In the dichotomized ozone exposure analysis, OLS estimates an association in the opposite direction of BART, while in the continuous ozone exposure analysis, BART estimates lower measures of association overall, and for 3 of the 6 cities compared to OLS. These differences suggest that OLS may not be equipped to handle the high correlation between ozone concentration and time-varying weather confounders, or that non-linear relationships or higher-order interactions are not being accounted for. Thus the flexible
controlof confounding offered by BART makes a difference in producing reliable estimates of association.

In order to confirm that ozone exposure is affecting violent crime specifically not just crime in general, we repeated the statistical analysis outlined in Section 3.2.5 on counts of two categories of non-violent crime (larceny and auto theft). Table 3.4.1 shows there is no strong evidence of an effect between ozone exposure and the non-violent crimes when ozone is dichotomized, but there is evidence of a negative relationship between ozone exposure and larceny when ozone is kept continuous. This is interesting, as BART does not estimate a strong negative relationship between continuous ozone and auto theft. One explanation may be that higher levels of ozone exposure may cause non-violent crimes like larceny to intensify into more violent crimes like aggravated assault, thereby lowering counts of the former and raising counts of the latter. This echoes a previous finding by Rotton and Frey [90].

One limitation of this study is our treatment of aggravated assault reports as a continuous, rather than count variable as modifications to BART for estimating count variable outcomes [74] do not have readily-available R packages. Figure 3.4.1 contains histograms utilized to visually assess the normality of daily AA counts within each city. Chicago, NYC and Philidelphia all exhibit moderate left skew in number of reported AA, and data from all cities rejected the null of Shapiro-Wilk tests, indicating significant departure from normality. The impact of this on OLS association estimates was accessed by performing a sensitivity analysis using Poisson regressions (Table 3.4.2). Measures of association using Poisson regression exhibit similar patterns to that of OLS in direction and significance. One direction for future study is to re-perform this analysis with extensions of BART formulated to estimate count data [74].

This study presented evidence establishing the need for flexible modeling when assessing the relationship between ozone exposure and violent crime. Our estimates of causal effect contribute to growing evidence that exposure to high concentrations of ozone has a damaging impact not only on health health, but on psychology and behavior as well.
Appendices
.1 Model specification for examined implementations

.1.1 Stratification

Let \( v_i \in [1, ..., S] \) represent the strata which observation \( i \) is assigned to. We model the outcome with a linear regression:

\[
E(Y_i) = \beta_0 + \beta_1 T_i + \sum_{s=2}^{S} \left[ \beta_{s1} 1(v_i=s) + \beta_{s2} T_i 1(v_i=s) \right]
\]  

(8)

Let \( P_{st} \) represent the proportion of those in treatment group \( t \) who were assigned to stratum \( s \), thus \( P_{st} = \frac{\sum_{i=1}^{n} 1(v_i=s)1(T_i=t)}{\sum_{i=1}^{n} 1(T_i=t)} \). Our estimate of \( \Delta \) is then a linear combination of \( \beta_s \):

\[
\Delta = \frac{\sum_{i=1}^{n} Y_i T_i}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n} Y_i (1 - T_i)}{\sum_{i=1}^{n} (1 - T_i)} = \beta_1 + \sum_{s=2}^{S} \left[ (P_{st} - P_{so}) \beta_{s1} + P_{so} \beta_{s2} \right]
\]  

(9)

Where the effect difference in level \( s = 1 \) is represented by \( \beta_1 \). In Sections 2.4 and 1.6, quintile stratification is utilized, setting \( S = 5 \). When performing standard Frequentist PSA, the MLE estimate for \( \hat{\beta} \) may be substituted for \( \beta \) to calculate \( \Delta \). The asymptotic form for \( Var(\Delta) \) was derived by Lunceford and Davidian [67], which is functionally equivalent to performing a transformation of the variance-covariance matrix for \( \beta \). We calculated the latter when performing PSA in Section 2.4.

Since a parametric likelihood may be recovered from this specification of the outcome model, when performing BPSA \( E(\Delta | T, X, Y, v) \) and \( Var(\Delta | T, X, Y, v) \) are simply the conditional posterior mean and variance of \( \Delta \), respectively, as outlined in Section 1.4.1. In practice, this involves performing a Bayesian linear regression conditional on strata assignments and obtaining the posterior mean and variance of multiple draws. We utilized a flat prior on \( \beta \).

.1.2 DR

Doubly-robust effect estimation is performed as described in Funk et. al. [37]. Let \( \hat{Y}_i \) represent counterfactual outcomes for an all-treated sample, estimated with the following
model:

\[
E(Y_i|T_i = 1) = \beta_0 + \sum_{q=1}^{P} \beta_q X_{iq}
\] (10)

\(\hat{\beta}\) is estimated by fitting this regression on the subsample of treated observations, then \(\hat{Y}_i\) is predicted for all observations. With a similar process, \(\hat{Y}_o\) is also predicted. The following formula is then utilized to estimate \(E(\Delta^*(\nu))\), where \(\nu\) is defined in Equation 1.5:

\[
\left[\frac{1}{n} \sum_{i=1}^{n} \frac{T_i Y_i - \hat{Y}_i(T_i - \hat{\epsilon}_i)}{\hat{\epsilon}_i} - \left[\frac{1}{n} \sum_{i=1}^{n} \frac{(1-T_i) Y_i - \hat{Y}_o(T_i - \hat{\epsilon}_i)}{1 - \hat{\epsilon}_i}\right]\right] = \left[\frac{1}{n} \sum_{i=1}^{n} \nu_i (Y_i - \hat{Y}_i) + \hat{Y}_i\right] - \left[\frac{1}{n} \sum_{i=1}^{n} \nu_i (Y_i - \hat{Y}_o) + \hat{Y}_o\right]
\] (11)

While the following formula from Lunceford and Davidian \cite{67} is utilized to estimate \(\text{Var}(\Delta^*(\nu))\):

\[
\frac{1}{n} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{T_i Y_i - \hat{\mu}_{IPW,i}}{\hat{\epsilon}_i} - \left(1 - T_i\right)(Y_i - \hat{\mu}_{IPW,o})\right] - \left[\frac{1}{n} \sum_{i=1}^{n} \sqrt{\frac{\hat{\epsilon}_i}{1 - \hat{\epsilon}_i} (\hat{Y}_i - \hat{\mu}_{DR,i}) + \sqrt{\frac{\hat{\epsilon}_i}{1 - \hat{\epsilon}_i} (\hat{Y}_o - \hat{\mu}_{DR,o})}\right]
\] (13)

\[
\frac{1}{n} \left[\frac{1}{n} \sum_{i=1}^{n} \nu_i T_i (Y_i - \hat{\mu}_{IPW,i}) - \nu_i (1 - T_i)(Y_i - \hat{\mu}_{IPW,o})\right] - \left[\frac{1}{n} \sum_{i=1}^{n} \left(\nu_i - 1\right) T_i^{-\frac{1}{2}} (\hat{Y}_i - \hat{\mu}_{DR,i}) + \left(\nu_i - 1\right) T_i^{-\frac{1}{2}} (\hat{Y}_o - \hat{\mu}_{DR,o})\right]
\] (14)

Where \(\hat{\mu}_{IPW,i} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} T_i}\), \(\hat{\mu}_{IPW,o} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} (1-T_i)}\), \(\hat{\mu}_{DR,i} = \frac{\sum_{i=1}^{n} T_i Y_i}{\sum_{i=1}^{n} T_i}\), \(\hat{\mu}_{DR,o} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} (1-T_i)}\).

1.3 IPW

When utilizing IPW with BPSA, the conditional distribution of \(\Delta\) is estimated asymptotically.
The following formula is used to estimate $E(\hat{\Delta}^*(v))$, where $v_i$ is defined in Equation 1.5:

\[
\Delta = \frac{\sum_{i=1}^{n} \frac{T_i Y_i}{v_i}}{\sum_{i=1}^{n} \frac{Y_i}{v_i}} - \frac{\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{v_i}}{\sum_{i=1}^{n} \frac{1-T_i}{v_i}}
\]

\[
= \frac{\sum_{i=1}^{n} v_i T_i Y_i}{\sum_{i=1}^{n} v_i T_i} - \frac{\sum_{i=1}^{n} v_i (1 - T_i)Y_i}{\sum_{i=1}^{n} (1 - T_i)v_i}
\]

(15)

This is equivalent to the estimation of $\hat{\beta}_1$ in the following regression, when weights are set equal to $v_i$:

\[
E(Y_i) = \beta_0 + \hat{\beta}_1 T_i
\]

(17)

For PSA, $\text{Var}(\hat{\Delta})$ is estimated with the Hubert-White sandwich estimator \[54\], which is known to be conservative. This variance estimator leads to double-counting of design uncertainty when used to estimate $\hat{\text{V}}(v)$ in BPSA.

We instead estimate $\text{Var}(\Delta^*(v))$ with the following estimator from Lunceford and Davidian \[67\] which assumes the “true” propensity score (and thus true $v_i$) is known:

\[
\frac{1}{n} \sum_{i=1}^{n} v_i \left[ T_i (Y_i - \mu_{IPW,i})^2 + (1 - T_i)(Y_i - \mu_{IPW,0})^2 \right]
\]

(18)

1.4 NN AND CALIPER MATCHING

When utilizing NN or caliper matching with BPSA, the conditional distribution of $\Delta$ is estimated asymptotically.

Both nearest neighbor and caliper matching were performed with the MatchIt package on R. $v_i$, defined as the frequency weight outputted by the MatchIt object (see Section 1.3.2), is utilized in a weighted regression of the same model as Equation 17. Estimation of $E(\hat{\Delta}^*(v))$, $\text{Var}(\Delta^*(v))$ (as utilized in BPSA) and $\text{Var}(\hat{\Delta})$ (as utilized in PSA) proceed as described in Appendix 1.3 where frequency weights as described in the MatchIt package documentation \[49\] replace IP weights for $v$. 

81
.2 Observations highly weighted under IPW and OW

Closer examination of both OW and IPW reveals differences in what each weighting scheme prioritizes and which observations recieve are up-weighted and down-weighted under each procedure.

Figure ?? plots the log of IP-weights against overlap weights calculated on a single data set simulated as described in Section 2.4.1 under each of the 3 and 5 time points settings. For reference, observations are colored by the number of times they appear in DGCOP.

Both IPW and OW are alike in that they down-weight observations with a high probability of receiving the observed treatment pattern. An observation with a low IPW would never have a high OW. However, observations with a low OW and high IPW are common.

In order for an observation to have a high OW, it must consistently display a high to moderate probability of receiving the opposite treatment than observed at each time point. This is due to the restricted range of overlap weights, which severely penalizes observations which have a near-zero probability of observing the opposite treatment at any time point. As seen on Figure ??, observations with high overlap weights tend to exist in an area of covariate overlap in the PS distribution at most of the time points, though are not necessarily all members of the DGCOP.

IPW performs differently to OW in two key ways. First, an observation does not have to exhibit consistently high probability of receiving the opposite treatment to be assigned an extreme IPW. Rather, due to the lack of upper bound on IPWs, an observation may recieve such an extreme weight at one time point that its behavior at other time points has little effect on its final weight. Second, observations which recieve extreme weights under IPW are in fact often in areas of low covariate overlap. IPW highly up-weights observations which, at any time point, are one of a few representing its own treatment group in an area of the PS distribution overwhelmingly populated by members of the opposite treatment group. As evidenced in Figure ??, observations which recieve extreme IPWs are often not in the DGCOP.

.3 BART algorithm description

We initialize each regression tree with a single node at depth $d = 0$. This node proposes a split into two child nodes of depths $d + 1$, and this split is accepted or rejected via a
Metropolis-Hastings step \cite{52}. \(T_t\) has the branching process prior where each node at depth \(d\) is non-terminal with probability \(q(d) = \gamma(1 + d)^{-\beta}\).

Each parent node is assigned a splitting rule of the form \([x_p \leq c]\) where \(p \in [1...P]\), and observations are assigned to children of the parent node based on whether or not \(X_p\) satisfies the splitting rule. \(X_p\) is chosen according to the probability vector \(s = [s_1...s_P]\).

Given that predictor \(p\) is selected, \(c\) is drawn uniformly from the observed values \(x_{1p}...x_{np}\) which lead to nontrivial splitting rules. A splitting rule is considered trivial if it contradicts or is redundant with regards to a rule higher in the tree. If no \(X_p\) exists where a nontrivial split is possible, the node becomes terminal.

At each iteration, each tree may split (a terminal node becomes a parent node), prune (a parent node becomes a terminal node) or change (two parent nodes on the same tree swap splitting rules). New trees are sampled with a Metropolis-Hastings step detailed in Kapelner and Bleich \cite{52}.

Each terminal node \(l\) is associated with a mean parameter \(\mu_{t,l} \sim N(0, \sigma^2_{\mu})\). Normalization by the number of trees \(T\) ensures that \(f\) tends to a Gaussian process as \(T \to \infty\) under mild conditions \cite{61}. Let \(M_t = [\mu_{t,1}, ..., \mu_{t,L_t}]\) where \(L_t\) is the total number of leaves in tree \(t\).

We may define \(g(Z, X; T_t, M_t)\) as the \(\mu_{t,l}\) obtained by sending an observation with \([Z, X]\) down tree \(t\). Our predictions are created from the sum-of-trees model:

\[
\hat{Y} = \frac{\sum_{t=1}^{T} g(Z, X; T_t, M_t)}{T}
\]

(19)

\(T_t, M_t\) and \(\sigma\) are treated as parameters in a statistical model. Draws from the joint posterior distribution are taken with a Metropolis-within-Gibbs sampler, computed via Bayesian backfitting \cite{52}, where a “good fit” for the \(t\)th tree is sought iteratively by holding the other \(T - 1\) trees constant. At each iteration, trees may grow, shrink and change in response to the “fit” captured by other trees.
References


