DISSERTATION

METHODOLOGY IN AIR POLLUTION EPIDEMIOLOGY FOR LARGE-SCALE EXPOSURE PREDICTION AND ENVIRONMENTAL TRIALS WITH NON-COMPLIANCE

Submitted by

Nathan Ryder

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2023

Doctoral Committee:

Advisor: Kayleigh Keller

Ander Wilson
Daniel Cooley
Andreas Neophytou

ABSTRACT

METHODOLOGY IN AIR POLLUTION EPIDEMIOLOGY FOR LARGE-SCALE EXPOSURE PREDICTION AND ENVIRONMENTAL TRIALS WITH NON-COMPLIANCE

Exposure to airborne pollutants, both long- and short-term, can lead to harmful respiratory, cardiovascular, and cardiometabolic outcomes. Multiple challenges arise in the study of relationships between ambient air pollution and health outcomes. For example, in large observational cohort studies, individual measurements are not feasible so researchers use small sets of pollutant concentration measurements to predict subject-level exposures. As a second example, inconsistent compliance of subjects to their assigned treatments can affect results from randomized controlled trials of environmental interventions. In this dissertation, we present methods to address these challenges.

We develop a penalized regression model that can predict particulate matter exposures in space and time, including penalties to discourage overfitting and encourage smoothness in time. This model is more accurate than spatial-only and spatiotemporal universal kriging (UK) models when the exposures are missing in a regular (semi-daily) pattern. Our penalized regression model is also faster than both UK models, allowing the use of bootstrap methods to account for measurement error bias and monitor site selection in a two-stage health model.

We introduce methods to estimate causal effects in a longitudinal setting by latent "at-the-time" principal strata. We implement an array of linear mixed models on data subsets, each with weights derived from principal scores. In addition, we estimate the same stratified causal effects with a Bayesian mixture model. The weighted linear mixed models outperform the Bayesian mixture model and an existing single-measure principal scores method in all simulation scenarios, and are the only method to produce a significant estimate for a causal effect of treatment assignment by strata when applied to a Honduran cookstove intervention study.

Finally, we extend the "at-the-time" longitudinal principal stratification framework to a setting where continous exposure measurements are the post-treatment variable by which the latent strata are defined. We categorize the continuous exposures to a binary variable in order to use our previous method of weighted linear mixed models. We also extend an existing Bayesian approach to the longitudinal setting, which does not require categorization of the exposures. The previous weighted linear mixed model and single-measure principal scores methods are negatively biased when applied to simulated samples, while the Bayesian approach produces the lowest RMSE and bias near zero. The Bayesian approach, when applied to the same Honduran cookstove intervention study as before, does not find a significant estimate for the causal effect of treatment assignment by strata.

ACKNOWLEDGEMENTS

Many thanks to my advisor, Kayleigh Keller, who has patiently provided guidance, encouragement, and (importantly) deadlines these last few years. To my wife Joy, the person who bore the brunt of our family's responsibilities when I could not, and who has been beside me for all of the ups and all of the downs. To family for not letting me be satisfied with until I know why, and to friends who made these years an adventure, not an ordeal.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Ambient air pollution, both outdoor and indoor, is a global health burden, leading to loss of life and health through respiratory, cardiovascular, and cardiometabolic pathways (Chuang et al., 2011; He et al., 2022; Mann et al., 2021; Murray et al., 2020; U.S. Environmental Protection Agency, 2019). Ideally, to analyze the causal effects of air pollution on health researchers would perform a randomized controlled trial (RCT), but there are clear ethical concerns with randomly assigning subjects to different exposure levels and measuring health outcomes. Even beyond the ethical concerns, such a study would need to be prohibitively large in sample size and in length of time in order to capture the chronic and/or rare effects of exposure to pollutant concentrations over time. Instead, a researcher in air pollution epidemiology may use a large-scale (national or larger) observational set of PM measurements and healthcare data. Or they may perform an RCT in a small sample and in a particular context for an intervention to improve household air quality, but subjects may all be assigned the same treatment by the end of the study, and some do not comply to their assigned treatment. Analysis of these examples requires the researcher to address measurement error and non-compliance.

1.1 Exposure Prediction for Large Cohort Studies

To investigate the relationships between ambient air pollution and health outcomes, we must measure and often also estimate subjects' exposures to pollutant concentrations at personal or aggregate levels. Consider the example of a large spatial and temporal set of exposure and health measurements, where subjects' exposure levels are not measured directly, but must be predicted from observed concentrations at monitoring sites in the region. We would then use the predicted subject-level exposures in an analysis associating subject exposures and health outcomes, the second stage of an overall two-stage method for investigating the relationships of ambient air pollution exposure and health. Prediction models for the exposure may incorporate spatial variables, satellite

data, and both spatial and temporal correlation structures. Uncertainty due to measurement error from the monitoring devices and monitoring site selection in the exposure prediction model induces a complex form of measurement error for the overall two-stage health effects model. Szpiro and Paciorek (2013) address this issue and correct for measurement error with a design-based nonparametric bootstrap. Predicting in space and time can lead to a complicated model, which is likely to have a high cost in computation time that worsens with the size of the dataset. High computational cost in the exposure prediction model can render bootstrap measurement error correction methods infeasible due to limited time and resources. Thus, a fast enough exposure prediction model for bootstrap measurement error correction may be a better choice for a two-stage model than a slower and possibly more accurate prediction model. In Chapter 2, we develop a method to quickly predict subject-level exposure to PM in space and time at a large scale for application to epidemiological two-stage health effects studies. The contents of Chapter 2 are published in the Journal of Agricultural, Biological, and Environmental Statistics (JABES) with the title *Spatiotemporal Exposure Prediction with Penalized Regression* (Ryder and Keller, 2023).

1.2 Causal Inference in the Presence of Non-Compliance

The large cohort analyses mentioned above are typically observational, where no subjects are randomly assigned their exposure level, and it is difficult or impossible to interpret causality. Causal effect estimates, rather than measures of just the overall exposure-response relationship, are important for use in policy decisions. When an unmeasured shared cause of both exposure and outcome is present, interventions or regulations to lower the exposure (which does not affect the outcome) can be a waste of resources or even harmful. Take the toy example of ice cream sales and shark attacks, which have strong positive correlation but no causal relationship. Regulating ice cream sales would be ineffective to lower shark attack rates, but would hurt the businesses affected. For the case at hand, the causal effect of particulate matter exposure on health outcomes is crucial knowledge for setting air quality standards which not only protect health in the general population but also do not needlessly hamper the industries that the standards restrict.

To achieve inference that is "causal" we rely on carefully made assumptions and often employ techniques to parse total effects into more specific components, such as direct and indirect effects or the effects partitioned by principal strata (Frangakis and Rubin, 2002; Robins and Greenland, 1992). In Chapters 3 and 4, we deal with a longitudinal RCT for the intervention of a ventilated "improved" cookstove in rural Honduran households (Young et al., 2019). Subjects in the trial did not always comply with their assigned treatment, some using their original stove after being given the improved one and others using the improved stove before they were assigned it. This non-compliance obfuscates the true efficacy of the treatment in a typical intent-to-treat analysis of the study data. We use a principal stratification framework, and the assumptions it requires, to estimate the effect of treatment specifically on the subjects who would comply with their assigned treatment. In Chapter 3, we apply a principal stratification framework to perform causal analysis of an intervention in a longitudinal randomized controlled trial. In Chapter 4 we extend the longitudinal principal stratification framework to use a continuous post-treatment variable to define the latent strata.

Chapter 2

Spatiotemporal Exposure Prediction with Penalized Regression

2.1 Introduction

Long- and short-term exposures to total particulate matter (PM) are causally related to adverse respiratory and cardiovascular health outcomes (U.S. Environmental Protection Agency, 2019), and PM contributes to 4.7% of disability-adjusted-life-years (DALYs) in all ages (95% uncertainty interval 3.8 to 5.5) (Murray et al., 2020). PM is categorized by size, typically into ranges 10 μm or 2.5 μm and smaller, denoted as PM_{10} and $PM_{2.5}$ respectively. PM is itself a mixture of many different components, such as nitrates, sulfates, organic matter, metals, and soil dust. The health effects of each component can vary, with sulfate (SO_4^{2-}) identified as one that is associated with respiratory and cardiovascular health effects (U.S. Environmental Protection Agency, 2019).

Epidemiological studies investigating relationships between air pollution and adverse health outcomes rely on the prediction of ambient PM concentrations for subjects across space and/or time. PM is subject to widespread regulatory monitoring in many countries and these measurements can be used to develop prediction models. In the United States (US), regulatory monitors are often placed preferentially in urban centers or near known sources and vary in both method and frequency of measurements. PM and its components, which are measured at a subset of monitors, are subject to seasonal trends that can lead to large and highly variable measurements in some regions of the US, making accurate predictions of ambient concentrations challenging. Other characteristics that can make predicting concentrations difficult include: differences in instrument tolerances or protocols, extreme events such as wildfires or dust storms, and the overall size of the dataset. Thus, predicting a spatiotemporal exposure surface requires efficient use of both spatial and temporal information and can benefit from computationally efficient methods. The structure

of monitoring data for specific components is similar to total PM, although components are often measured with more sparsity in space and time.

There are a variety of spatial or spatiotemporal models that may be used to predict ambient pollutant concentrations. These include land-use regression (Beelen et al., 2013; Hoek et al., 2008), universal kriging (Sampson et al., 2013; Xu et al., 2019), penalized regression (Bergen and Szpiro, 2015; Keet et al., 2018; Paciorek et al., 2009), Gaussian Processes (GP) (Datta et al., 2016; Pati et al., 2011), Gaussian Markov Random Fields with Stochastic Partial Differential Equations (INLA-SPDE) (Cameletti et al., 2013), quantile methods (Reich et al., 2011), spectral approaches (Reich et al., 2014), convolutional neural networks (Di et al., 2016), and models using deterministic atmospheric chemistry simulation output (Berrocal et al., 2010, 2012; Wang et al., 2016) and satellite measurements (Berrocal et al., 2020; Young et al., 2016). Different modeling approaches can also be combined into ensemble models (Di et al., 2019, 2020).

In a review of spatiotemporal exposure modeling approaches, Berrocal et al. (2020) compared the performance of common methods using daily $PM_{2.5}$ measurements across the US. They found that universal kriging (UK), when fit separately to each day of data, outperformed all other tested models, including neural networks, random forests, and inverse distance weighting. The density of $PM_{2.5}$ monitoring is sufficient that a large number of observations are available each day, allowing the empirical Best Linear Unbiased Predictor from a kriging model to perform well (Schabenberger and Gotway, 2004). However, fitting UK separately on each day of measurements ignores temporal information, which provides an opportunity for improvement. A spatiotemporal model developed by Lindström et al. (2014) extends universal kriging by combining a smooth temporal trend with spatially varying coefficients. This more flexible and complex model is implemented in the `SpatioTemporal` package in `R`.

One weakness of more complicated models is poor scaling in computation time. For both UK and the `SpatioTemporal` (ST) model, the computational burden lies in an iterative optimization technique to estimate covariance parameters (e.g. sill or range). High computation time can prohibit the use of bootstrap-based predictive variance estimates or measurement error corrections.

These bootstrap approaches can capture uncertainty from monitoring site selection in addition to the uncertainty from parameter estimation that is found in prediction variances (Bergen et al., 2016; Keller and Peng, 2019; Szpiro and Paciorek, 2013). In this paper we propose a penalized regression model that penalizes overfitting and smoothes over predictions at adjacent timepoints. The proposed model is computationally fast and thereby feasible for bootstrapping while also providing accurate predictions over time and space.

Our motivation for fitting a spatiotemporal surface comes from ambient air pollution exposures, so we approach model description and performance with respect to this application, but the method could be applied to other contexts. In Section 2.2, we introduce our model and its methods of fitting. In Section 2.3, we evaluate the method in a set of simulations under a variety of conditions and compare against the predictive accuracy of UK and ST. In Section 2.4, we apply the method to daily measurements of total $PM_{2.5}$, total PM_{10}, sulfate, and silicon concentrations from 2017 and in Section 2.5 we provide a discussion.

2.2 Model

We propose a penalized regression model where, in addition to the typical overfitting penalty, there is a penalty that smoothes over adjacent timepoints. By smoothing temporally, our method can take advantage of data from the previous and following days where spatial-only methods cannot. Furthermore, unlike the `SpatioTemporal` model, this penalization approach does not require the assumption of a specific smooth time trend or temporal covariance function. With our approach we aim to match or improve on the predictive accuracy of other common methods while being computationally faster.

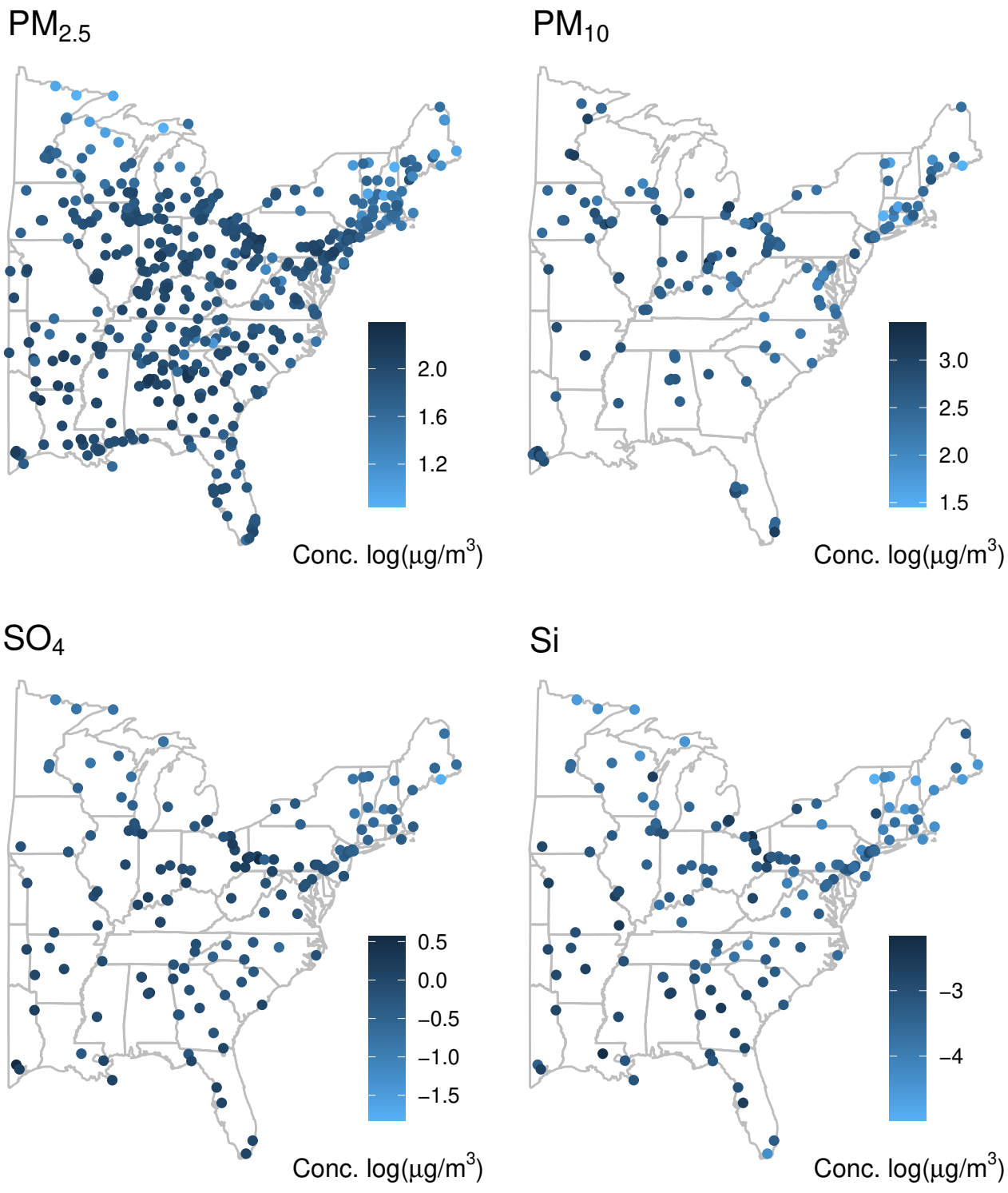

Figure 2.1: Average logged concentrations in 2017 of $PM_{2.5}$, PM_{10}, sulfate, and silicon at monitoring site locations in the Eastern United States.

2.2.1 Penalized Regression Model

The objective function for our model is:

$$\underset{\boldsymbol{\beta}_t \in \mathbb{R}^p \text{ for } 1 \le t \le T}{\text{minimize}} \sum_{i=1}^{n} \sum_{t=1}^{T} I_{it} \left(x_{it} - \boldsymbol{r}_{it}^{\top} \boldsymbol{\beta}_t \right)^2 + \sum_{t=1}^{T} g_1(\lambda_1, \boldsymbol{\beta}_t) + \lambda_2 \sum_{i=1}^{n} \sum_{t=2}^{T} g_2(\boldsymbol{r}_{it}^{\top} \boldsymbol{\beta}_t,\ \boldsymbol{r}_{i(t-1)}^{\top} \boldsymbol{\beta}_{t-1}). \tag{2.1}$$

The first term in (2.1) is quadratic loss and uses the indicator value I_{it} to only include time points (t) and locations (i) where x_{it} exists, e.g. where an ambient air pollutant concentration was measured. There are n total unique sites (locations) and T total dates (time points) where we could have observed x_{it}. The vector $\boldsymbol{r}_{it}$ contains p spatiotemporal covariates (Section 2.2.2) for date t and site i, including an intercept, while $\boldsymbol{\beta}_t$ is a p-vector of model coefficients for each date. We center and scale the covariates in $\mathbf{r}_{it}$ by time point. The second term, g_1, discourages overfitting, and may be an L_2, L_1, or Elastic Net penalty. We use an L_2 penalty here, i.e. $g_1(\lambda_1, \boldsymbol{\beta}_t) = \boldsymbol{\beta}_t^{\top} \boldsymbol{\Gamma}_1 \boldsymbol{\beta}_t$, where $\boldsymbol{\Gamma}_1$ is a $p \times p$ diagonal matrix comprised of only the values λ_1 and 0. This matrix allows penalization of all or specific model predictors using the value λ_1, and always excludes the intercept from penalization. See Section B.1 in the Appendix for specific construction.

The third term in (2.1) is $g_2(\boldsymbol{r}_{it}^{\top} \boldsymbol{\beta}_t,\ \boldsymbol{r}_{i(t-1)}^{\top} \boldsymbol{\beta}_{t-1}) = (\boldsymbol{r}_{it}^{\top} \boldsymbol{\beta}_t - \boldsymbol{r}_{i(t-1)}^{\top} \boldsymbol{\beta}_{t-1})^2$. This term smoothes over predictions that are adjacent in time by penalizing their differences and makes our model similar to trend-fitting or fused lasso (Petersen and Witten, 2019). Every prediction is smoothed to its immediate temporal neighbors, even if the observation at that time and place is missing. When $\lambda_2 > 0$, our model must predict using information across consecutive time points, but without the complication of spatiotemporal interaction and increased parameterization.

To make a spatiotemporal exposure surface, we must be able to predict exposure levels at any location and on any day in the spatiotemporal domain. The vector $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^{\top}, \widehat{\boldsymbol{\beta}}_2^{\top}, \ldots, \widehat{\boldsymbol{\beta}}_T^{\top})^{\top}$ that minimizes the objective function (2.1) includes p model coefficients for every date in the temporal domain. We can then predict an exposure level at any date and at any location for which we have the same covariate information used to fit the model, and can aggregate the resulting predictions to any desired spatial unit.

2.2.2 Spatial and Temporal Covariates

All spatial information for the model is provided through a set of spatial predictors included in $\mathbf{r}_{it}$. For the sake of flexibility and simplicity, we use Thin Plate Regression Splines (TPRS) (Wood, 2003). By fitting TPRS to site locations, we have a set of predictors that can vary in quantity through specification of degrees of freedom, i.e. the number of basis functions produced, and do not require manual tuning of knot placement. TPRS are also scaleable to a large number of site locations and provide covariate values at any location we want to predict. With this predictive flexibility comes the cost of the extra parameterization and computation time of adding a possibly large set of predictors to the model. We can also provide spatiotemporal covariates to the model. For our analysis of ambient pollutant concentrations in Section 2.4, we use just atmospheric chemical model output and meteorological data, but other variables such as land use measures could be included if desired.

2.2.3 Parameter Estimation

We rewrite the objective function (2.1) in matrix notation for simplicity:

$$\min_{\boldsymbol{\beta}} \left[(\boldsymbol{x} - \boldsymbol{R}_{obs}\boldsymbol{\beta})^\top(\boldsymbol{x} - \boldsymbol{R}_{obs}\boldsymbol{\beta}) + \boldsymbol{\beta}^\top\boldsymbol{\Lambda}_1\boldsymbol{\beta} + \lambda_2\boldsymbol{\beta}^\top\boldsymbol{R}^\top\boldsymbol{D}^\top\boldsymbol{D}\boldsymbol{R}\boldsymbol{\beta}\right]. \tag{2.2}$$

The first sum in Equation (2.2) is standard least squares loss which we write as the inner product of the difference in predicted and observed exposure vectors. To do this, we combine the covariate vectors ($\mathbf{r}_{it}$) for every observed time point at a site i in a row-wise manner to create a block-diagonal matrix. The n resulting block-diagonal matrices are stacked to form the matrix $\mathbf{R}_{obs}$ (see Section B.1 in the Appendix). Then the vector $\boldsymbol{R}_{obs}\boldsymbol{\beta}$ is the set of fitted values. The second sum in (2.2) is an L_2 penalty on the full coefficient vector $\boldsymbol{\beta}$. We create a set of T $\boldsymbol{\Gamma}_1$ matrices (as in Equation (2.1)) and stack them to form the block diagonal matrix $\boldsymbol{\Lambda}_1$.

To allow the temporal smoothing of predictions at unobserved date and site combinations, a "full" $\boldsymbol{R}$ matrix can be made by including covariate values for all site and date combinations,

instead of for just those that were observed. The third sum in (2.2) is then written as the inner product of the vector $\boldsymbol{DR\beta}$, where $\boldsymbol{D}$ is a diagonal block matrix and each of its n blocks is a first-order difference matrix. Every row in a block of the $\boldsymbol{D}$ matrix contains just zeroes and the pair of values -1 or 1 where the vector $\boldsymbol{R\beta}$ contains two predictions "adjacent" in time. The specific arrangements of all objects listed here are provided in Section B.1 of the Appendix.

When using an L_2 penalty for the overfitting term in the objective from Equation (2.2), a closed form solution for $\boldsymbol{\beta}$ exists. Since the matrix $\boldsymbol{R}_{obs}^\top \boldsymbol{R}_{obs} + \boldsymbol{\Lambda}_1 + \lambda_2 \boldsymbol{R}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{R}$ is positive definite for sufficiently large λ_1 and λ_2, we can find its inverse to produce:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{R}_{obs}^\top \boldsymbol{R}_{obs} + \boldsymbol{\Lambda}_1 + \lambda_2 \boldsymbol{R}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{R}\right)^{-1} \boldsymbol{R}_{obs}^\top \boldsymbol{x}. \tag{2.3}$$

The closed form solution is fast to compute when taking advantage of the sparseness of the matrices $\boldsymbol{R}$, $\boldsymbol{R}_{obs}$, and $\boldsymbol{D}$. If instead of the L_2 penalty, we use a non-convex overfitting penalty, or if we were to add further penalty terms, the optimization problem may no longer have a closed form solution. A general optimization strategy, such as Alternating Direction Method of Multipliers (ADMM) (Boyd et al., 2010) is then required, which increases computational cost.

2.2.4 Selection of Penalty Values

We select the penalty parameters λ_1 and λ_2 using 10-fold cross validation (CV), which provides an estimate of out-of-sample Root Mean Squared Error (RMSE). When cross-validating to select λ_1 and λ_2, we can use a coarse grid, e.g. every combination of 5 values for each penalty a factor of 10^2 apart. If a basin in the cross-validated RMSE values is identified, we can resume the search on a finer grid, e.g. combinations of values a factor of 10 apart, and repeat until we converge upon an approximately optimal fit. Using all combinations of values for the two penalties is comprehensive but slow. The alternative is to select λ_1 first, by setting $\lambda_2 = 0$ and choosing the λ_1 value that provides the lowest cross-validated RMSE. Then we repeat the process for values of λ_2 while setting λ_1 to the previously selected value. This sequential method of selecting the two

penalties sacrifices some predictive accuracy if the chosen values differ from what is selected by jointly searching over every combination of values, but it requires considerably fewer model fits. In Section 2.4 we find that the two methods select the same set of penalty values when modeling ambient $PM_{2.5}$ concentrations.

2.3 Simulation

2.3.1 Setup

For a variety of simulated data conditions, we compare our method to universal kriging and the `SpatioTemporal` model. Since the motivating application for our method is PM and its components, we create data sets with some characteristics matching the concentrations modeled in Section 2.4. Specifically, the simulated data have periodic (every three or six days) measurement by some proportion of "monitoring sites" and are correlated with a spatiotemporal observed covariate.

We simulate exposure data spatially over a [0,1] x [0,1] square using a 64 x 64 grid of points and temporally over a set of 60 time points. The model for simulated mean exposures at any grid points $\boldsymbol{s}$ and time point t is

$$\boldsymbol{\mu}(\boldsymbol{s},t) = \boldsymbol{\mu}_0(\boldsymbol{s},t) + \boldsymbol{Z}_1(\boldsymbol{s},t) + \boldsymbol{Z}_2(\boldsymbol{s},t), \tag{2.4}$$

where $\boldsymbol{\mu}_0(\boldsymbol{s},t)$ is a baseline spatiotemporal surface that is considered an observed covariate, and is unchanged for every simulated sample. All three spatiotemporal surfaces $\boldsymbol{\mu}_0(\boldsymbol{s},t)$, $\boldsymbol{Z}_1(\boldsymbol{s},t)$, and $\boldsymbol{Z}_2(\boldsymbol{s},t)$ are Gaussian processes generated with a Gneiting-style non-separable spatiotemporal covariance function (Gneiting, 2002). We use the `R` package `RandomFields` to simulate the processes using the specific covariance structure

$$C(h,u) = \frac{1}{(\psi(u)+1)} \ \phi\left(\frac{h}{(\psi(u)+1)^{1/2}}\right), \tag{2.5}$$

for spatial distance h and change in time u (Schlather et al., 2015). We use an Exponential covariance function for $\phi()$, with a range and sill both equal to 1 (i.e. $\phi(w) = e^{-w}$). For $\psi()$ we use fractional Brownian motion, a generalized random walk that depends on a Hurst index (H). Setting $1/2 < \text{H} < 1$ produces walks with positive correlation between increments, while for $0 < \text{H} < 1/2$ increments are negatively correlated and the trend will alternate more often in shorter time spans. We generate $\boldsymbol{\mu}_0(\boldsymbol{s}, t)$ using H = 0.25, allowing for some day-to-day fluctuation, and fix H = 0.95 for $\boldsymbol{Z}_1(\boldsymbol{s}, t)$, resulting in a more consistent long-term trend. The Hurst index for $\boldsymbol{Z}_2(\boldsymbol{s}, t)$ we vary across simulations to be H = 0.5 (standard Brownian motion with independent increments) or H = 0.05 (high daily fluctuation).

From the 4,096 grid points, we randomly select 500 training and 1000 testing locations as "monitors" and assign them simulated mean values for each of the 60 time points. To represent measurement error in the training set, we add the error term $\boldsymbol{\epsilon}_t \sim N(0, \boldsymbol{I}\sigma^2)$ to each set of "monitor locations" for each time point. Thus the training data follow the model $\boldsymbol{Y}(\boldsymbol{s}, t) = \boldsymbol{\mu}(\boldsymbol{s}, t) + \boldsymbol{\epsilon}_t$. To evaluate predictive accuracy, we compare each model's predictions for the testing data with the "true" simulated mean values $\boldsymbol{\mu}(\boldsymbol{s}, t)$ at the testing data time-locations.

We simulate while adjusting three settings: the error standard deviation (σ = 0.5 or 1.5), the temporal relation of $\boldsymbol{Z}_2(\boldsymbol{s}, t)$ (H = 0.5 or 0.05), and the amount of temporal missingness in the training data. As previously mentioned, some AQS monitors record values every third or every sixth day. These monitors generally follow synchronized schedules, so that each monitor observed every third day will record a measurements on the same schedule, leaving periods when only monitors on a daily schedule are observed. We then control the daily missingness in our training data by the proportion of monitoring locations that are observed daily, every third, or every sixth day, each matching the schedule of others in its scheme. For the sake of comparison, we perform another simulation study using the same proportions of sites by observation frequency, but stagger the observation schedules so that on any given day there may be sites that record daily, every third day, and every sixth day. In this alternative set of simulations, we repeat all settings adjustments as in the original, so that the only difference is that the non-daily schedules of monitoring sites

are evenly distributed across their possible starting dates. We report the results of this secondary simulation study in Figure B.1 of the Supplementary Material in the Appendix.

Under each distinct set of data conditions, we take 100 different seeded samples of training and testing data. The testing data are predicted from the training data using our model from Equation (2.1), UK, and ST. Each model uses the spatiotemporal covariate $\boldsymbol{\mu}_0(\boldsymbol{s}, t)$, unchanged from sample to sample, as a predictor. To fit our penalized regression model, the number of TPRS basis functions used as additional spatial covariates are selected via 10-fold cross-validation from the possible values 5, 10, 20, 50, 100, or 175. The penalities λ_1 and λ_2 are also chosen via 10-fold CV from the sets of possible values 0.1, 1, 10, 50, 100, 200, 300, 400, 500, and 0.001, 0.01, 0.1, 1, 10, 50, 100 respectively. Here we select λ_1 first before selecting λ_2, as mentioned in Section 2.2.4. We use UK per Berrocal et al. (2020), with a median set of exponential covariance parameter values from maximum likelihood fits at each time point. The `SpatioTemporal` model we fit with a single basis function and an exponential covariance structure with nugget for both the β-fields and the residual process ν (Lindström et al., 2014).

2.3.2 Results

We report the Root Mean Square Error (RMSE) values for each set of 100 simulated samples and fits as boxplots in Figure 2.2. There are results for the three tested models: penalized regression, universal kriging, and `SpatioTemporal`, as well as for penalized regression with only the L_2 penalty ($\lambda_2 = 0$), which is exactly ridge regression. We provide exact median RMSE values and squared correlations between predictions and observations (R^2) in Table B.1 in the Appendix.

Our model provides lower or matching RMSE values to universal kriging (UK) and ridge regression (denoted by $\lambda_2 = 0$) in every scenario, demonstrating good general predictive accuracy for a spatiotemporal model and the usefulness of the time-smoothing penalty λ_2. The `SpatioTemporal` (ST) model produces the lowest RMSE values of all models in each of the data scenarios except those when all locations are only observed every third or every sixth day,

where instead our penalized regression model outperforms all others. When none of the monitors are observed daily, our penalized regression model will disregard the completely missing dates and fit every third day as if it were daily (see Section 2.5 for discussion on interpolation in this scenario). Here we see that despite requiring greater than forty times the computation time (see Section 2.3.3), the ST model is more affected than our penalized smoother when there are time points without any observations. If monitors on an every third day and every sixth day schedule are measured in a staggered fashion, i.e. each day had some monitors of each frequency (daily, every third, and every sixth) and every date has a similar number of observations, then we do not see our model gain a predictive advantage over ST (see Supplementary Material Figure B.1 for the "staggered" simulation results). So it is when dates are completely unobserved that our model is able to outperform ST.

The results in Figure 2.2 highlight additional trends across different simulation settings. As non-spatial error (σ) increases, the fit of each model worsens. Similarly, more fluctuation from day-to-day (H = 0.05) reduces predictive accuracy for every model, and it should be noted that the ridge regression fits ($\lambda_2 = 0$) lose more accuracy than our model with its time-smoothing penalty ($\lambda_2 > 0$). UK and both penalized regression models have larger error when monitors are observed in all three frequencies (daily, every third day, and every sixth day) than if every monitor is observed only every third or sixth day. Since UK is a spatial-only model and fit onto the data from each day separately, having some days with only a few measured locations can pose a serious issue for prediction.

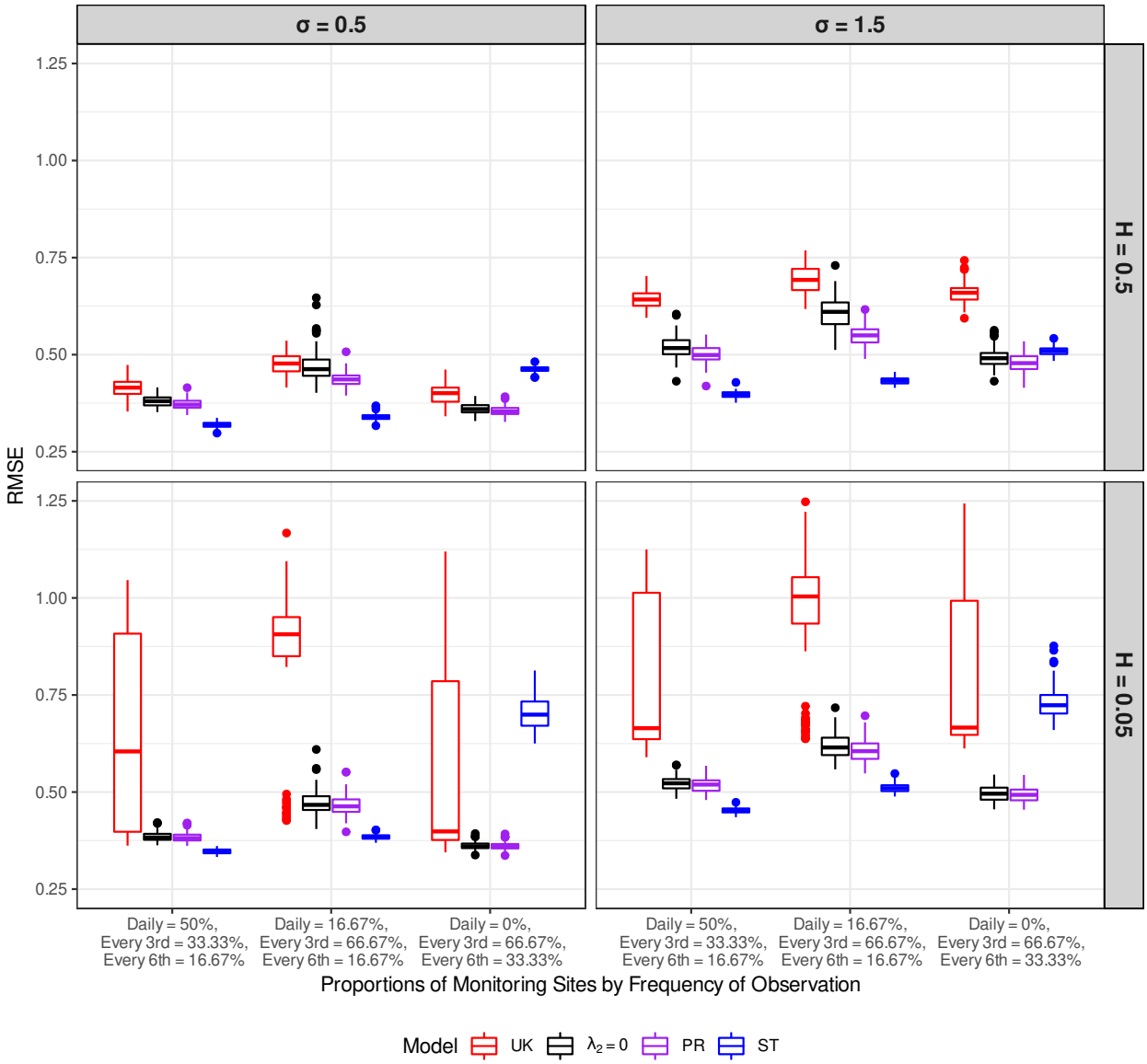

Figure 2.2: Boxplots of RMSE values from each model on 100 replicate samples for each simulation scenario. In order, the boxplots correspond to universal kriging (UK), our penalized regression model without its temporal smoothing penalty ($\lambda_2 = 0$), our model with the penalty (PR), and the `SpatioTemporal` (ST) model. Note that H is the Hurst index for $Z_2(s,t)$, which results in a more variable temporal trend as H approaches zero, and that σ is the standard deviation of non-spatial error added to the training data. The selected monitoring locations are observed daily, every third day, or every sixth day, according to the proportions listed on the x-axis.

2.3.3 Computation Times

In Equation (2.3), we see that a sparse pT x pT matrix must be inverted to estimate $\boldsymbol{\beta}$. To use UK, we must first invert an nT by nT matrix (or T n x n matrices) to estimate covariance parameters. Thus, our model scales better in computation time than UK with increasing site locations n. Figure 2.3 shows median computation times for sets of 10 replicate samples of each specified size ($N_{train} = 500$, 1000, 1500, or 2000) and compares our penalized regression model with UK. These values are also reported in Table B.2 in the Supplementary Material. Each of the fits were computed on the RMACC Summit Supercomputer, with a Intel Xeon E5-2680 v3 processor at 2.50GHz, using a memory cap of 50 GB of RAM. We see in Figure 2.3 that the need for cross-validating over penalty values and amounts of TPRS basis functions slows our model down to a speed similar to UK for some smaller sample sizes. However, as N_{train} increases, our method is faster. We include the median computation times for the one-time fit of sample training data that occurs after selecting penalty values and the number of TPRS basis functions to include as predictors. When using non-parametric bootstrap to correct measurement error, we would reuse the same model predictors and penalty values so that only a single fit would be run on each resampling, which would require much less time than re-fitting the UK model.

The `SpatioTemporal` model uses an optimisation algorithm to produce maximum likelihood estimates for its parameters in both space and time, which takes considerable time. The median computation time of the ST model to fit the sample sets with N_{train} = 500 is 3.02 hours (181.2 minutes) and is 91.44 hours (5,486.5 minutes) when N_{train} = 2,000. Thus, if resources are limited and the number of training site locations is high, ST becomes infeasible when the other models would not.

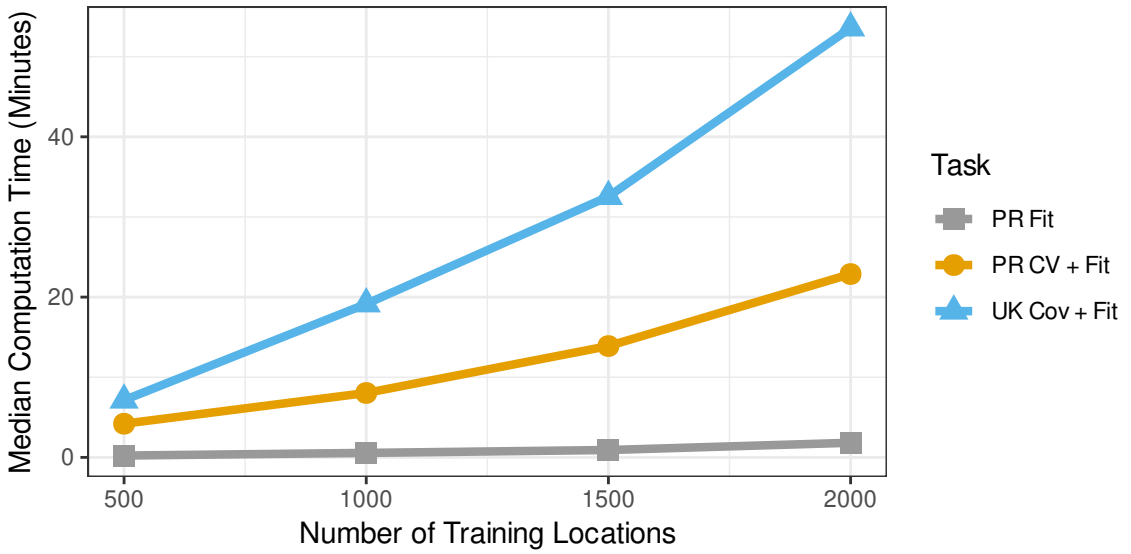

Figure 2.3: Median computation times from 10 replicate simulated data sets with increasing numbers of monitoring locations. For the "PR CV + Fit" time, we used the penalized regression model to select over penalty values and number of TPRS basis functions, then fit the resulting best model ("PR Fit"). The universal kriging fitting time includes estimation of covariance parameters ("UK Cov + Fit").

2.4 Analysis of Ambient Air Quality

2.4.1 Monitoring Data

We demonstrate our method from Section 2.2 by predicting daily ambient total $PM_{2.5}$, total PM_{10}, sulfate (SO_4^{2-}), and silicon (Si) concentrations for the eastern portion of the contiguous United States in 2017 (Figure 2.1). By including only the eastern US, we limit our analysis to monitors that are spatially dense and a region with a similar set of ambient pollution sources. Sulfate is a component of PM that is by itself associated with respiratory and cardiovascular health effects, while silicon is a component that is less studied (U.S. Environmental Protection Agency, 2019). Both species are observed less frequently than total PM, and by different methods of measurement. We use Air Quality System monitoring data obtained September 18, 2019 to provide observed daily measurements of $PM_{2.5}$ and PM_{10} with temporal and geographical metadata. Similarly, we use AQS monitoring data obtained October 16, 2019 for concentrations of sulfate and silicon at the $PM_{2.5}$ size. In addition, we use total $PM_{2.5}$ concentrations (obtained June 22, 2021)

as well as sulfate and silicon $PM_{2.5}$ concentrations (obtained January 14, 2022) from a different monitoring network, the Interagency Monitoring of Protected Visual Environments (IMPROVE) program (Malm et al., 1994). The locations measured by IMPROVE are largely rural, while the AQS data come from mostly urban areas.

Nearly all sulfate and silicon concentrations are observed on every third or every sixth day. Total $PM_{2.5}$ and PM_{10} are measured by some monitors more often than every three days. See Table 2.1 for the distribution of monitoring sites for each pollutant that record values one sixth of the year or less or between one sixth and one third of the year. Figure B.2 in the Supplementary Material depicts the by-date monitoring frequency for each pollutant.

2.4.2 Spatiotemporal Predictors

We fit the data using our model from Equation (2.3) with TPRS as well as exactly one or two other types of predictors. The Community Multiscale Air Quality Modeling System (CMAQ) is a mathematical model that uses atmospheric dispersion and emissions to estimate air quality levels, providing a rich though uncalibrated spatiotemporal predictor for our model (Reff, A. et al., 2020). The EPA provides daily predictions of $PM_{2.5}$ from CMAQ on a 12 km grid across the United States. We use data acquired June 30, 2020 and match each monitor site with the closest grid centroid. For an additional spatiotemporal predictor we use a grid of estimated 3-hour average surface temperature values from the North American Regional Reanalysis (NARR) obtained on July 9, 2021 (Mesinger et al., 2006). We again match each monitoring site location to the closest grid point and estimate daily average temperature values for each site.

2.4.3 Data Filtering and Transformation

Before estimating a spatiotemporal surface, we pre-process both the $PM_{2.5}$ and PM_{10} concentration data in the following ways. We use only the Federal Reference Method (FRM) monitors, which use 24-hour gravimetric measurements (i.e. based on weighing mass accumulated on a fil-

Table 2.1: Summary statistics for ambient pollutant concentrations in 2017 after log tranformation, along with the distribution of monitoring sites by frequency of observation. The monitoring sites are split into groups that were observed for 61 days (one sixth of a year) or less, observed between 61 and 122 days (one third of a year), or observed between 123 and 365 days.

									Number of monitoring sites by days observed		
Pollutant	N	Mean	SD	Min	Q_1	Median	Q_3	Max	0-61 days	62-122 days	123-365 days
log($PM_{2.5}$)	69,693	1.88	0.568	-2.78	1.55	1.93	2.28	4.47	85 (15.3%)	402 (72.2%)	70 (12.6%)
log(PM_{10})	11,383	2.60	0.633	0	2.20	2.64	3.00	6.16	137 (81.5%)	26 (15.5%)	5 (3.0%)
log(SO_4^{2-})	9,160	-0.30	0.860	-10.82	-0.68	-0.24	0.19	2.61	130 (80.7%)	29 (18.0%)	2 (1.2%)
log(Si)	8,868	-3.48	1.351	-10.82	-4.07	-3.35	-2.73	1.31	132 (82.0%)	29 (18.0%)	0 (0%)

ter) and are the basis of assessing compliance with the National Ambient Air Quality Standards. We remove concentrations of value zero, which are likely invalid measurements. If monitors are collocated, we remove all but one measurement per day at that site. We keep the measurement with the lowest "parameter occurrence code", which should be the earliest registered monitor at that site. Finally, to account for the skewedness of PM concentrations, we natural-log transform the observed concentrations and CMAQ values. See Table 2.1 for the summary statistics of the log-transformed $PM_{2.5}$ and PM_{10} data and Figure 2.1 for their average log-transformed concentrations by monitor location in 2017.

Similar filtering is performed on the sulfate and silicon concentrations, although the components are measured with different approaches. The sulfate concentrations are measured via ion chromatography or pulsed fluorescence, while all silicon concentrations are calculated using X-ray fluorescence. As with PM, we remove the values at zero, natural-log transform all concentrations, and use only one measurement at a site per day, with preference to the older or non-IMPROVE monitors. We replace negative values (21 sulfate and 78 silicon measurements) with the lowest observed positive value before being log-tranformed. See Table 2.1 and Figure 2.1 for summary statistics and the average spatial distribution over 2017.

2.4.4 Model Fits

To predict each pollutant, we fit the model from Equation (2.1) with daily average temperatures and TPRS basis functions as predictors. To predict $PM_{2.5}$ we use logged CMAQ values as an additional predictor. Since CMAQ is itself a multisource, deterministic estimate of $PM_{2.5}$, which we have transformed to the same scale as our observations, we exclude its coefficients from the over-fitting penalization $\mathbf{\Gamma}_1$. In contrast, we penalize the temperature and TPRS coefficients because we expect them to explain variation in pollution levels, but they are not direct predictions of pollution levels. As described in Section 2.2.4, we fix $\lambda_2 = 0$ and select the penalty λ_1 by lowest cross-validated RMSE, then select a λ_2 value with λ_1 fixed at the previously chosen value. Reusing this penalty selection process, we calculate cross-validated RMSE with each amount of TPRS basis functions from a chosen set of values (e.g. 100, 200, or 300 for $PM_{2.5}$). Our best model fit is the predictor and penalty set with lowest overall cross-validated RMSE. In the case of $PM_{2.5}$, we verify the accuracy of selecting penalties one at a time by also fitting models for every combination of λ_1 and λ_2 values (e.g. all combinations of the values 0.01, 0.1, 1, 10, and 100) and find that the same penalties are selected by either method. To compare with our model, we also fit UK and the ST model with 10-fold cross validation, using the same procedures as in Section 2.3.

2.4.5 Results

After selection through 10-fold CV on the $PM_{2.5}$ data, the model from Equation (2.3) is fit with 200 TPRS basis functions and the penalty values of $\lambda_1 = 30$ and $\lambda_2 = 0.01$. Our penalized regression approach is matched by UK and outperformed by ST when estimating $PM_{2.5}$. Table 2.2 shows that the cross-validated RMSE over all dates and sites for our model (2.184 μg/m^3) lies between that of UK (2.389 μg/m^3) and ST (1.973 μg/m^3). For UK, the cross-validated RMSE over the yearly average predictions and observations for every site is lower than ST or penalized regression (0.875 μg/m^3 vs. 0.899 and 0.917 respectively). UK likely performs well on an annual average due to its excellent use of spatial structure but blindness to temporal structure. We can see

similar behavior in Table 2.3, where we report summary statistics for daily cross-validated RMSE across all sites. UK has a lower daily CV RMSE on average than our model, but median daily CV RMSE for UK is higher than for penalized regression. The `SpatioTemporal` model has lower mean and median daily CV RMSE than the other two models.

For PM_{10} we select with cross validation 50 TPRS basis functions and the penalty values $\lambda_1 = 15$ and $\lambda_2 = 0.5$. The ST model has lower overall CV RMSE (9.275 μg/m^3) than our model (9.569 μg/m^3) and UK (13.503 μg/m^3), but our penalized regression model produces the lowest annual average cross validated RMSE (4.375 μg/m^3 vs. 4.446 for ST and 4.746 for UK). The ST model also has the lowest daily average and median CV RMSE values of any model for PM_{10}. Our model has lower overall, annual average, daily mean, and daily median cross validated RMSE values than universal kriging (Tables 2.2 and 2.3). From Figure 2.1, we see that the PM_{10} data have fewer observations and fewer unique locations than $PM_{2.5}$. There are 40 days of the year having two or fewer observed concentrations, which are not predicted by any model due to the lack of data. See Section 2.5 for a discussion of possible interpolation work-arounds for this issue. With fewer observed sites each day and greater variance (Table 2.1), UK may not have enough spatial information on each day to outperform our model.

On the sulfate concentrations, we choose via cross validation 45 TPRS basis functions and the penalty values $\lambda_1 = 15$ and $\lambda_2 = 0.001$. Our penalized regression model demonstrates the best predictive accuracy in overall cross validated RMSE (0.456 μg/m^3 vs. 0.535 for ST and 0.771 for UK), annual average CV RMSE (0.193 μg/m^3 vs. 0.223 for ST and 0.279 for UK), and both the daily average and daily median. (Tables 2.2 and 2.3). From Table 2.1 we see that nearly all monitors that measure sulfate and silicon record concentrations every third day or less frequently. So in Table 2.3 we see that only 173 days of 2017 are being predicted on by our model and ST. UK predicts on 124 days of the year since it requires that a cross validation fold contain at least two observations on a given date to create a distance matrix and estimate covariance parameters. The temporal missingness in sulfate leads to the same outcome as in Section 2.3, where our penalized smoother is able to retain the greater accuracy than its competitors. Figure B.3 in the Supple-

Table 2.2: Cross-validated RMSE and R^2 values across all dates and sites ("Overall") and from by-site annual average predictions and observations ("Annual Average"). We fit penalized regression (PR) per Equation (2.3) using daily average temperature values and some number of TPRS basis functions (see Section 2.4.5) as predictors. For $PM_{2.5}$ we also added logged CMAQ values as predictors.

		Overall		Annual Average	
Pollutant	Model	RMSE	R^2	RMSE	R^2
$PM_{2.5}$	PR	2.184	0.731	0.917	0.629
	ST	1.973	0.780	0.899	0.634
	UK	2.389	0.677	0.875	0.651
PM_{10}	PR	9.569	0.311	4.375	0.260
	ST	9.275	0.350	4.446	0.235
	UK	13.503	0.009	4.746	0.130
SO_4^{2-}	PR	0.456	0.530	0.193	0.627
	ST	0.535	0.373	0.223	0.592
	UK	0.771	0.057	0.279	0.434
Si	PR	0.119	0.455	0.038	0.508
	ST	0.137	0.308	0.047	0.309
	UK	0.138	0.308	0.041	0.472

mentary Material depicts the observations and predictions over time for sulfate concentrations at four randomly chosen monitoring sites. In East Baton Rouge, LA, where observed concentrations are more variable than at the other sites shown, our model fits more closely than the ST model. It may be that the smooth time trend applied to the ST model makes following highly variable observations difficult.

For the silicon concentrations we cross-validate and select 55 TPRS basis functions with the penalties $\lambda_1 = 5$ and $\lambda_2 = 0.001$. Similar to sulfate, our model performs well under temporal missingness, producing the lowest overall CV RMSE (0.119 μg/m^3 vs. 0.137 for ST and 0.138 for UK) and lowest annual average CV RMSE (0.038 μg/m^3 vs. 0.047 for ST and 0.041 for UK) (Table 2.2). In daily CV RMSE, the average of our model lies above UK and below ST, while the median of our model is higher than both that of ST and of UK (Table 2.3).

Table 2.3: Summary statistics for daily cross-validated RMSE values across all monitoring sites. We fit penalized regression (PR) per Equation (2.3) using daily average temperature values and some number of TPRS basis functions (see Section 2.4.5) as predictors. For $PM_{2.5}$ we also added logged CMAQ values as predictors.

Pollutant	Model	T	Mean	SD	Min	Q_1	Median	Q_3	Max
$PM_{2.5}$	PR	365	2.339	0.912	1.076	1.719	2.145	2.740	7.166
	ST	365	2.117	0.838	0.951	1.572	1.951	2.43	6.2
	UK	365	2.329	0.844	1.021	1.776	2.187	2.588	6.249
PM_{10}	PR	325	11.768	8.463	1.641	6.380	9.234	15.029	59.477
	ST	325	11.211	8.162	1.072	6.354	8.974	13.088	57.135
	UK	325	15.380	8.248	2.215	10.239	13.127	18.169	61.267
SO_4^{2-}	PR	173	0.420	0.279	0.069	0.264	0.374	0.519	3.007
	ST	173	0.464	0.269	0.102	0.316	0.420	0.523	2.760
	UK	124	0.712	0.287	0.163	0.533	0.676	0.845	1.942
Si	PR	122	0.079	0.085	0.012	0.030	0.049	0.087	0.441
	ST	122	0.084	0.106	0.010	0.026	0.046	0.087	0.502
	UK	122	0.076	0.088	0.009	0.026	0.044	0.080	0.464

2.5 Discussion

We have presented a penalized regression model for spatiotemporal prediction that penalizes overfitting and smoothes over predictions at adjacent timepoints. Using spatiotemporal covariates and TPRS basis functions as predictors, we predict daily values anywhere on a spatial domain. In Section 2.3, we demonstrate in simulations that our smoothing method solving Equation (2.1) can outperform day-by-day universal kriging under a variety of data conditions and outperform the `SpatioTemporal` model when observations are less frequent than daily. When the data increase in spatial locations, we see that our model is faster than the each-day application of UK while the ST model is more than forty times slower than either. In Section 2.4, we find that our model performs with cross-validated predictive accuracy close to that of day-by-day UK and worse than that of the ST model on total $PM_{2.5}$ and PM_{10} concentrations. But on sulfate and silicon concentrations, our penalized regression model achieves the best accuracy of the models in all but two metrics, where it performs similarly.

Our model proposed in Equation (2.1) has spatiotemporal predictive accuracy for $PM_{2.5}$ concentrations that is on par with UK, a method shown to be more accurate than the several other approaches tested in Berrocal et al. (2020). Since UK is an excellent spatial model, it is difficult to outperform when, for each time point, there is a nearly spatially complete set of observations with a strong spatial signal. In Section 2.3, both increasing non-spatial error (lowering the signal to noise ratio) and increasing daily fluctuation in the latent Gaussian processes ($H = 0.05$) in our simulations results in a larger drop in accuracy for UK than for our model. When concentrations are observed less frequently at some sites, as in the case of PM_{10}, sulfate, and silicon, there are dates when only a few measurements are available and UK drops in accuracy behind our model. For sulfate and silicon, our model is more accurate than the more complex and computationally expensive `SpatioTemporal` model, showing that we have developed a useful prediction method for less-studied ambient air pollutants.

In both the simulations (Section 2.3) and the ambient concentrations analysis (Section 2.4), our model provides predictions only on dates where at least one measurement is observed. If a timepoint t^* is never observed in the data, there will be no p-vector of coefficients $\boldsymbol{\beta}_{t^*}$ estimated via Equation (2.3). Furthermore, our model is treating any two dates with no observations between them as "adjacent in time". In the case of the silicon concentrations in Section 2.4, all monitoring sites recorded measurements every third day, so our model penalized the differences between predictions three days apart, ignoring the two days in between them. We can define any interval of time in our data as "adjacency" by altering the matrix $\boldsymbol{D}$ so that values at a desired amount of time apart are differenced. To obtain predictions for a date that is never observed, we may use some form of interpolation, such as simply averaging between predictions at each site before and after the missing date. Alternatively, we could average the coefficients $\boldsymbol{\beta}_{t^*-1}$ and $\boldsymbol{\beta}_{t^*+1}$ and use the interpolated $\widehat{\boldsymbol{\beta}}_{t^*}$ to predict values over the unobserved date.

The selection of penalty values and predictor sets for our model is time-consuming and can require manual tuning. In our simulations and applications to observed ambient concentrations, we use a faster sequential method described in Section 2.2.4 to select penalty values and repeat

for each set of TPRS basis functions. This sequential method only requires the manual input of candidate value sets for the penalties and amounts of TPRS basis functions. After the penalties and number of TPRS basis functions for our model are chosen, to predict a spatiotemporal surface we need to calculate the inverse of a sparse pT x pT matrix only once. To perform measurement error correction with non-parametric bootstrapping, we may reuse the same chosen penalty values and TPRS basis functions on each bootstrapped sample, making the computation very feasible. This is important because a non-parametric bootstrap measurement error correction can account for site selection, in addition to parameter estimation.

Extensions to the model, such as predicting multiple pollutants at once, may be possible through clever application of new penalty terms. Although, if a non-convex penalty is used, the closed form solution from Equation (2.3) would no longer apply. Generalized optimization techniques, such as ADMM (Boyd et al., 2010) should be able to utilize the sparse structure of the matrices and produce an accurate $\boldsymbol{\beta}$ estimate relatively fast, but not as fast as Equation (2.3).

The penalized regression model proposed in this study showed spatiotemporal predictive ability that is competitive with both universal kriging applied daily and the `SpatioTemporal` model. When the data are missing observations for whole time points, our model achieves better predictive accuracy than either of the other models. The proposed model is also unlike more complicated counterparts in that it can be used with bootstrap-based measurement error correction for epidemiological health effects models.

Chapter 3

Principal Stratification in Longitudinal Trials with Treatment Crossover for Application to Indoor Air Pollution Interventions

3.1 Introduction

Exposure to household air pollution is a global health burden, contributing through respiratory and cardiovascular diseases to 3.6% of disability-adjusted-life-years (DALYs) in all ages (95% uncertainty interval 2.7 to 4.6) (Murray et al., 2020). As of 2010, 41 % of the world's population burns primarily solid fuels such as wood and charcoal for cooking in their households (Bonjour et al., 2013). The emissions of these cookstoves and their contribution to household air pollution have motivated randomized controlled interventions of cleaner, better ventilated, or better filtered stoves. We address one such randomized controlled trial (RCT) with a stepped-wedge design for a biomass-burning cookstove intervention in Honduras (Young et al., 2019). With already established links between traditional biomass-burning cookstoves and cardiovascular diseases (Baumgartner et al., 2011; Clark et al., 2013; McCracken et al., 2007), Young et al. (2019) aimed to investigate the relationships of indoor air pollution and broader cardiometabolic health. As an intervention, the researchers installed improved biomass cookstoves (called the Justa stove) in rural Honduran households, which were designed to have reduced indoor emissions. We aim to estimate the causal effect of the cookstove intervention on percent of glycated hemoglobin (HbA1c) for the primary cook in the household.

Other RCTs that have analyzed the intent-to-treat (ITT) effects of cookstove interventions on household air pollution health burdens have often produced mixed or even null results. Improved cookstove interventions performed in Guatemala, Nepal, Mexico, India, and Peru each failed to find significant ITT effects, likely due to low compliance to treatment that resulted in household

air pollution exposures that were not sufficiently reduced (Clasen et al., 2022; Romieu et al., 2009; Smith et al., 2011; Tielsch et al., 2016). Non-compliance is common in human subjects research, with participants taking no treatment when assigned or taking a treatment other than their own. Presence of "non-compliers" leads to attenuation in the estimate of an ITT effect, since some who are not ultimately treated will be analyzed as part of the treatment group and vice versa. Compliance can be particularly difficult in environmental studies which are typically unblinded and during which treatment adherence is only monitored at a few time points (e.g. quarterly). To measure the efficacy of a cookstove intervention on household air pollution related health outcomes, we endeavor to find an alternative to the ITT effect which does not suffer the same attenuation from non-compliance.

We consider principal effects, such as an average causal effect for only those who would always comply with their assigned treatment (i.e. compliers). Frangakis and Rubin (2002) explain that principal effects are in fact causal effects, and introduce the method of principal stratification to identify them. Hidden principal strata are defined by the potential outcomes (under each of the treatments being compared) of a post-treatment variable, such as compliance, and are unaffected by the observed treatment itself. The earlier method developed by Frangakis and Rubin (2002) as well as other non-parametric (Ding and Lu, 2017), Bayesian (Hackstadt et al., 2014; Jin and Rubin, 2008; Peng et al., 2015), and two-stage (Woo et al., 2023) principal stratification methods have been applied to parallel-group RCTs with a single outcome measure, where each subject is assigned to a group that receives only the treatment or control and the subject's outcome is summarized in a single value. For parallel-group RCTs such as the CLEAN AIR study (Woo et al., 2023) and PREACH study (Peng et al., 2015), the latent principal strata are defined by the post-treatment variable of reduction in indoor particulate matter (PM) and individuals' estimated probabilities of stratum membership are used to estimate stratum-specific average causal effects. In contrast, Hackstadt et al. (2014) impute the potential PM and potential outcome values first, from which they calculate stratum-specific average causal effects.

Environmental trials with human subjects may often follow non-parallel group designs, as reasons such as ethical concerns prevent researchers from withholding treatment from any subjects. Instead they may use a longitudinal crossover design, where subjects are repeatedly measured and all eventually receive the treatment. In this setting an ITT estimate still suffers from attenuation, but principal stratification methods such as the one introduced in Ding and Lu (2017) cannot be directly applied to the full data. Parallel group or not, few principal stratification methods have been applied to measurements over time in a longitudinal study. Frangakis et al. (2004) work with a time-to-event outcome and both treatment and compliance that vary in time. Subjects' treatment exposure is influenced by their distance from a clinic, which leads to assumptions of multilevel monotonicity and compound exclusion restriction. Dai et al. (2012) use a hidden Markov model with a time-to-event outcome, having subjects transition through "principal states" and estimating a cumulative risk ratio and discrete hazard ratio by state. We do not have a time-to-event outcome and are more comparable to Lin et al. (2008) and Lin et al. (2009), who have developed a Bayesian and hidden markov approach using "compliance superclasses", which summarize the varying compliance classes of subjects over time into strata such as "decreasing compliers" or "high compliers". These superclasses still contain mixed compliance, so the issue of attenuation may persist and interpretation of the intervention's effect can be difficult for policy implementation.

We propose two methods applicable to longitudinal settings that estimate "at-the-time" compliance class causal effects. That is, after imputing the latent principal stratum for each subject and timepoint, we estimate an average causal effect for subjects who were in the same compliance class "at-the-time". In Section 3.2 we review principal strata and describe a set of assumptions, a method of principal scores, and covariate-adjusted estimands all adapted from Ding and Lu (2017) to a longitudinal setting. In Section 3.3 we introduce our extension of principal scores to a weighted linear mixed model and describe a Bayesian approach to principal stratification for a longitudinal study. We test model performance with simulated samples in Section 3.4 and apply the models to the motivating study, a Honduran cookstove intervention (Young et al., 2019), in Section 3.6. We discuss the paper's conclusions, method limitations, and next steps in Section 3.7.

3.2 Principal Stratification with Principal Scores in a Longitudinal Setting

We summarize the setup, weight formulation, and estimands from Ding and Lu (2017), adding the dimension of time t where applicable to extend their work to a longitudinal setting. We can apply this estimation strategy on repeated measures, ignoring their correlation. In Section 3.3 we introduce two models which implement principal stratification for a crossover design and include a participant-specific effect.

3.2.1 Notation and Setting

Consider a randomized trial comparing a treatment to control, i.e. an experiment in which the assignment of subject $i \in \{1, \ldots, n\}$ to one of two study arms is random. Here we refer to a treatment and intervention (i.e. installation of a ventilated cookstove) interchangeably. In a parallel study, each arm would receive only treatment or only control over the duration of the study, while a crossover design would assign each arm both control and treatment over the course of the time points $t \in \{1, \ldots, T\}$. For the stepped-wedge crossover design, researchers start with assigning each arm the control then begin assigning treatment to one arm before the other (Brown and Lilford, 2006). We consider the case with only two study arms for simplicity, although the methods we introduce could be applied without adjustment to a crossover study with more than two arms. We denote Z_{it} as the assignment of treatment (1) or control (0) to subject i at time t and $\mathbf{Z}_i$ as a vector of length T with the treatment assignments for subject i over all time points. A subject's treatment assignments over time will follow one of only two possible patterns ($\mathbf{z}_1$ or $\mathbf{z}_2$) determined by their study arm. We define Y_{it} as the continuous or binary outcome value for subject i at time t and define the post-treatment variable S_{it} as the actual treatment usage ($S_{it} = 0$ if subject i did not take the treatment at time t and $S_{it} = 1$ if they did). Let $\mathbf{x}_i$ denote covariates that are measured only pre-randomization. We use a potential outcomes framework, signifying $S_{it}(z)$ as the potential treatment used and $Y_{it}(z)$ the potential response value for subject i at time t under the assignment of treatment ($z = 1$) or not ($z = 0$). The vectors $\mathbf{Y}_i(\mathbf{z}_k)$ and $\mathbf{S}_i(\mathbf{z}_k)$ are the sets of

potential response values and potential actual treatments used for all timepoints of subject i under the treatment patterns $k = 1$ or 2.

3.2.2 Principal Strata

We define four principal strata for the subjects' compliance behavior at time t, defined by the potential outcomes $S_{it}(0)$ and $S_{it}(1)$. These principal strata are the same as developed by Frangakis and Rubin (2002), now also indexed by time. If subject i at time t is a *complier* (c), then $S_{it}(0) = 0$ and $S_{it}(1) = 1$. An *always-taker* (a) has $S_{it}(0) = S_{it}(1) = 1$ and a *never-taker* (n) has $S_{it}(0) = S_{it}(1) = 0$. Finally, a *defier* ($d$) would do the opposite of what they are assigned, i.e. $S_{it}(0) = 1$ and $S_{it}(1) = 0$. We denote a subject's latent principal stratum at time t as $U_{it} \in \{c, n, a, d\}$ and the vector of their principal strata over time as $\mathbf{U}_i$. Strata may change over time, as practically we may see a subject comply, take treatment when unassigned, and neglect to take treatment when assigned, all in the same experiment. In addition, strata for a subject may change in time while their behavior does not, e.g. a complier assigned to treatment becoming an always taker assigned to treatment. We estimate average causal effects for each stratum "at-the-time", i.e. across observations at any time from any subject that belong to a specific stratum. These average causal effects (ACEs) can be calculated as $\mathrm{ACE}_u = E\{Y_{it}(1) - Y_{it}(0) \mid U_{it} = u\}$ for $u = \{c, n, a, d\}$.

3.2.3 Assumptions

We make the following assumptions to ensure identifiability of the causal effects.

Assumption 1: Stable Unit Treatment Value (Rubin, 1980) — we assume there is only a single version of treatment i.e. that the potential outcomes $S_{it}(0)$, $S_{it}(1)$, $Y_{it}(0)$, and $Y_{it}(1)$ may be written as single values. Further, we assume no interference between subjects, so that the potential outcome of a subject is not affected by the treatment assignment of a different subject.

Assumption 2: Randomization — i.e. $\mathbf{Z}_i \perp\!\!\!\perp \{\mathbf{S}_i(\mathbf{z}_1), \mathbf{S}_i(\mathbf{z}_2), \mathbf{Y}_i(\mathbf{z}_1), \mathbf{Y}_i(\mathbf{z}_2), \mathbf{U}_i, \mathbf{x}_i\}$ for any i and the two treatment assignment patterns $\mathbf{z}_1$ and $\mathbf{z}_2$. Since the assigments $\mathbf{Z}_i$ in a crossover design are randomized to be either $\mathbf{z}_1$ or $\mathbf{z}_2$, they are independent at all timepoints of the potential actual

treatments $\mathbf{S}_i(\cdot)$, the potential outcomes $\mathbf{Y}_i(\cdot)$, the pre-randomization covariates $\mathbf{x}_i$, and the latent principal strata $\mathbf{U}_i$, which determine compliance behavior.

Assumption 3: Monotonicity — i.e. $S(1)_{it} \geq S(0)_{it}$ for all i and t. When S is binary, this means that a subject would not at any time be a defier ($S(1)_{it} = 0$ and $S(0)_{it} = 1$), leaving the just three of the strata: compliers, always-takers, and never-takers.

Assumption 4: General Principal Ignorability (GPI) — i.e. $Y_{it}(z) \perp\!\!\!\perp U_{it} \mid \mathbf{x}_i$ for $z = 0$ and $z = 1$. The distribution of potential outcomes for Y_{it} under treatment and control are the same across the latent principal strata when controlling for $\mathbf{x}_i$. More specifically, for $z = 0$ and $z = 1$:

$$E\{Y_{it}(z) \mid U_{it} = c,\ \mathbf{x}_i\} = E\{Y_{it}(z) \mid U_{it} = n,\ \mathbf{x}_i\} = E\{Y_{it}(z) \mid U_{it} = a,\ \mathbf{x}_i\}.$$

3.2.4 Non-parametric estimation for a single outcome measure

We rely on the assumptions in Section 3.2.3 and principal scores $e_u(\mathbf{x}_i) = \Pr(U_i = u \mid \mathbf{x}_i)$ for $u \in \{c, n, a\}$ to identify and estimate average causal effects. Here we review the method developed in Ding and Lu (2017) to calculate these principal scores and use them to create weighted samples. We retain the index in time t where applicable even though the method of Ding and Lu (2017) does not explicitly account for repeated measures.

First we note that under the assumption of monotonicity in Section 3.2.3, the observed data can be categorized so that $Z_{it} = 0$ and $S_{it} = 0$ indicate that subject i is either a never-taker or complier at time t. Similarly, ($Z_{it} = 1$, $S_{it} = 1$) is the behavior of an always-taker or complier, while ($Z_{it} = 1$, $S_{it} = 0$) must be a never-taker and ($Z_{it} = 0$, $S_{it} = 1$) must be an always-taker. Then to form principal scores we make use of the probabilities of S_{it} by assignment, $p_1 = \Pr(S_{it} = 1 \mid Z_{it} = 1)$ and $p_0 = \Pr(S_{it} = 1 \mid Z_{it} = 0)$, as well as those conditional on the covariates $\mathbf{x}_i$, $p_1(\mathbf{x}_i) = \Pr(S_{it} = 1 \mid Z_{it} = 1, \mathbf{x}_i)$ and $p_0(\mathbf{x}_i) = \Pr(S_{it} = 1 \mid Z_{it} = 0, \mathbf{x}_i)$. Thus the principal scores as a function of the covariates $\mathbf{x}_i$ are $\boldsymbol{e}_c(\mathbf{x}_i) = p_1(\mathbf{x}_i) - p_0(\mathbf{x}_i)$ for compliers, $\boldsymbol{e}_n(\mathbf{x}_i) = 1 - p_1(\mathbf{x}_i)$ for never-takers, and $\boldsymbol{e}_a(\mathbf{x}_i) = p_0(\mathbf{x}_i)$ for always-takers. As done by Ding and Lu (2017), we estimate $p_1(\mathbf{x}_i)$ and $p_0(\mathbf{x}_i)$ by modeling $\Pr(\mathbf{U} \mid \mathbf{X})$ as a three-level multinomial

logistic model and using maximum likelihood estimation via the EM algorithm. The overall proportions of the strata in the data are $\pi_c = p_1 - p_0$, $\pi_n = 1 - p_1$, and $\pi_a = p_0$. Note that the principal scores and stratum proportions are without respect to time t, so the resulting weights applied to the data will not change across timepoints within the same subject.

Combining the principal scores and proportions we produce the weights

$$\begin{aligned} w_{1,c}(\mathbf{x}_i) &= \frac{e_c(\mathbf{x}_i)}{e_c(\mathbf{x}_i) + e_a(\mathbf{x}_i)} \Big/ \frac{\pi_c}{\pi_c + \pi_a}, & w_{0,c}(\mathbf{x}_i) &= \frac{e_c(\mathbf{x}_i)}{e_c(\mathbf{x}_i) + e_n(\mathbf{x}_i)} \Big/ \frac{\pi_c}{\pi_c + \pi_n}, \\ w_{1,a}(\mathbf{x}_i) &= \frac{e_a(\mathbf{x}_i)}{e_c(\mathbf{x}_i) + e_a(\mathbf{x}_i)} \Big/ \frac{\pi_a}{\pi_c + \pi_a}, \quad \text{and} & w_{0,n}(\mathbf{x}_i) &= \frac{e_n(\mathbf{x}_i)}{e_c(\mathbf{x}_i) + e_n(\mathbf{x}_i)} \Big/ \frac{\pi_n}{\pi_c + \pi_n}. \end{aligned} \tag{3.1}$$

Then the complier average causal effect (CACE), never-taker average causal effect (NACE), and always-taker average causal effect (AACE) may be estimated

$$\begin{aligned} \text{CACE} &= E\left\{w_{1,c}(\mathbf{x}_i) Y_{it} \mid Z_{it} = 1, S_{it} = 1\right\} - E\left\{w_{0,c}(\mathbf{x}_i) Y_{it} \mid Z_{it} = 0, S_{it} = 0\right\} \\ \text{NACE} &= E\{Y_{it} \mid Z_{it} = 1, S_{it} = 0\} - E\left\{w_{0,n}(\mathbf{x}_i) Y_{it} \mid Z_{it} = 0, S_{it} = 0\right\} \\ \text{AACE} &= E\left\{w_{1,a}(\mathbf{x}_i) Y_{it} \mid Z_{it} = 1, S_{it} = 1\right\} - E\{Y_{it} \mid Z_{it} = 0, S_{it} = 1\} \end{aligned} \tag{3.2}$$

Ding and Lu (2017) formulate an asymptotically more efficient estimator using covariate adjustment. For simplicity, we present only the CACE version of the estimator here. We first estimate the coefficient $\boldsymbol{\beta_{1,c}}$ from weighted least squares regression of $\mathbf{Y}$ on $\mathbf{X}$ using only data with ($Z_{it} = 1$, $S_{it} = 1$) and the weights $w_{1,c}(\mathbf{X})$. This estimate written explicitly is $\widehat{\boldsymbol{\beta}}_{1,c} = \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Y}$, on the subsetted data with with $w_{1,c}(\mathbf{X})$ the diagonal of $\mathbf{W}$. Similarly we estimate $\boldsymbol{\beta}_{0,c}$ using only data with ($Z_{it} = 0$, $S_{it} = 0$) and the weights $w_{0,c}(\mathbf{X})$. Then the estimand for the complier average causal effect is as follows:

$$
\begin{aligned}
\text{CACE} = E\big\{ w_{1,c}(\mathbf{x}_i) \left(Y_{it} - \widehat{\boldsymbol{\beta}}_{1,c}^{\mathrm{T}} \mathbf{x}_i \right) \mid Z_{it} = 1, S_{it} = 1 \big\} & \\
- E\left\{ w_{0,c}(\mathbf{x}_i) \left(Y_{it} - \widehat{\boldsymbol{\beta}}_{0,c}^{\mathrm{T}} \mathbf{x}_i \right) \mid Z_{it} = 0, S_{it} = 0 \right\} & \\
+ \left(\widehat{\boldsymbol{\beta}}_{1,c} - \widehat{\boldsymbol{\beta}}_{0,c} \right)^{\mathrm{T}} E\left\{ w_{z,c}(\mathbf{x}_i) \mathbf{x}_i \mid Z_{it} = S_{it} = z \right\}, &
\end{aligned}
\tag{3.3}
$$

where in the third term, $z = 0$ and 1 so that $w_{z,c}(\mathbf{x}_i)$ is $w_{1,c}(\mathbf{x}_i)$ when $Z_{it} = S_{it} = 1$ and is $w_{0,c}(\mathbf{x}_i)$ when $Z_{it} = S_{it} = 0$. The covariate-adjusted estimators for the NACE and AACE are provided in Section C.1 of the Supplementary Material. To measure uncertainty for the estimands in Equations (3.2) or (3.3) we create bootstrap samples and construct 95% confidence intervals from the sample estimate quantiles. Resampling is done by participant rather than by observation.

We can employ this covariate-adjusted principal scores (CAPS) method for estimating average causal effects by strata in a longitudinal setting, but there are limits to its application. The estimand in Equation (3.3) does not account for any trend in time, such as seasonality. Using repeated measurements when weights and effect estimates are ignorant of time and individual will result in bias from any existing temporal trends. We can prevent this bias by restricting to a single time point where the study arms are not assigned to the same treatment, but such a reduction in the data set would reduce the power of the analysis. Balancing between temporal bias and sample size, we may include more than one timepoint in the data set and add the time index as categorical variables in the covariate adjustment. In a stepped-wedge study, where all participants are assigned the control at first and all are assigned the treatment at the end, we cannot adjust for every time point in the subsetted datasets since, for example, there are no observations where $Z = 0$ during the last time point. We would like to extend this principal scores method from Ding and Lu (2017) to a longitudinal setting with more explicit consideration of time and repeated measures.

3.3 Longitudinal Principal Stratification Methods with Random Effects

We present two methods for estimating the effect of treatment using principal stratification when compliance is defined "at-the-time" and while accounting for repeated measures. In the first we create the principal score weights and estimate causal effects by subsetting the data and applying a weighted linear mixed model. In the second we extend a Bayesian approach from Peng et al. (2015), a model with mixture by latent stratum membership and accounting for repeated measures in the outcome model. Each method builds on an existing model and applies them to a previously not well-studied context. Both methods can be applied to a longitudinal crossover or parallel-group study which satisfies the principal stratification and model assumptions.

3.3.1 Linear Mixed Subset Model Weighted Via Principal Scores

We now extend the causal framework developed by Ding and Lu (2017) to account for repeated measures. Our approach uses the same principal score weights from Equation (3.1), but instead of Equations (3.2) or (3.3), we estimate each average causal effect as a coefficient from a principal score weighted linear mixed (PS-WLM) model on subsets of the data grouped by possible stratum. The data is subsetted in the same way as in Section 3.2.4 to obtain $\widehat{\boldsymbol{\beta}}_{1,c}$ and $\widehat{\boldsymbol{\beta}}_{0,c}$. For example, we estimate the complier average causal effect via a PS-WLM model using only data with $(Z_{it} = 0, S_{it} = 0)$ or $(Z_{it} = 1, S_{it} = 1)$, which represent be all possible "at-the-time" compliers. Similarly, we estimate the AACE with data such that $(Z_{it} = 0, S_{it} = 1)$ or $(Z_{it} = 1, S_{it} = 1)$ and the NACE with data such that $(Z_{it} = 0, S_{it} = 0)$ or $(Z_{it} = 1, S_{it} = 0)$. The assumed model fit to each subset is

$$(Y_{it} \mid U_{it} = u, Z_{it} = z, \mathbf{x}_i, \nu_i, w_{z,u}) \sim N(\alpha_t + \mathbf{x}_i^\top \boldsymbol{\beta} + z\beta' + \nu_i,\ w_{z,u}(\mathbf{x}_i)\sigma^2). \tag{3.4}$$

The conditional mean of Y_{it} includes a fixed effect by time point, the dot product of the pre-randomization covariates $\mathbf{x}_i$ with their respective coefficients $\boldsymbol{\beta}$, the true effect of treatment as-

signment β', and a random intercept by subject $\nu_i \overset{iid}{\sim} N(0, \sigma_\nu)$. For the example of the complier average causal effect (i.e. $U_{it} = c$), the weight $w_{z,u}(\mathbf{x}_i)$ takes the value $w_{1,c}(\mathbf{x}_i)$ when $Z_{it} = 1$ or $w_{0,c}(\mathbf{x}_i)$ when $Z_{it} = 0$. Similarly, $w_{z,u}(\mathbf{x}_i) = 1$ or $w_{0,n}(\mathbf{x}_i)$ for the never-taker subset, and $w_{z,u}(\mathbf{x}_i) = w_{1,a}(\mathbf{x}_i)$ or 1 for the always-taker subset. Our estimate for the average causal effect of treatment assignment within a stratum $U_{it} = u$ is $\widehat{\beta}'$. We measure the uncertainty of the fixed effect $\widehat{\beta}'$ with a Wald confidence interval, and in just the first simulation scenario of Section 3.4.2, we also use the much slower parametric bootstrap.

The covariate-adjusted estimands in Equation (3.3) cannot make use of all time points in a data set without suffering bias from trends in time. Our linear mixed model weighted with principal scores can make use of all time points and account for repeated measures with random effects. The principal score weights from Equation (3.1) form a pseudo population by strata when applied to the data subsets, from which we estimate average causal effects. Instead of using differences in means or residual means as in Section 3.2.4, we estimate these effects with a weighted linear mixed model.

Here we point out that our assumed model bears similarity to that of Lin et al. (2008) and (2009), although the estimation approaches are not alike and ultimately we are concerned with a different estimand. Because we are targeting causal effects for compliance at-the-time, we make the strong assumption that treatment does not affect subsequent time points, while Lin et. al. account for cumulative effects within their model. They develop compliance "superclasses", latent strata for participants over all timepoints in a study, and thereby summarize the time-varying compliance classes. Instead of estimating an average causal effect for those who are compliers, they consider "high compliers", who are probable to be complying over time. With some non-compliance still present in the estimate, attenuation is still possible and the full efficacy of the treatment on those who use it could remain unknown, weakening the interpretation of results for policy implementation. For further discussion of our limited treatment effect time assumption and its implications, see Section 3.7.

3.3.2 Principal Stratification in Bayesian Framework

Alternative to the principal scores approach, we also use a Bayesian mixture model (BM) to estimate average causal effects by latent strata "at-the-time". Our model follows that of Peng et al. (2015), where the outcome Y_{it} has a normal distribution with mean μ_{zu} and standard deviation σ_u^2. We define Y_{it} as the difference in outcome for subject i at time t from the overall average and prepare the indices $z \in \{0, 1\}$ and $u \in \{c, n, a\}$. Building further on Peng et al. (2015), we consider the latent principal stratum by subject and time, $U_{it} = u$, as a multinomial random variable with the density

$$f\left(U_{it} \mid \mathbf{q}_i\right) \propto \pi_{ic}^{\mathbf{1}\{U_{it}=c\}} \pi_{ia}^{\mathbf{1}\{U_{it}=a\}} \pi_{in}^{\mathbf{1}\{U_{it}=n\}}, \quad \text{where} \quad \pi_{iu} = \frac{\exp\left(\mathbf{q}_i^\top \boldsymbol{\gamma}_u\right)}{\sum_{u'} \exp\left(\mathbf{q}_i^\top \boldsymbol{\gamma}_{u'}\right)}.$$

Note that $\mathbf{q}_i^\top$ is some set of covariates, $\boldsymbol{\gamma}_u$ are their coefficients and $\sum_{u'}$ is the sum over $u' \in \{c, n, a\}$. Similar to the principal scores approach, the probabilities π_{iu} rely on compliance information from Z_{it} and S_{it}, so that when for example $Z_{it} = 0$ and $S_{it} = 1$, then $\pi_{ia} = 1$ and $\pi_{in} = \pi_{ic} = 0$.

The model for the outcome is

$$\left(Y_{it} \mid Z_{it} = z, U_{it} = u\right) \sim N\left(\alpha_t + \mu_{zu} + \mathbf{x}_i^\top \boldsymbol{\beta} + \nu_i,\, \sigma_u^2\right),$$

and we estimate the average causal effect for a stratum u as $E\left[Y_{it}(1) - Y_{it}(0) \mid U_{it} = u\right] = \mu_{1u} - \mu_{0u}$. The random effect is distributed $\nu_i \overset{iid}{\sim} N\left(0, \sigma_\nu^2\right)$, and we use a normal prior distribution for each of γ_u, σ_u^2, $\boldsymbol{\beta}$ and μ_{zu}. We use the probabilistic programming language Stan via the `R` package `rstan` to perform Markov chain Monte Carlo with the NUTS algorithm (Stan Development Team, 2023a,b), and do not need conjugate priors. We list the prior values for each parameter in Table C.1 of the Supplementary Material.

3.4 Simulation

3.4.1 Setup

To compare our proposed methods of estimating the "at-the-time" stratified causal effects, we perform a simulation study under several data scenarios. We create a stepped-wedge design similar to the motivating study we analyze in Section 3.6. Participants are randomized to one of two study arms and are each measured for six time points ($t = 1, 2, 3, 4, 5, 6$), with participants in one arm first receiving treatment at the third time point and in the other arm receiving it first at the fifth time point. Thus, both arms begin under the control and both are receiving treatment by the fifth time point.

To generate the latent principal stratum for each subject and time point, U_{it}, we first draw the starting quantities of stratum members from a multinomial distribution with the probabilities 2/3, 1/6, and 1/6 for compliers, always-takers, and never-takers respectively. Once we have randomized an initial stratum for each participant by the drawn quantities, we use a Markov chain to produce strata at each time point for each participant. The transition matrix for this Markov chain is symmetric with the values 0.6 in its diagonal and 0.2 in its off-diagonal. The stratum U_{it} is then used to determine S_{it}, the value of the actual treatment for participant i and time t.

The model for the simulated outcomes is as follows,

$$Y_{it} = \mathbf{x}_i^\top \boldsymbol{\beta} + \beta' S_{it} + \sin(t) + \nu_i + \epsilon_{it}.$$

Here $\beta' = -0.5$ is the true at-the-time effect of received treatment (S_{it}) on the outcome, $\epsilon_{it} \overset{iid}{\sim} N(0, 1)$ is the error across all measurements, and $\nu_i \overset{iid}{\sim} N(0, \sigma_\nu^2)$ is participant-specific effect over repeated measures. We generate five covariates to include in $\mathbf{x}_i$ ($\mathbf{g}_1$, $\mathbf{g}_2$, $\mathbf{g}_3$, $\mathbf{g}_4$, and $\mathbf{g}_5$) that are constant in time. These covariates are each distributed independently across subjects, with $\mathbf{g}_1$ and $\mathbf{g}_5$ distributed $\mathrm{Bern}(0.5)$ and $\mathbf{g}_2$, $\mathbf{g}_3$, and $\mathbf{g}_4$ distributed $N(0, 1)$. We fix the five coefficient values in $\boldsymbol{\beta}$ to 0.5, 0.5, 1, 1, and 0.25, respectively for the five covariates in $\mathbf{x}_i$. Finally, we include seasonality across the time points via the term $\sin(t)$.

We use the covariate-adjusted principal score (CAPS) estimands from Section 3.2.4, the estimated coefficient from the principal score weighted linear mixed (PS-WLM) model in Section 3.3.1, and the posterior mean from the Bayesian mixture (BM) model in Section 3.3.2 to estimate average causal effects on a simulated sample. For each dataset and method we fit two models: one for all six time points and one for just the third and fourth time points (when both treatment assignments are present). We measure uncertainty in the CAPS method using 95% bootstrap confidence intervals from 1,000 samples and apply the method to all six time points using the strategy described in the final paragraph of Section 3.2.4. For the BM model we use the quantiles of posterior samples to contruct 95% credible intervals.

In order to measure model performance beyond the primary setting, we use three variations on the simulation to show the sensitivity of each model to increases in misspecification of $\mathbf{X}$, in participant-error variance σ_ν^2, and in data missingness. For the basic simulation scenario, we use a participant error variance of $\sigma_\nu^2 = 0.25$, no missingness, and no misspecification in $\mathbf{X}$. To create misspecification of $\mathbf{X}$, we remove the covariate $\mathbf{g}_5$, both covariates $\mathbf{g}_5$ and $\mathbf{g}_4$, or the three covariates $\mathbf{g}_5$, $\mathbf{g}_4$, and $\mathbf{g}_3$ from $\mathbf{X}$ to perform the model fits. For example, we remove $\mathbf{g}_5$ and the models are provided with four covariates in $\mathbf{X}$ when five covariates were used to in the outcome generating function. In the missingness study we remove a proportion of all observations from each sample at random. For every scenario in the studies we create 500 replicate samples of 300 participants each. To measure model performance for a certain average causal effect (CACE, NACE, or AACE), we use root mean square error (RMSE), bias, power, and coverage.

3.4.2 Results

Basic Setting

Boxplots of model estimates from 500 replicate samples under the basic settings of no missingness, no misspecification, and $\sigma_\nu^2 = 0.25$ are depicted in Figure 3.1, with the true complier average causal effect -0.5 marked by a dotted line. The RMSE, bias, power, and coverage of the simulation fits are reported in Table 3.1. In addition to the Wald 95% confidence intervals used to

Table 3.1: Model fitting results of 500 simulated samples under the basic setting (no missingness, no misspecification, static random effect variance σ_ν^2). The models used are a Bayesian mixture (BM) model, principal score weighted linear mixed (PS-WLM) model, and covariate-adjusted principal score (CAPS) method.

σ_ν^2	Model	Time Points Used	RMSE	Bias	Power	Coverage
0.25	BM	123456	0.137	-0.019	0.236	1.000
0.25	BM	34	0.178	0.036	0.014	1.000
0.25	PS-WLM	123456	0.115	0.005	0.992	0.940
0.25	PS-WLM	34	0.125	0.007	0.986	0.936
0.25	CAPS	123456	0.922	0.909	0.734	0.000
0.25	CAPS	34	0.128	0.010	0.960	0.956

form Table 3.1, we construct parametric bootstrap 95% confidence intervals and find that the two methods differ less than a percent in overall coverage and power for the 500 simulated samples. The PS-WLM model on the full dataset (all time points) has the lowest RMSE and highest power of any model. Both the PS-WLM and BM models produce lower RMSE and higher power on the full dataset than on just phases 3 and 4. The CAPS method for the full dataset is strongly biased from the seasonality term $\sin(t)$, but the method remains competitive with the other two models on the subset to phases 3 and 4. The BM model has only slightly larger bias and RMSE than the PS-WLM model, but has very low power and perfect coverage, indicating its credible intervals (quantiles of the posterior samples) are very large. Poor sampling of the posterior distribution may be due to small effect size, low sample size, and/or non-informative priors.

Increasing Participant Error Variance

We next test model tolerances to the size of random effect variance by simulating with increasing values of σ_ν^2. The results of this study are listed in Table 3.2, with the $\sigma_\nu^2 = 0.25$ case previously listed in Table 3.1. The PS-WLM model has the lowest RMSE and highest power of any model for each level of σ_ν^2, and the BM model is second best at each level. RMSE tends to increase as σ_ν^2 increases for all the models on the subset to phases 3 and 4, while those on the full dataset remain nearly unchanged. Similarly, power for each model decreases as σ_ν^2 increases, but more dramatically for the those on only two time points.

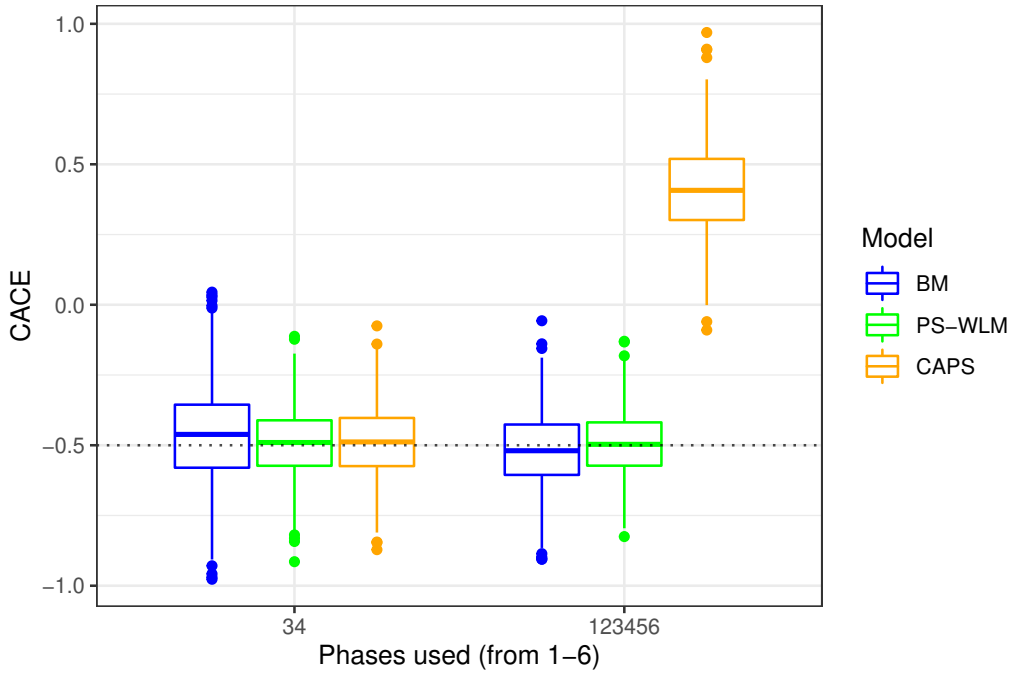

Figure 3.1: Boxplots of estimates for the complier average causal effect (CACE) in 500 replicate samples from the basic simulation setting. The models used are a Bayesian mixture (BM) model, principal score weighted linear mixed (PS-WLM) model, and covariate-adjusted principal score (CAPS) method.

Increasing Misspecification of X

For a study on misspecification of $\mathbf{X}$, we provide only two, three, or four covariates in $\mathbf{X}$ to the models being fit when in truth all five covariates were used to generate the outcome. The results of this study are reported in Table 3.3, and the results when correctly specified in $\mathbf{X}$ are reported in Table 3.1. As we provide fewer covariates in $\mathbf{X}$ to the models, power decreases and RMSE increases for every method, while bias and coverage do not follow any obvious trends. The models using the full dataset do not increase in RMSE as much as those only using phases 3 and 4, and the power of the PS-WLM model with all six phases decreases very little compared to any other method.

Increasing Missingness at Random

To test robustness of each model to missingness completely at random, we remove at random increasing proportions of data from the replicate samples. The results of this study are reported in Table 3.4 and the results when no data is missing are reported in Table 3.1. As missingness

Table 3.2: Model fitting results of 500 simulated samples with increasing participant error variance σ_ν^2. The models used are a Bayesian mixture (BM) model, principal score weighted linear mixed (PS-WLM) model, and covariate-adjusted principal score (CAPS) method.

σ_ν^2	Model	Time Points Used	RMSE	Bias	Power	Coverage
1.0	BM	123456	0.141	-0.022	0.216	1.000
1.0	BM	34	0.218	0.043	0.008	1.000
1.0	PS-WLM	123456	0.124	0.004	0.982	0.936
1.0	PS-WLM	34	0.168	0.010	0.836	0.932
1.0	CAPS	34	0.176	0.012	0.736	0.954
2.25	BM	123456	0.144	-0.022	0.204	1.000
2.25	BM	34	0.255	0.050	0.008	1.000
2.25	PS-WLM	123456	0.127	0.004	0.976	0.942
2.25	PS-WLM	34	0.220	0.014	0.598	0.942
2.25	CAPS	34	0.233	0.015	0.482	0.964
4.0	BM	123456	0.143	-0.023	0.190	1.000
4.0	BM	34	0.285	0.056	0.004	1.000
4.0	PS-WLM	123456	0.128	0.0033	0.970	0.940
4.0	PS-WLM	34	0.275	0.017	0.402	0.948
4.0	CAPS	34	0.295	0.017	0.304	0.966

Table 3.3: Model fitting results for simulation study with increasing misspecification of $\mathbf{X}$. The models used are a Bayesian mixture (BM) model, principal score weighted linear mixed (PS-WLM) model, and covariate-adjusted principal score (CAPS) method.

Covariates Provided	Model	Time Points Used	RMSE	Bias	Power	Coverage
g_1, g_2, g_3, and g_4	BM	123456	0.134	-0.023	0.226	1.000
	BM	34	0.173	0.042	0.018	1.000
	PS-WLM	123456	0.116	0.004	0.990	0.938
	PS-WLM	34	0.126	0.007	0.978	0.928
	CAPS	34	0.129	0.010	0.962	0.946
g_1, g_2, and g_3	BM	123456	0.144	-0.023	0.098	1.000
	BM	34	0.216	0.051	0.004	1.000
	PS-WLM	123456	0.124	0.003	0.978	0.944
	PS-WLM	34	0.185	0.006	0.764	0.940
	CAPS	34	0.191	0.008	0.732	0.952
g_1 and g_2	BM	123456	0.155	-0.017	0.032	1.000
	BM	34	0.247	0.057	0.004	1.000
	PS-WLM	123456	0.127	0.001	0.970	0.944
	PS-WLM	34	0.232	-0.002	0.622	0.934
	CAPS	34	0.240	-0.002	0.590	0.936

Table 3.4: Model fitting results for simulation study with increasing missingness at random. The models used are a Bayesian mixture (BM) model, principal score weighted linear mixed (PS-WLM) model, and covariate-adjusted principal score (CAPS) method.

Proportion Missing	Model	Time Points Used	RMSE	Bias	Power	Coverage
0.2	BM	123456	0.145	-0.026	0.156	1.000
0.2	BM	34	0.197	0.042	0.016	1.000
0.2	PS-WLM	123456	0.126	0.003	0.982	0.936
0.2	PS-WLM	34	0.137	0.004	0.958	0.932
0.2	CAPS	34	0.144	0.009	0.878	0.958
0.4	BM	123456	0.162	-0.042	0.118	1.000
0.4	BM	34	0.199	0.043	0.008	1.000
0.4	PS-WLM	123456	0.140	0.006	0.930	0.958
0.4	PS-WLM	34	0.150	0.008	0.886	0.940
0.4	CAPS	34	0.159	0.017	0.694	0.974
0.6	BM	123456	0.194	-0.056	0.062	1.000
0.6	BM	34	0.241	0.063	0.008	0.998
0.6	PS-WLM	123456	0.173	0.015	0.802	0.952
0.6	PS-WLM	34	0.196	0.013	0.736	0.936
0.6	CAPS	34	0.202	0.031	0.338	0.990

at random increases, RMSE and bias increase, both with a roughly constant rate for all models. At the same time, power decreases, although not as much for the PS-WLM models than for the others. Increasing coverage and decreasing power for the CAPS method may indicate its bootstrap confidence intervals are lengthening as the missingness increases.

3.5 Additional Simulation with Latent Strata Dependent on Covariates $\mathbf{x}_i$

3.5.1 Setup

We perform an additional simulation study, altering the construction of the latent strata U_{it} to depend on two covariates in $\mathbf{x}_i$. In Section 3.4.1, we drew the intial quantities of members in each stratum from a multinomial distribution with probabilities 2/3, 1/6, and 1/6 for compliers always-takers, and never-takers respectively. According to these generated quantities, we designated subjects in the first time point to the different strata at random. As a result, the strata U_{it} and

the treatment use S_{it} were each independent of the subjects' covariates $\mathbf{x}_i$. Each of the models presented in Sections 3.2.4, 3.3.1, and 3.3.2 is specified to account for a relationship between $\mathbf{X}$ and U_{it}, via the principal scores (CAPS and PS-WLM) or as part of the mixture model (BM). In this section, we create simulated samples to test model performance when the latent principal strata are dependent $\mathbf{X}$, and to test sensitivity to misspecification of the covariates $\mathbf{X}$ in the model fits under these conditions.

To construct the principal strata U_{it} with dependency on $\mathbf{X}$, we draw from a multinomial distribution by subject for each time point, with the probabilities of membership to the different strata determined by the covariates $\mathbf{g}_1$ and $\mathbf{g}_4$. Specifically, the probability of being a complier at-the-time for subject i is $p_{c,i} = \text{expit}_{(0.1,\,0.7)}(\gamma_0 + \gamma \mathrm{g}_{4,i}) + 0.2\mathrm{g}_{1,i}$, where $\text{expit}_{(0.1,\,0.7)}()$ is the expit function ($\text{expit}(a) = 1/(1 + e^{-(a)})$) truncated so that any value lower than 0.1 is replaced with 0.1 and any value higher than 0.7 replaced with 0.7. We fix the values $\gamma_0 = \text{logit}(0.4)$ and $\gamma = \frac{1}{3}(\text{logit}(0.7) - \gamma_0)$, where the logit function is defined $\text{logit}(p) = \ln(p/(1-p))$. With these fixed values for γ_0 and γ, the range of -3 to 3 for the standard normally distributed covariate $\mathbf{x}_4$ is converted by the typical expit function to the the range of 0.1 and 0.7. Then by the empirical rule, 99% of the values which $\mathbf{x}_4$ may take on will fall in the range 0.1 to 0.7, and the rest are truncated by the $\text{expit}_{(0.1,\,0.7)}$ function. The probabilities of being a never-taker or always-taker at-the-time are split equally from $1 - p_{c,i}$, so that $p_{n,i} = p_{a,i} = (1 - p_{c,i})/2$. Thus, we draw the stratum U_{it} for subject i from the same multinomial distribution for all six time points, defined by the probabilities $p_{c,i}$ for compliers, $p_{n,i}$ for never-takers, and $p_{a,i}$ for always-takers.

We create simulated samples with the same model and parameter values as in Section 3.4.1, only modifying the method of generating the latent strata U_{it} which determine treatment use S_{it}. We apply the same three models (CAPS, PS-WLM, and BM) on the same data subsets (all six time points or just time points 3 and 4). To test sensitivity of the models to misspecification in $\mathbf{X}$, we again remove the covariate $\mathbf{g}_5$, both covariates $\mathbf{g}_5$ and $\mathbf{g}_4$, or the three covariates $\mathbf{g}_5$, $\mathbf{g}_4$, and $\mathbf{g}_3$ from $\mathbf{X}$ to perform the model fits.

3.5.2 Results

We report the results from the additional simulation study in Table 3.5. The first row of model results in the table ($\mathbf{g}_1$, $\mathbf{g}_2$, $\mathbf{g}_3$, $\mathbf{g}_4$, and $\mathbf{g}_5$) corresponds to the basic setting simulated samples and results in Table 3.1. When we have added dependency between U_{it} and $\mathbf{X}$ in the basic setting, we find that the RMSEs for all models are lower except for the BM model on the subset of time points 3 and 4. Same as in Section 3.4.2, when we remove the binary fifth covariate $\mathbf{g}_5$, the RMSE of the BM model on the subset of only time points 3 and 4 decreases and all model RMSEs remain nearly unchanged. When we remove $\mathbf{g}_5$ and the normally distributed fourth covariate $\mathbf{g}_4$ (which we used along with $\mathbf{g}_1$ to generate U_{it}), the RMSEs of all models increase, especially those using the subset of only time points 3 and 4. In the final row of model results ($\mathbf{g}_1$ and $\mathbf{g}_2$), all model RMSEs again increase, although the RMSE of the PS-WLM model on all time points increases by only 0.03 (2.5%). The power of each model in the stages of misspecification of $\mathbf{X}$ corresponds with the trends in RMSE, i.e. power decreases when RMSE increases and vice versa. Power and RMSE across most models and stages of misspecification when the samples have dependent principal strata on $\mathbf{X}$ are better (power is higher, RMSE is lower) than when $\mathbf{X}$ and U_{it} are independent. Bias and coverage of the models do not show obvious trends in Table 3.5.

In general, the model results in Table 3.5 have slightly lower RMSE and slightly higher power than the corresponding results in Tables 3.1 and 3.3. There is a relatively large increase in RMSE and decrease in power when we remove both covariates $\mathbf{g}_5$ and $\mathbf{g}_4$ compared with only removing $\mathbf{g}_5$, but there is a similarly large reduction in model performance in Table 3.3 as well, likely due to the fact that $\mathbf{g}_5$ is a binary covariate and has less of an effect on the outcome than $\mathbf{g}_4$, which comes from a standard normal distribution. The model RMSEs in Table 3.5 drop slightly more than in Table 3.3 across the two scenarios, indicating some impact of misspecifying the relationship between $\mathbf{X}$ and U_{it}. Overall, this additional simulation study shows that each model can account and even benefit from dependency of the covariates $\mathbf{X}$ on U_{it}, and misspecification of this dependency only leads to relatively small drops in the model performances.

Table 3.5: Model fitting results for additional simulation study with latent principal strata that are dependent on the covariates $\mathbf{g}_1$ and $\mathbf{g}_4$. We provide only the listed covariates (of five) in the first column to each model fit, increasing the misspecification of $\mathbf{X}$ by removing covariates. We fit the Bayesian mixture (BM) model, principal score weighted linear mixed (PS-WLM) model, and covariate-adjusted principal score (CAPS) method.

Covariates Provided	Model	Time Points Used	RMSE	Bias	Power	Coverage
$\mathbf{g}_1$, $\mathbf{g}_2$, $\mathbf{g}_3$, $\mathbf{g}_4$, and $\mathbf{g}_5$	BM	123456	0.128	-0.028	0.250	1.000
	BM	34	0.179	0.026	0.010	1.000
	PS-WLM	123456	0.111	0.002	0.996	0.944
	PS-WLM	34	0.117	0.005	0.992	0.950
	CAPS	34	0.120	0.005	0.984	0.952
$\mathbf{g}_1$, $\mathbf{g}_2$, $\mathbf{g}_3$, and $\mathbf{g}_4$	BM	123456	0.129	-0.031	0.266	1.000
	BM	34	0.171	0.023	0.010	1.000
	PS-WLM	123456	0.111	0.002	0.996	0.944
	PS-WLM	34	0.117	0.004	0.992	0.950
	CAPS	34	0.121	0.004	0.982	0.948
$\mathbf{g}_1$, $\mathbf{g}_2$, and $\mathbf{g}_3$	BM	123456	0.145	-0.036	0.134	1.000
	BM	34	0.223	0.036	0.008	1.000
	PS-WLM	123456	0.120	0.001	0.994	0.952
	PS-WLM	34	0.170	0.000	0.852	0.940
	CAPS	34	0.178	0.000	0.810	0.938
$\mathbf{g}_1$ and $\mathbf{g}_2$	BM	123456	0.157	-0.027	0.054	1.000
	BM	34	0.247	0.035	0.008	1.000
	PS-WLM	123456	0.123	0.000	0.992	0.942
	PS-WLM	34	0.212	-0.004	0.690	0.948
	CAPS	34	0.225	-0.003	0.644	0.944

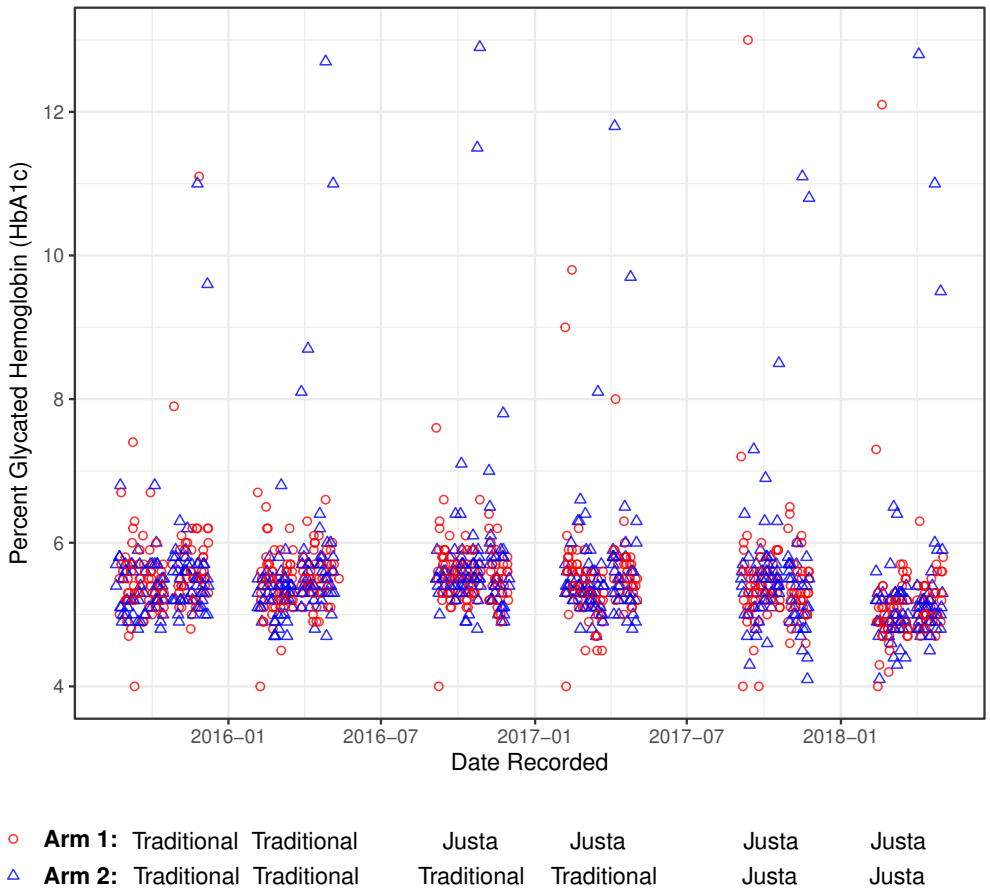

Figure 3.2: Percent glycated hemoglobin (HbA1c) of 230 primary cooks measured across six study phases, with the assigned stoves (Traditional or Justa) for study arms 1 and 2 listed below the plot. Using a stepped wedge design, primary cooks are randomized to a study arm and receive the Justa stove before phase 3 or phase 5.

3.6 Analysis of Honduran Stepped-Wedge Cookstove Trial

We now apply our principal stratification models to the study that motivated this work, a randomized controlled trial of a biomass burning cookstove intervention in rural Honduras (Young et al., 2019). Researchers replaced existing cookstoves with ventilated (Justa) biomass cookstoves in every household according to a stepped-wedge design. Stove use, health outcomes, and other covariates were collected for 230 primary cooks (all women) over 6 study visits in a span of 3 years (roughly every 6 months). Between the second and third study visits (phases), participants were randomly assigned to a study arm and the household either received the Justa cookstove before phase 3 (Arm 1) or before phase 5 (Arm 2), so that every household eventually received the

intervention, but in a staggered fashion (Figure 3.2). In our analysis, we will estimate the stratified average causal effects of stove assignment (traditional or Justa) on the outcome, percent of glycated hemoglobin (HbA1c) that is depicted in Figure 3.2.

3.6.1 Setup and Filtering

From the data, we take Z, the assigned cookstove at each phase, which was predetermined when the participant was assigned to a study arm. The outcome Y is HbA1c and to create S, the actual cookstove usage, we consider that a household i at time t may use the traditional stove ($S_{it} = 0$) or the improved (Justa) stove ($S_{it} = 1$). The existing traditional stoves in the households were destroyed when the Justa stove was installed, but the study did not prevent participants from making and using another traditional stove in addition to the newly installed Justa. In these cases, we define the usage of both stoves in a phase the same as when they were not receiving the intervention, i.e. $S_{it} = 0$. Here we note that an Exclusion Restriction (ER), the assumption in this context that $Y_{it}(z = 1, s) = Y_{it}(z = 0, s) = Y_{it}(s)$ for $s = 0$ or 1 (Angrist et al., 1996), is likely violated. Specifically, when under $z = 0$ a household does not have access to a Justa stove, but under $z = 1$ they do. Then the value $Y_{it}(z = 1, s = 0)$ may differ from the value $Y_{it}(z = 0, s = 0)$, since the household would use both the Justa and traditional stove in the $(z = 1, s = 0)$ case, but only the traditional stove(s) in the $(z = 0, s = 0)$ case. A violation of ER prevents the use of instrumental variable analysis in our example and indicates a need for models without this assumption, such as those discussed in Sections 3.2 and 3.3 (Greenland, 2000).

We filter the study observations to complete cases, removing any observations that are missing Z_{it} ($n = 28$), S_{it} ($n = 126$), Y_{it} ($n = 168$), or some part of $\mathbf{x}_i$ ($n = 6$) for $n = 175$ total removed and $n = 1,211$ remaining. Defining Z and S by stove assignment and stove use as we have above, we may categorize all study observations by their possible stratum membership(s) at-the-time. There are 6 at-the-time always-takers ($Z_{it} = 0$, $S_{it} = 1$) , 375 at-the-time never-takers ($Z_{it} = 1$, $S_{it} = 0$), 205 possible compliers or always-takers ($Z_{it} = 1$, $S_{it} = 1$), and 625 possible compliers or never-takers ($Z_{it} = 0$, $S_{it} = 0$).

3.6.2 Covariates Chosen/Sensitivity Checks

An important assumption of principal scores for principal stratification is that the covariates in $\mathbf{X}$ are balanced, specifically that Assumption 4 is satisified with respect to $\mathbf{X}$. For any function of the covariates $h(\mathbf{X})$, we can check for violations of the balancing conditions below, which follow from Assumption 4 and Equation (3.2):

$$\begin{aligned} E\left\{w_{1,c}(\mathbf{x}_i)h(\mathbf{x}_i) \mid Z_{it}=1, S_{it}=1\right\} &= E\left\{w_{0,c}(\mathbf{x}_i)h(\mathbf{x}_i) \mid Z_{it}=0, S_{it}=0\right\}, \\ E\{h(\mathbf{x}_i) \mid Z_{it}=1, S_{it}=0\} &= E\left\{w_{0,n}(\mathbf{x}_i)h(\mathbf{x}_i) \mid Z_{it}=0, S_{it}=0\right\}, \\ E\left\{w_{1,a}(\mathbf{x}_i)h(\mathbf{x}_i) \mid Z_{it}=1, S_{it}=1\right\} &= E\{h(\mathbf{x}_i) \mid Z_{it}=0, S_{it}=1\}. \end{aligned} \tag{3.5}$$

We estimate the average differences for each of the balancing conditions and use bootstrapped standard errors to produce standardized t-statistics for each covariate. We select the binary pre-randomization covariates of whether or not the primary cook is over 40 years of age, has a BMI over 25, has metabolic syndrome, or has electricity available in their household, and also the continuous pre-randomization covariate of peripheral systolic blood pressure. The standardized t-statistic for untransformed peripheral systolic blood pressure is large ($> |2|$), so we apply a natural log transformation and find that all covariates are then balanced. We report the values used to establish covariance balance in Table C.2 in the Appendix.

3.6.3 Results

We apply the covariate-adjusted estimands from Section 3.2.4 and the principal score weighted linear mixed model from Section 3.3.1 to estimate average causal effects from the Honduran cook-stove study data. We fit the PS-WLM model on both the full dataset and the subset of just the third and fourth phases, in each case including the indicator variable(s) for phase in the outcome model (the fixed effects by time point α_t). We fit the CAPS estimands to just the subset of phases 3 and 4, without accounting for phase at all and calculating confidence intervals as the quantiles of 1,000 bootstrap samples. We exclude the Bayesian mixture model from the analysis due to concerns with

Table 3.6: Estimates and confidence intervals of the principal score weighted linear mixed (PS-WLM) and covariate-adjusted principal score (CAPS) methods on the Honduran cookstove study data.

				CI Bounds	
ACE	Model	Phases Used	Estimate	Lower	Upper
CACE	PS-WLM	123456	-0.150	-0.219	-0.082
	PS-WLM	34	-0.171	-0.430	0.088
	CAPS	34	-0.121	-0.750	0.306
NACE	PS-WLM	123456	-0.114	-0.180	-0.047
	PS-WLM	34	-0.070	-0.295	0.155
	CAPS	34	-0.082	-0.404	0.590
AACE	PS-WLM	123456	-0.419	-1.284	0.447
	PS-WLM	34	-0.233	-0.516	0.050
	CAPS	34	0.222	-0.339	1.027

correctly sampling the posterior when it appears that the effect magnitude and sample size are even smaller than those used in the simulations (-0.5 and 300 respectively). Similarly, we do not fit the CAPS estimands on the full dataset, as it would likely result in strong bias.

The estimates and confidence intervals of the models for each average causal effect are listed in Table 3.6. All estimates of the complier average causal effect are negative, indicating a reduction in HbA1c for compliers when assigned the Justa stove. Each of the never-taker and two of the three always-taker estimates are also negative. Only the complier and never-taker average causal effect estimates from the PS-WLM on all six time points are significant. A significant never-taker effect is not surprising when participants may be receiving partial treatment (using both Justa stove and traditional stove) but considered untreated (see Section 4.6.1). The point estimates of all methods are relatively close for the complier and never-taker average causal effects. The always-taker effect estimates are more variable, likely due to the small amount of possible always-taker observations in the study (17% of observations)

3.7 Discussion

Non-compliance to treatment assignment is a prevalent issue in randomized controlled trials measuring the effect of indoor cookstove-related air pollution on respiratory or cardiometabolic

health. To combat attenuation in effect estimates from mixed compliance, we have turned to the method of principal stratification and aim to estimate stratum-specific average causal effects, instead of the overall intent-to-treat effect. Cookstove intervention studies such as the one used in Section 3.6 often rely on repeated measures and may assign the improved cookstove intervention to both study arms, but in a staggered manner. We applied the existing method of principal stratification using principal scores developed by Ding and Lu (2017) in a longitudinal setting, but faced bias from temporal trends when ignoring the repeated measures structure. We introduced two extensions of principal stratification to a longitudinal setting that each include random effects. We used principal scores to weight a linear mixed model on the data subsetted to only possible members of a certain stratum. The estimated coefficient for assignment to the improved stove in this model is then an estimate of the average causal effect for that stratum. In the second method, we used a Bayesian mixture model to estimate averages effects of differing assignments with respect to the latent strata.

We generated simulated samples in three sensitivity analyses and under an additional simulation model to test the performance of the covariate-adjusted principal score (CAPS) estimands from Ding and Lu (2017), the principal score weighted linear mixed (PS-WLM) model, and the Bayesian mixture (BM) model. In every setting, the PS-WLM model on the full data set produced the lowest RMSE and highest power. The BM and PS-WLM models using the full data set (all six time points) were more robust than any model using only time points 3 and 4 to increasing random intercept variance and to the removal of covariates in $\mathbf{X}$ from the fits, while uniform random missingness worsened the fit of every model similarly. When including dependence between $\mathbf{X}$ and U_{it} in the simulation model, the fits of each model were slightly improved and removal of the covariate $\mathbf{g}_4$ (which was affecting U_{it}) led to a relatively large decrease in model performances. The BM model had competitive RMSE and bias across all simulation settings, but yielded credible intervals that were too large and had low power and near-perfect coverage as a result. Uncertainty in the posterior distribution may be carried over from uncertainty in the imputation of the latent strata, while the frequentist principal scores methods translate possible stratum membership through con-

tinuously valued weights. In our motivating application (Section 3.6), the PS-WLM model for the full data set found the complier and never-taker average causal effects to be significant, and no other model found significant effects.

It is clear from the performance of the PS-WLM model that the principal scores weights in Equation (3.1) are effective in providing useful information on stratum membership with regard to the covariates $\mathbf{X}$. The weights for a subject do not change in time, however, as they are formulated using only pre-randomization covariates in $\mathbf{X}$. We also cannot calculate the weights across subjects for each time point separately, since in a stepped-wedge design, subjects all receive the same assignment for the starting time points and ending time points. The probabilities and principal scores derived in Section 3.2.4 would be too numerically extreme to allow calculation of the weights in Equation (3.1). Thus, extensions to time-varying or cumulative principal scores weights would require reconsideration of the weight formulas themselves.

An important characteristic to each of the models we have developed is that they estimate average causal effects for "at-the-time" principal strata, i.e., the average effects from assignment when belonging to a certain principal stratum for the current period of time. These results translate more directly to the efficacy of an intervention than those based off, for example, the percentage of compliance behavior of a subject over time. Specifically, the intervention's effect in the same time period it is taken can be measured by the complier average causal effect, or the difference in outcome due to assignment of treatment when the subject would comply with their assignment. This ease of interpretation, however, does require the additional assumption. Both the PS-WLM and the Bayesian model require the strong assumption that the effect of intervention is restricted to the same time point in which it occurred. In other words, neither model accounts for a carryover effect of intervention to the following time point. Clearly there are many situations where this assumption would be violated, and the resulting average causal effect estimates from either model would be inaccurate and sensitive to the ordering of the study design. We can account for lagged effects in our model, possibly following the example of Lin et al. (2009), but would lose the simple interpretation of "at-the-time" compliance. Instead of a single effect of intervention, we would be

estimating an accumulation over lags, which adds considerable complication. Moreover, the task of establishing the presence of lagged causal effects in a dataset is not trivial. Simply including lag terms (such as $S_{i(t-1)}$) to the weighted linear mixed model introduces the trouble of temporal confounding, since in the stepped-wedge design $S_{i(t-1)} = 1$ is more likely at later time points.

Implementation of lagged effects in model structure is a potential future work for longitudinal principal stratification methods. The PS-WLM model could include a random effects structure of higher complexity, and already has the advantage of much lower computational time than the methods reliant on posterior sampling or bootstrapping. Both the PS-WLM and Bayesian models presented in this study provide relevant extensions of principal stratification to a longitudinal setting, and are forward steps in handling the issue of mixed compliance in enviromental intervention trials.

Chapter 4

Principal Stratification Defined by a Continuous Exposure in a Longitudinal Setting

4.1 Introduction

In this chapter we are once again interested in assessing the causal effect of an environmental intervention on measures of health. We approach this problem using principal stratification methods, but now we consider principal strata that are based on a continuous intermediate variable. This work is motivated by the study of an improved cookstove intervention and the health outcomes of Honduran woman in the longitudinal trial introduced in Chapter 3 (Young et al., 2019). The underlying rationale for the intervention is that we expect a ventilated stove to reduce indoor air pollution compared to an open cooking fire, and this pollution reduction to then improve a subject's health (Figure 4.1). Previously in Chapter 3, we used a subject's assignment to the improved or traditional cookstove to define potential outcomes (potential value under assignment of treatment or control, respectively). The subject's actual stove usage, revealing their compliance to treatment assignment, was the binary post-treatment variable we used to form principal strata. While stove use is highly correlated with exposure in this study (Witinok-Huber et al., 2022), it does not fully capture exposure differences between and within subjects. It is possible that exposure levels may not decrease, or only decrease by a small amount, when using the improved stove. Furthermore, from an epidemiological perspective, it is important to know the specific relationship between indoor exposure to particulate matter (PM) and health outcomes. The total effect of the stove intervention on the health outcome, may also include a direct effect from the stove assignment itself. In this chapter, we consider continuous exposure to indoor particulate matter (PM) as the continuously-valued post-treatment variable defining principal strata.

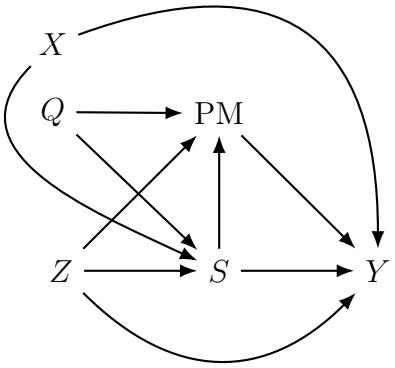

Figure 4.1: Directed acyclic graph of the assignment (Z), treatment usage(S), exposure (PM), outcome (Y), outcome-related covariates (X) and exposure-related covariates (Q) of the study setting. The covariate sets X and Q may be comprised of some of the same covariates.

We consider two different ways of estimating principal causal effects using longitudinal exposures. Firstly, by categorizing exposure simply as lowered or not, we can apply the PS-WLM model developed in Section 3.3.1. Second, we extend a Bayesian approach from Hackstadt et al. (2014) that also estimates causal effects dependent on a threshold L for exposure reduction, but does not require conversion of continuous exposure values to binary. We apply this approach to a longitudinal setting while accounting for repeated measures.

The remainder of this chapter is outlined as follows. We first introduce the notation and setting in Section 4.2, then in Section 4.3, we extend the PS-WLM model from Section 3.3.1 to using continuous post-treatment variable for defining the principal strata. We introduce and extend the Bayesian approach from Hackstadt et al. (2014) to a longitudinal setting in Section 4.4. In Section 4.5 we perform a simulation study to compare model performances on simulated samples, and in Section 4.6 we analyze the cookstove intervention study from Young et al. (2019) with the Bayesian approach. Finally, we provide a discussion in Section 4.7.

4.2 Notation and Setting

We adopt the same notation and setting as in Section 3.2.1, with the addition of the post-treatment exposure variable PM. Consider N subjects measured over T time points with a stepped-

wedge crossover design, where both study arms begin with control and are eventually assigned treatment, but one arm receives treatment before the other (Brown and Lilford, 2006). For subject $i = 1, \ldots, N$ at time $t = 1, \ldots, T$, we denote Z_{it} as assignment to treatment ($Z_{it} = 1$) or control ($Z_{it} = 0$) and the vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ as the predetermined vectors of assignment for subjects in study arm 1 and study arm 2, respectively. When a subject i actually uses the treatment at time t we denote $S_{it} = 1$ and $S_{it} = 0$ if they use control. The observed outcome value we write as Y_{it} and the exposure as PM_{it}. We consider two sets of possibly overlapping sets of pre-randomization covariates: $\mathbf{x}_i$ are related to the outcome and $\mathbf{q}_i$ are related to the exposure (PM). Both $\mathbf{x}_i$ and $\mathbf{q}_i$ may be related to the latent principal strata U_{it} and thereby to the actual treatment use S_{it}. We depict a directed acyclic graph including all the observed variables mentioned above in Figure 4.1. Note that both the treatment assignment Z and actual treatment use S may have direct effects on the outcome Y, such as in an unblinded study.

We define the potential outcome values under control and under treatment as $Y_{it}(0)$ and $Y_{it}(1)$, respectively. Similarly, $S_{it}(0)$ and $S_{it}(1)$ are the potential treatment-use values and $\mathrm{PM}_{it}(0)$ and $\mathrm{PM}_{it}(1)$ are the potential exposure values under control and treatment, respectively. Finally, the vectors $\mathbf{Y}_i(\mathbf{z}_k)$, $\mathbf{S}_i(\mathbf{z}_k)$, and $\mathbf{PM}_i(\mathbf{z}_k)$ are the sets of potential outcome values, potential actual treatments, and potential exposure values used for all timepoints of subject i under the treatment patterns $k = 1$ or 2.

4.3 Principal Score Weighted Linear Mixed Model with Compliance Defined by a Continuous Exposure

4.3.1 Conversion of Continuous Exposure to Binary Variable

To estimate the stratified causal relationships between exposure and outcome, we would like to implement the model from Section 3.3.1. Previously in the PS-WLM framework, a subject was a complier "at-the-time" when their potential stove usage matched their assigned stove, i.e. that $S_{it}(1) = 1$ and $S_{it}(0) = 0$. Always-takers had $S_{it}(1) = S_{it}(0) = 1$, never-takers had $S_{it}(1) =$

$S_{it}(0) = 0$, and defiers had $S_{it}(1) = 0$ and $S_{it}(0) = 1$. As there are infinite potential values that the exposure PM may take, there are infinite principal strata that could be defined by it. To use the PS-WLM model with only four possible latent strata, we first convert the continous exposures PM_{it} to a binary measure of "lowered" or "not lowered" exposure with respect to a numerical threshold L. The resulting variable PM_{it}^{L} is equal to 1 when $\mathrm{PM}_{it} < L$ and is equal to 0 otherwise.

The variable PM_{it}^{L} may not reflect all observations that are causally lowered due to treatment assignment, as unmeasured covariates and exogenous temporal factors may also be affecting the exposure. To control for some of the variability in PM_{it} unexplained by the treatment, we fit a preliminary model:

$$\mathrm{PM}_{it} = \alpha_t + \mathbf{q}_{it}^{\top}\boldsymbol{\alpha} + \beta_S S_{it} + \eta_i + \epsilon'_{it}, \tag{4.1}$$

where $\mathbf{q}_{it}$ are relevant covariates for PM_{it}, which do not have to be measured pre-randomization, and $\boldsymbol{\alpha}$ are their model coefficients. We assume $\epsilon'_{it} \overset{iid}{\sim} N(0, \sigma')$ and $\eta_i \overset{iid}{\sim} N(0, \sigma_\eta)$ is a random intercept by subject. The fixed effect α_t is meant to capture variation by each time point t that is not captured in $\mathbf{q}_{it}$. Since we control for treatment assignment S_{it}, the estimated coefficients for each time point $\widehat{\alpha}_t$, for the model coefficients $\widehat{\boldsymbol{\alpha}}$, and participant-specific random intercepts $\widehat{\eta}_i$ are explaining variability that is not due to treatment. Then the Best Linear Unbiased Predictor (BLUP) for subject i at time t under control ($S_{it} = 0$) is $\widehat{\alpha}_t + \mathbf{q}_{it}^{\top}\widehat{\boldsymbol{\alpha}} + \widehat{\eta}_i$. We create a set of residual exposures by removing the predictions under control from each observation $\left(\widehat{\mathrm{PM}}_{it}^{adj} = \mathrm{PM}_{it} - (\widehat{\alpha}_t + \mathbf{q}_{it}^{\top}\widehat{\boldsymbol{\alpha}} + \widehat{\eta}_i)\right)$, so that we are left with a mixture of variation which is explained by use of treatment and variation which is not explained by the time point and participant-specific effects. We then convert the residual exposure values to a binary variable PM_{it}^{L}, that signifies when a subject's exposure is lowered ($\widehat{\mathrm{PM}}_{it}^{adj} < L \rightarrow \mathrm{PM}_{it}^{L} = 1$) or not lowered ($\widehat{\mathrm{PM}}_{it}^{adj} \geq L \rightarrow \mathrm{PM}_{it}^{L} = 0$).

4.3.2 Model Fitting

Once we have defined a binary exposure variable, we can apply the PS-WLM model from Section 3.3.1 with the only modification that in the place of actual treatment use S_{it} we use PM_{it}^{L}.

In other words, $Z_{it} = 1$ means subject i is assigned to have a lowered exposure at time t and PM_{it}^{L} reveals whether they complied to this assignment. When using PM_{it}^{L} as the post-treatment to define principal strata, we would consider subject i at time t to be a complier at-the-time if $\mathrm{PM}_{it}^{L}(1) = 1$ and $\mathrm{PM}_{it}^{L}(0) = 0$, a never-taker if $\mathrm{PM}_{it}^{L}(1) = \mathrm{PM}_{it}^{L}(0) = 0$, an always-taker if $\mathrm{PM}_{it}^{L}(1) = \mathrm{PM}_{it}^{L}(0) = 1$, and a defier if $\mathrm{PM}_{it}^{L}(1) = 0$ and $\mathrm{PM}_{it}^{L}(0) = 1$.

The assumptions listed in Section 3.2.3 must still apply. Each is untestable, but we specifically address the assumption of monotonicity (Assumption 3), which requires that the data do not contain members of the defier stratum. Using the assumption of General Principal Ignorability (GPI), the potential exposures of a defier would only occur in response to the treatment assignment, not due to some other factor. Then our assumption of monotonicity in this context is that there are no time points when a subject would have lowered exposure because they are assigned the control and have raised exposure because they are assigned treatment. We consider this behavior, responding to treatment with higher exposure than to control, to be unrealistic and we assume monotonicity for the PS-WLM model as a result.

While an attractive extension of methods developed in Section 3.3.1, this approach has potential drawbacks. The conversion of a continuous valued exposure to a binary variable by some threshold is a possible source of bias. Unexplained variation in the exposure measurements could mean that even at a time when a subject's exposure is causally lowered by the treatment, their observed exposure measurement may still be higher than observed measurements under control. Even if this variation is non-differential, the distribution of all exposures lowered causally by treatment and distribution of all exposures not casually lowered by treatment can overlap. Then any reasonable threshold L would lie in the ranges of both distributions such that the higher measurements of causally lowered exposures were above L, thus mislabelled. The resulting binary variable then identifies only casually lowered exposures which are low with respect to their full distribution. These right-truncated exposures, correspond to outcome values influenced only by relatively low exposures and lead to estimates that are biased to the left.

4.4 Extension of the Bayesian Approach in Hackstadt et al. (2014) to a Longitudinal Setting

As an alternative to the PS-WLM method, we consider an extension to a Bayesian approach developed by Hackstadt et al. (2014), involving the estimation of all potential values for the exposure and outcome. In this framework we first estimate the counterfactual exposures ($\mathrm{PM}_{it}(0)$, $\mathrm{PM}_{it}(1)$) and counterfactual outcome values ($Y_{it}(0)$, $Y_{it}(1)$), then compute the difference in potential exposure values $\mathrm{PM}_{it}(1) - \mathrm{PM}_{it}(0)$ for every subject i and time t. On these differences we use a threshold value L to define strata and estimate average causal effects from the differences in potential outcomes ($Y_{it}(1) - Y_{it}(0)$) in each strata.

4.4.1 Assumed Model

For this approach, we make Assumptions 1 (SUTVA) and 2 (Randomization) from Section 3.2.3. We do not make Assumption 3 and allow subjects to be defiers at-the-time, although the principal strata in this model are not the same as in Chapter 3. We replace Assumption 4 with the assumption of "ignorability of treatment", so that conditional on the observed covariates, there is no unmeasured confounding between the treatment and potential values for PM and Y.

We assume the following models for potential exposures and potential outcomes that are each taken from Hackstadt et al. (2014) except now with the addition of a random intercept ν_i or η_i and the index of time t. For the potential outcomes $Y_{it}(0)$ and $Y_{it}(1)$ we use the model:

$$\begin{bmatrix} Y_{it}(0) \\ Y_{it}(1) \end{bmatrix} \sim N\left(\begin{bmatrix} \alpha_t + \mathbf{x}_i^\top \boldsymbol{\beta} + \beta_{\mathrm{PM}} \mathrm{PM}_{it}(0) + \nu_i \\ \alpha_t + \mathbf{x}_i^\top \boldsymbol{\beta} + \beta_z + (\beta_{\mathrm{PM}} + \beta_{z\times\mathrm{PM}}) \mathrm{PM}_{it}(1) + \nu_i \end{bmatrix}, \begin{bmatrix} \xi^2 & 0 \\ 0 & \xi^2 \end{bmatrix} \right). \quad (4.2)$$

And for the potential exposures $\mathrm{PM}_{it}(0)$ and $\mathrm{PM}_{it}(1)$ we use the model:

$$\begin{bmatrix} \mathrm{PM}_{it}(0) \\ \mathrm{PM}_{it}(1) \end{bmatrix} \sim N\left(\begin{bmatrix} \alpha_t + \mathbf{q}_i^\top \boldsymbol{\gamma} + \eta_i \\ \alpha_t + \mathbf{q}_i^\top \boldsymbol{\gamma} + \eta_i + \delta_1 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{bmatrix} \right). \quad (4.3)$$

The p covariates $\mathbf{x}_i$ are pre-randomization predictors of the outcome Y with coefficients $\boldsymbol{\beta}$, and the r covariates $\mathbf{q}_i$ are pre-randomization predictors of the exposure PM with coefficients $\boldsymbol{\gamma}$. The fixed effect α_t incorporates a temporal trend by time point. We consider the effect of treatment assignment on the exposure PM as δ_1, while the relationship between assignment Z and outcome Y we assume to have a direct effect β_z and an effect modified by PM: $(\beta_{\mathrm{PM}} + \beta_{z\times\mathrm{PM}})\,\mathrm{PM}_{it}$. The random intercepts by subject are normally distributed so that $\nu_i \overset{iid}{\sim} N(0, \sigma_\nu)$ and $\eta_i \overset{iid}{\sim} N(0, \sigma_\eta)$. Note that ρ is a fixed value for the unknown correlation between $\mathrm{PM}_{it}(0)$ and $\mathrm{PM}_{it}(1)$, which will never be observed simultaneously. Here we set $\rho = 0.1$.

4.4.2 Causal Effects Estimates

The post-treatment variable that we would use to define principal strata, the exposure PM, is continuously valued, meaning there are infinite latent strata that could be defined by the potential outcomes $\mathrm{PM}(0)$ and $\mathrm{PM}(1)$. To deal with this, Hackstadt et al. (2014) define causal effect estimates by grouping observations by their difference in potential exposure $\mathrm{PM}(1) - \mathrm{PM}(0)$, and we adopt the same strategy here. Dependent on a numerical threshold L we define a complier average causal effect,

$$\mathrm{CACE}_L = E\left\{Y_{it}(1) - Y_{it}(0) \mid \mathrm{PM}_{it}(1) - \mathrm{PM}_{it}(0) < L\right\}, \tag{4.4}$$

and thereby consider a participant to be a complier "at-the-time" for an observation where their exposure would reduce (with respect to L) when assigned the treatment.

Unlike in Chapter 3, we do not estimate always- and never-taker effects in this framework. An always-taker would be defined for $\mathrm{PM}_{it}(1) < L$ and $\mathrm{PM}_{it}(0) < L$ simultaneously, and a never-taker would have $\mathrm{PM}_{it}(1) \geq L$ and $\mathrm{PM}_{it}(0) \geq L$. Instead of these separate strata, we capture the effect of treatment assignment on all subjects whose potential exposure would be unchanged according to the threshold L at-the-time. The unchanged average causal effect (UACE) is:

$$\mathrm{UACE}_L = E\left\{Y_{it}(1) - Y_{it}(0) \mid |\mathrm{PM}_{it}(1) - \mathrm{PM}_{it}(0)| \leq |L|\right\}. \tag{4.5}$$

An always- or never-taker would fit in this grouping, having both raised or both lowered potential exposures level and resulting in no difference in the potential values.

Also unlike Chapter 3, we do not require the assumption of monotonicity for our extension of the method from Hackstadt et al. (2014). A difference in potential exposures that is positive and not within the threshold of "unchanged" ($\mathrm{PM}(1) - \mathrm{PM}(0) > |L|$) is behavior of a defier at-the-time. We denote the defier average causal effect (DACE) as:

$$\mathrm{DACE}_L = E\left\{Y_{it}(1) - Y_{it}(0) \mid \mathrm{PM}_{it}(1) - \mathrm{PM}_{it}(0) > |L|\right\}. \tag{4.6}$$

4.4.3 Bayesian Analysis and Priors

To estimate the causal effects, we need to impute all missing potential exposures and outcomes. For example, under treatment assignment ($Z_{it} = 1$), $\mathrm{PM}_{it}(0)$ and $Y_{it}(0)$ are unknown and must be estimated. In addition to the missing potential values, we must estimate the model parameters from Equations (4.2) and (4.3), including $\boldsymbol{\theta} = \left(\boldsymbol{\gamma}^\top, \sigma_0, \sigma_1, \delta_1, \sigma_\eta, \boldsymbol{\beta}^\top, \beta_z, \beta_{\mathrm{PM}}, \beta_{z\times\mathrm{PM}}, \xi, \sigma_\nu\right)$ and the vectors of random intercepts for all subjects, $\boldsymbol{\nu}$ and $\boldsymbol{\eta}$. The quantity of potential values and parameters to be estimated simultaneously motivates us to use a Bayesian approach. We estimate the average causal effects in (4.4), (4.5), and (4.6) as posterior means, using the probabilistic programming language Stan via the `R` package `rstan` to perform Hamiltonian Monte Carlo posterior sampling with the NUTS algorithm (Stan Development Team, 2023a,b).

We implement the same prior distributions as Hackstadt et al. (2014) and use the following hyperparameter values across all simulation and environmental intervention model fits. The independent error variance ξ^2 has an inverse gamma prior with shape $\alpha_\xi = 0.01$ and scale $\beta_\xi = 0.01$. For potential exposure standard deviations σ_0 and σ_1 we use zero-mean log-normal prior distributions, both with a standard deviation of 5. The normally distributed random intercepts ν_i and η_i depend on the variances σ_ν and σ_η, which have normal distrubutions each with the prior mean 1 and prior standard deviation 1. For all other parameters ($\gamma_1, \ldots, \gamma_r$, δ_1, $\beta_1, \ldots, \beta_p$, β_z, β_{PM}, and

$\beta_{z\times\mathrm{PM}}$), we use less informative independent normal priors, each with mean zero and standard deviation 20.

4.5 Simulation

4.5.1 Setup

We create simulated samples with similarity to the data from the motivating cookstove intervention study, using them to test the performance of the models proposed in Section 4.3 and 4.4. The assumed model for the simulated outcome Y_{it} is the same as that of Section 3.4 with the addition of an exposure term PM_{it}. Simulated participants are randomized to one of two study arms in a stepped-wedge design and are measured for six time points ($t = 1, 2, 3, 4, 5, 6$). Both study arms are assigned control until the first study arm is assigned the treatment at the third time point, and the second arm is assigned treatment at the fith time point. We generate a latent principal stratum U_{it} for every participant at every time point. For the first time point, we draw starting quantities of strata from a multinomial distribution with probabilities of 2/3 for compliers, 1/6 for always-takers, and 1/6 for never-takers. Then we use a Markov chain to generate latent strata for each of the following time points by subject. The transition matrix for the chain is symmetric with 0.6 in its diagonal entries and 0.2 in the off-diagonal entries. From U_{it} the subject's behavior at each time point, S_{it}, is determined. The assumed model for simulated outcomes including PM as a mediator is as follows:

$$Y_{it} = \mathbf{x}_i^\top \boldsymbol{\beta} + \beta_1 S_{it} + \beta_2 \mathrm{PM}_{it} + \sin(t) + \nu_i + \epsilon_{it} \tag{4.7}$$

$$\text{where} \quad \mathrm{PM}_{it} = \beta_3 S_{it} + (0.5)\cos(2t) + \eta_i + \epsilon'_{it}. \tag{4.8}$$

Here $\beta_1 = -0.1$ is the direct at-the-time effect of received treatment (S_{it}) on the outcome, while $\beta_2 \cdot \beta_3 = 0.5 \cdot (-0.5) = 0.25$ is the at-the-time effect of S_{it} on the outcome that is mediated by PM. The participant-specific random effects ν_i and η_i are both i.i.d. $N(0, 0.5)$, while the error terms ϵ_{it} and ϵ'_{it} are both i.i.d. $N(0, 1)$. We use the same five covariates in $\mathbf{x}_i$ as described in Section 3.4.1,

which are constant in time. The five coefficient values in $\boldsymbol{\beta}$ are again fixed as 0.5, 0.5, 1, 1, and 0.25, respectively. Finally, we include a time trend for the outcome via the term $\sin(t)$ and a trend in time with shorter periods for the exposure with the term $(0.5)\cos(2t)$. Note that the simulation model does not include covariates $\mathbf{q}_{it}$ specifically related to the exposure PM_{it} or dependency of covariates $\mathbf{X}$ on the latent strata U_{it}. In Chapter 3, we conducted an additional simulation study with dependency of U_{it} on $\mathbf{X}$, and found similar trends in the model results to the main simulation study.

We create 200 replicate samples each with 300 participants and estimate complier average causal effects (CACE_L) for thresholds L from every sample in four ways. We fit the longitudinal models from Sections 4.3 and 4.4 to the full simulated samples, and fit both the unmodified approach (without random intercepts or the fixed effect α_t) from Hackstadt et al. (2014) and the principal scores (PS) estimands from Equation (3.2) on single exposure and outcome measurements. To reduce the repeated measures to single measures, we first compute the averages of exposure for every subject during the first two time points when all participants are assigned control ($\overline{\mathrm{PM}}_i^{12}$). Next we compute averages during time points 3 and 4 for each subject ($\overline{\mathrm{PM}}_i^{34}$), when the first study arm is receiving treatment and the second arm is still assigned control. The difference in averages (DA) ($\overline{\mathrm{PM}}_i^{34} - \overline{\mathrm{PM}}_i^{12} = \mathrm{PM}_i^{DA}$) is then the exposure of interest for the PS estimands and non-longitudinal Bayesian approach. Similarly, we compute $\overline{Y}_i^{34} - \overline{Y}_i^{12} = Y_i^{DA}$ for each subject. Before using the PS-WLM model and PS estimands, we convert PM_{it} and PM_i^{DA} respectively to binary exposure variables using a threshold L, considering subject i exposed if their exposure is less than L and unexposed otherwise. For each of the four model variants, we apply all possible threshold values ranging from $L = -5$ to $L = 1$ by increments of 0.05. For the principal scores methods, some of these thresholds values result in an extreme case of the binary exposure variable PM_{it}^L, having very few of a value (1 or 0) or none at all. These extreme exposure variables can then lead to numerically unstable weights so that we are unable to produce an estimate. For this reason, we only report a subset of the thresholds between -5 and 1 for the estimates in Figure 4.2. Finally,

Table 4.1: Selected model fit results from 200 simulated samples. For each model type, we include the best fits with respect to RMSE and to bias, which may be obtained under different thresholds L. We also include the PS-WLM fits when using treatment use (S_{it}) to define principal strata instead of the exposure PM_{it}^{L}. Power and coverage are measured for 95% confidence intervals.

Model	Criteria	L	RMSE	Bias	Power	Coverage
Bayesian Longitudinal	Lowest RMSE	-0.35	0.053	0.0164	1.000	0.920
	Least Bias	0.05	0.059	0.0033	1.000	0.930
Bayesian Single Measure	Lowest RMSE	-0.50	0.070	0.0177	1.000	0.925
	Least Bias	0.40	0.089	0.0001	0.995	0.885
PS-WLM (Longitudinal)	Lowest RMSE	-0.35	0.585	-0.5782	1.000	0.000
	Least Bias	-0.40	0.585	-0.5774	1.000	0.000
		(Treatment Use)	0.122	0.0050	0.785	0.955
PS (Single Measure)	Lowest RMSE	0.00	0.816	-0.5444	0.384	0.939
	Least Bias	0.00	0.816	-0.5444	0.384	0.939

we also find the CACE estimates for all samples by applying the PS-WLM model with potental actual treatment use ($S_{it}(0)$ and $S_{it}(1)$) defining the principal strata, the same as in Chapter 3.

4.5.2 Results

The results from model fits on the 200 simulated samples are reported in Figures 4.2 and 4.3 for the principal scores (PS-WLM and PS) methods and Bayesian approach from Section 4.4 respectively. The Bayesian approach has high power, high coverage, and low bias for thresholds L near 0. The longitudinal (including random effects) version of the Bayesian model produces lower RMSE and higher power than the summarized single measure version (on Y^{DA} and PM^{DA}). The thresholds L which correspond to the best performances in RMSE and bias for each model are reported in Table 4.1. The Bayesian models both have lower RMSE and higher power than any of the principal scores-based methods. The application of PS-WLM when using actual treatment S_{it} (same as in Chapter 3), is more accurate than the principal scores methods that use the actual exposure PM_{it}, which are both negatively biased. The PS-WLM model performs better than the PS estimands in RMSE and bias, but has lower coverage and higher power, indicating its Wald confidence intervals are shifted away from the true effect by biased estimates.

We attribute the negative bias in the principal scores methods to the issue of overlapping exposure distributions under the causal effects of treatment and control, which we discussed in Section 4.3. The average causal effect of the treatment on exposure level may be small relative to the overall variation of the exposure. In this case, consider all observed exposures (in subject and time) which have been causally lowered by the treatment. The distribution of measurements for these exposures, and the distribution of measurements for the exposures not causally lowered by treatment are overlapping. This overlap means that although the exposure of subject i at time t with $Z_{it} = 1$ may belong to the distribution of exposures causally lowered by treatment, their measurement PM_{it} is not lower than L and $\mathrm{PM}_{it}^{L} = 0$. Then subject i is considered a never-taker at time t and is not included in the estimation of CACE_L for the PS-WLM and PS methods. Similarly, when the exposure of a subject with $Z_{it} = 0$ belongs to the distribution of exposures not causally lowered by treatment but $\mathrm{PM}_{it}^{L} = 1$, they are considered an always-taker and excluded from the calculation of CACE_L. Assuming a negative true effect of treatment on exposure, the observations excluded by stratum membership belong to the right tail of the treated distribution ($S_{it} = 1$) and the left tail of the untreated distribution ($S_{it} = 0$). This systematic removal of observations from the estimation for CACE_L results in a strong negative bias for the longitudinal and single measure principal scores models which rely on PM_{it}^{L}, as seen in Table 4.1 and Figure 4.2.

4.6 Re-Analysis of Honduran Stepped-Wedge Cookstove Trial

Once again, our motivating study is the Honduran randomized controlled cookstove intervention from Young et al. (2019), who used a stepped-wedge design to provide improved (Justa) stoves to 230 households and measured stove use, exposure to PM, health outcomes, and other covariates from primary cooks over six study visits in three years. Between the second and third study visits (phases), researchers randomly assigned participants to a study arm and the household either received the Justa cookstove before phase 3 (Arm 1) or before phase 5 (Arm 2), so that every household eventually received the intervention, but in a staggered fashion (Figure 3.2). For our analysis, we again estimate the stratified average causal effects of stove assigment (traditional or

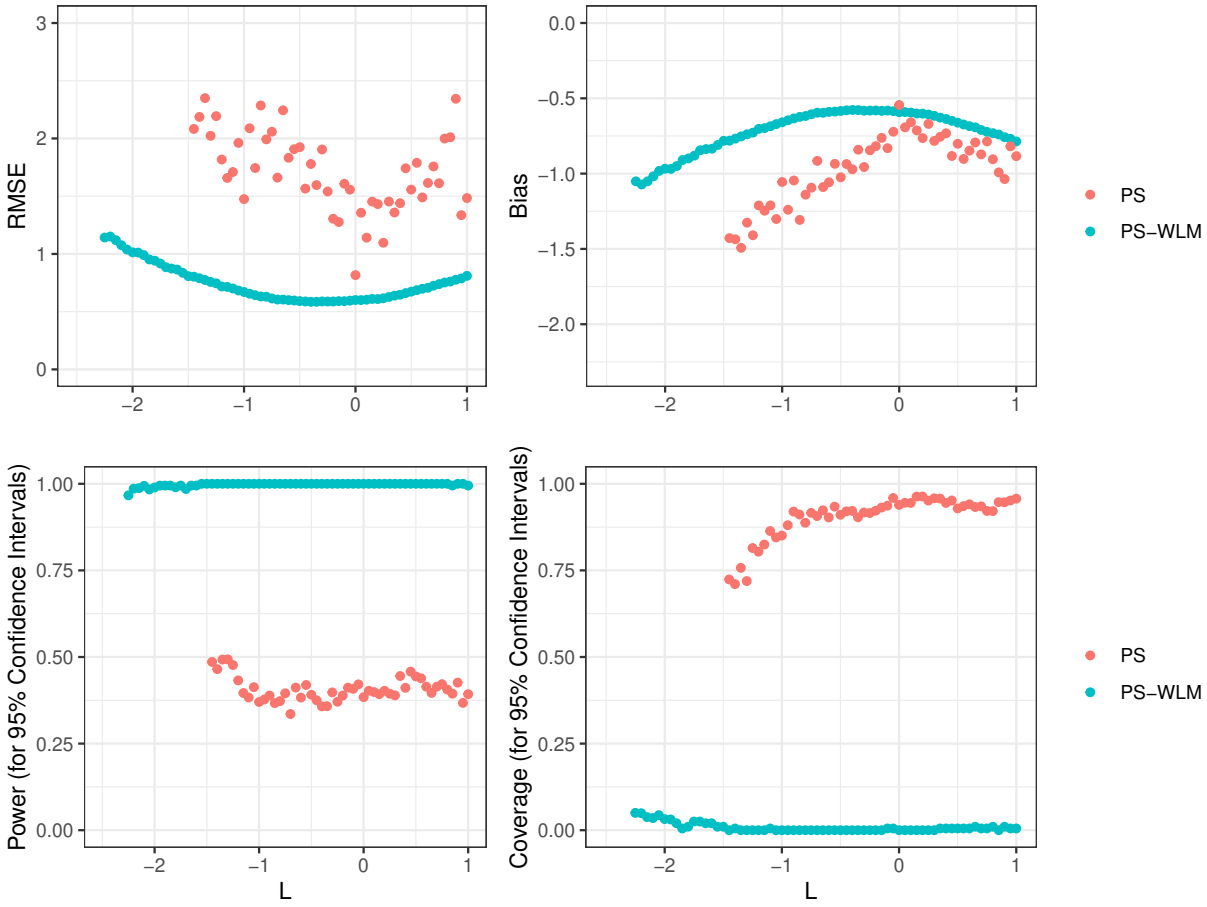

Figure 4.2: Results from the PS-WLM model and PS estimands estimates on 200 simulated samples, measuring RMSE, bias, power, and coverage with respect to the true total effect of stove assignment on the outcome (-0.35). For each sample, the model was fit using several thresholds L (on the x-axis) to define the binary exposure variable PM_{it}^{L}. Not all thresholds could be used on any given sample without leading to numerically extreme weights, so each point in the plot is a summary of estimates from between 100 and 200 samples.

Justa) on the outcome, percent of glycated hemoglobin (HbA1c), that is depicted in Figure 3.2. As a post-treatment variable to define the principal strata, we no longer use the subject's stove use S_{it}, but instead use their personal 24 hour time-weighted average exposure to PM. We apply the Bayesian approach from Hackstadt et al. (2014) and its extension from Section 4.4 to the study data, estimating a complier average causal effect of stove assignment on HbA1c.

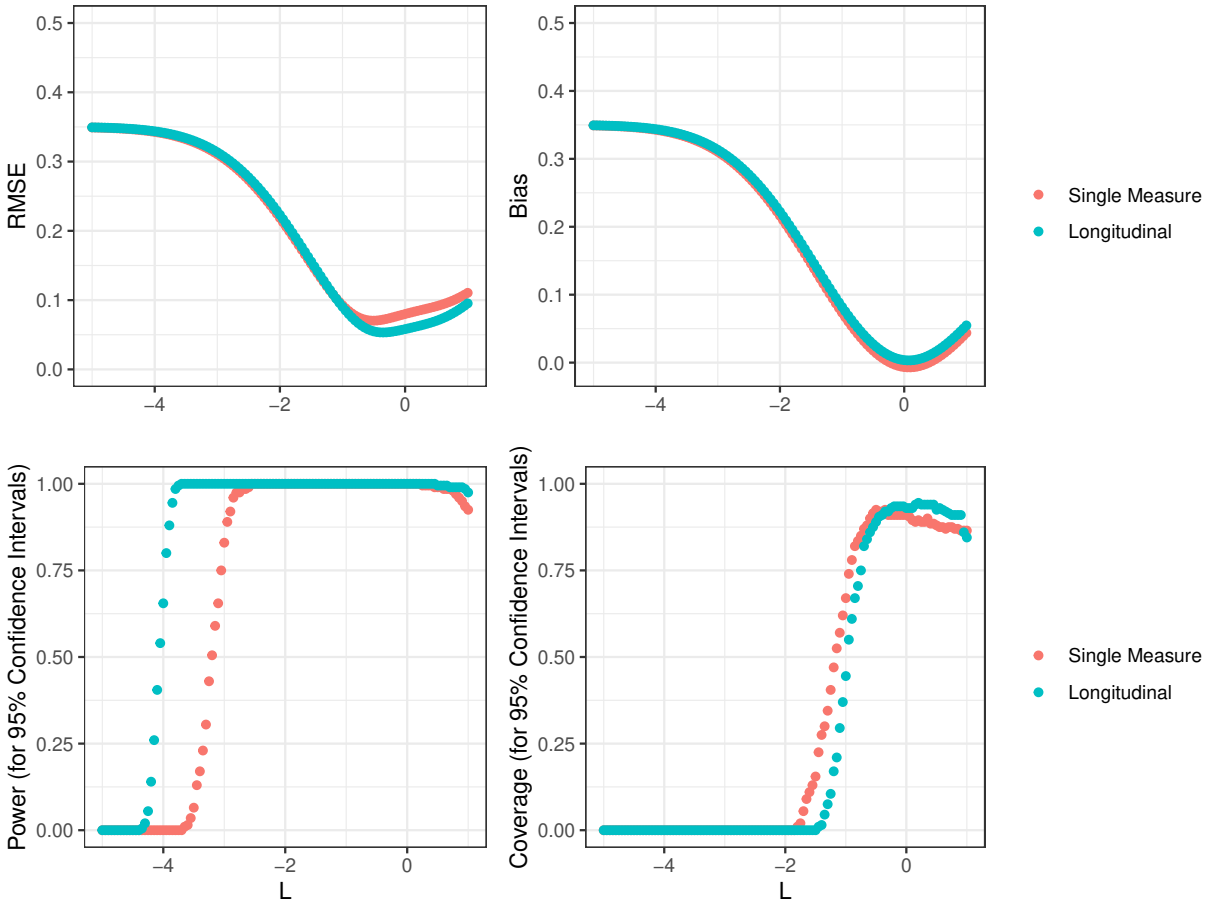

Figure 4.3: Results from model fits on 200 simulated samples using the method developed in Section 4.4 (Longitudinal) and the unmodified method from Hackstadt et al. (2014) (Single Measure), measuring RMSE, bias, power, and coverage with respect to the true total effect of stove assignment on the outcome (-0.35). For each sample, the model was fit using several thresholds L (on the x-axis) to define the complier average causal effect $\text{CACE}_L = E\left\{Y_{it}(1) - Y_{it}(0) \mid \text{PM}_{it}(1) - \text{PM}_{it}(0) < L\right\}$.

4.6.1 Setup and Filtering

For subject i at phase t we define Z_{it} as the assigned cookstove, Y_{it} as the outcome HbA1c, and PM_{it} as log-transformed 24 hour time-weighted average $\text{PM}_{2.5}$ (micrograms per cubic meter) measured by a personal monitor (Figure 4.4). We use the same outcome covariates in $\mathbf{X}$ as in Section 2.4. These are the binary covariates of whether or not the primary cook is over 40 years of age, has a BMI over 25, has metabolic syndrome, or has electricity available in their household, and also the log-transformed continuous measurements of peripheral systolic blood pressure, all measured pre-randomization. For covariates $\mathbf{Q}$ related to the exposure PM_{it}, we use the binary

covariates of whether electricity is available in the household and whether kerosene is used as a source of light. We restrict the study observations to complete cases, removing any observations that are missing Z_{it} ($n = 28$), PM_{it} ($n = 178$), Y_{it} ($n = 168$), or some part of $\mathbf{x}_i$ or $\mathbf{q}_i$ ($n = 6$) for 215 total removed and $n = 1,171$ observations remaining.

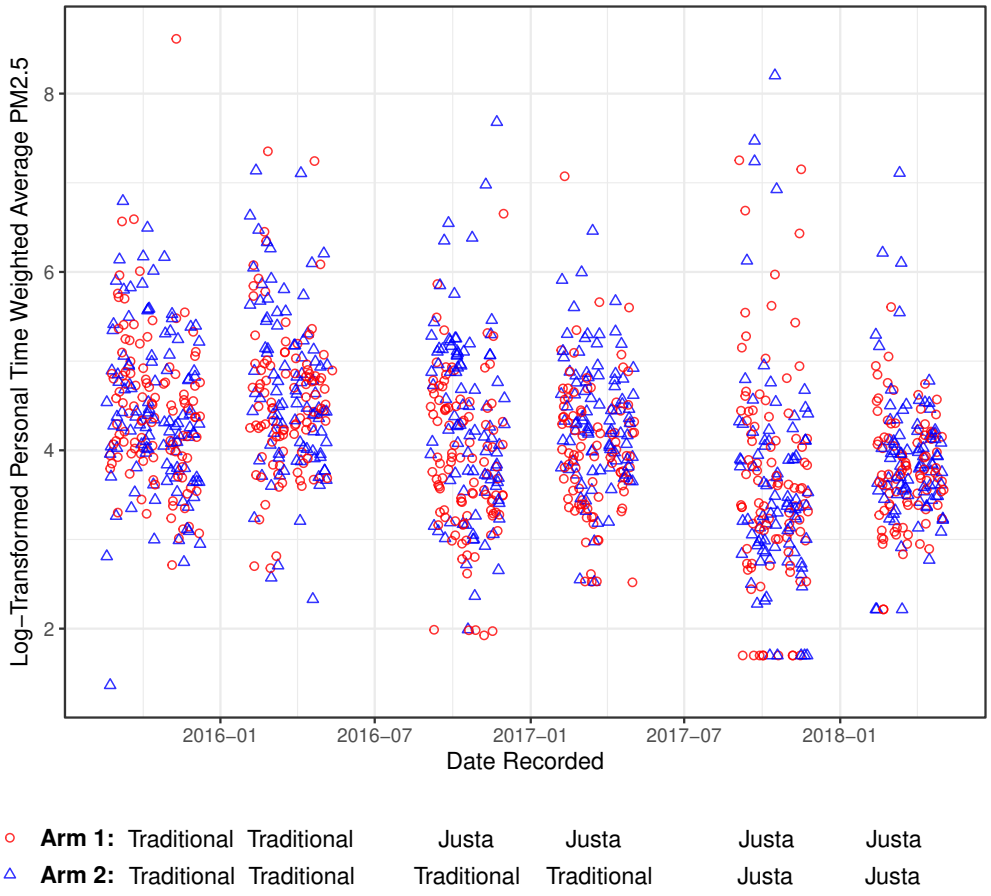

Figure 4.4: Personal $\text{PM}_{2.5}$ measurements ($\mu g/m^3$) from of 230 primary cooks measured across six study phases, with the assigned stoves (Traditional or Justa) for study arms 1 and 2 listed below the plot. The measurements are time-weighted averages and log-transformed.

4.6.2 Results

We report the confidence bands and estimates for CACE_L by the longitudinal and single measure Bayesian approaches in Figure 4.5. We find no significant estimate from either model among

the thresholds $L = -5$ to $L = 2$ by increments of 0.05. For the same thresholds, we found no significant DACE_L or UACE_L estimates.

Recall that in Section 3.6.3, we reported significant complier (-0.15 with 95% confidence interval -0.22 to -0.08) and never-taker (-0.11 with 95% confidence interval -0.18 to -0.05) effects from the same data with the PS-WLM model using stove use (S_{it}) to define the principal strata. The effects $\mathrm{CACE} = E\left\{Y_{it}(1) - Y_{it}(0) \mid S_{it}(1) = 1, S_{it}(0) = 0\right\}$ and $\mathrm{CACE}_L = E\left\{Y_{it}(1) - Y_{it}(0) \mid \mathrm{PM}_{it}(1) - \mathrm{PM}_{it}(0) < L\right\}$ are the same except for two cases. First, consider a subject at time t who would use the stove assigned to them but would not experience a difference less than L in exposure when using the improved stove. If we assume PM has the largest effect of any variable on the outcome, the potential outcomes values of these subjects would have little to no difference. Since these observations would be included in the estimation of CACE, but not in CACE_L, we would expect to see more attenuation in the former. Second, consider a subject at time t who would not comply with their stove assignment, but would experience a difference less than L in exposure when using the improved stove. The potential reduction in PM_{it} means that the potential outcomes of these subjects would differ. These observations would contribute to the estimation of CACE_L but not to CACE.

Reasons for the discrepancy between the estimates here and in Section 3.6.3 may include measurement error or inconsistencies in the individual $\mathrm{PM}_{2.5}$ exposures (Figure 4.4), or a strong direct effect of stove assignment on the outcome HbA1c that is due to the unblindedness of the study and not mediated by PM exposure.

In Figure 4.6, we depict the estimated differences in potential values for the exposure and outcome, where posterior means from the Bayesian model in Section 4.4 are used in place of the unobserved counterfactuals. The points are plotted with lines indicating no difference for each axis. The threshold L can be represented by a vertical line, with points to its right having $\mathrm{PM}_{it}^L = 0$ and points to its left having $\mathrm{PM}_{it}^L = 1$. For comparison, consider the threshold values from the x-axis of Figure 4.5 as they would be drawn in Figure 4.6. If PM has a reducing causal effect on the outcome Y, then more negative differences in the exposures potential values should correspond

with more negative differences in potential outcomes. This would be indicated by a greater relative concentration of points in the lower left quadrant of the plot defined by the intersecting lines. In Figure 4.6, there are more points and some strong negative differences of potential outcome values in lower left quadrant, but the trend is not obvious and there are also positive outliers in $Y_{it}(1) - Y_{it}(0)$ on the left hand side of the plot.

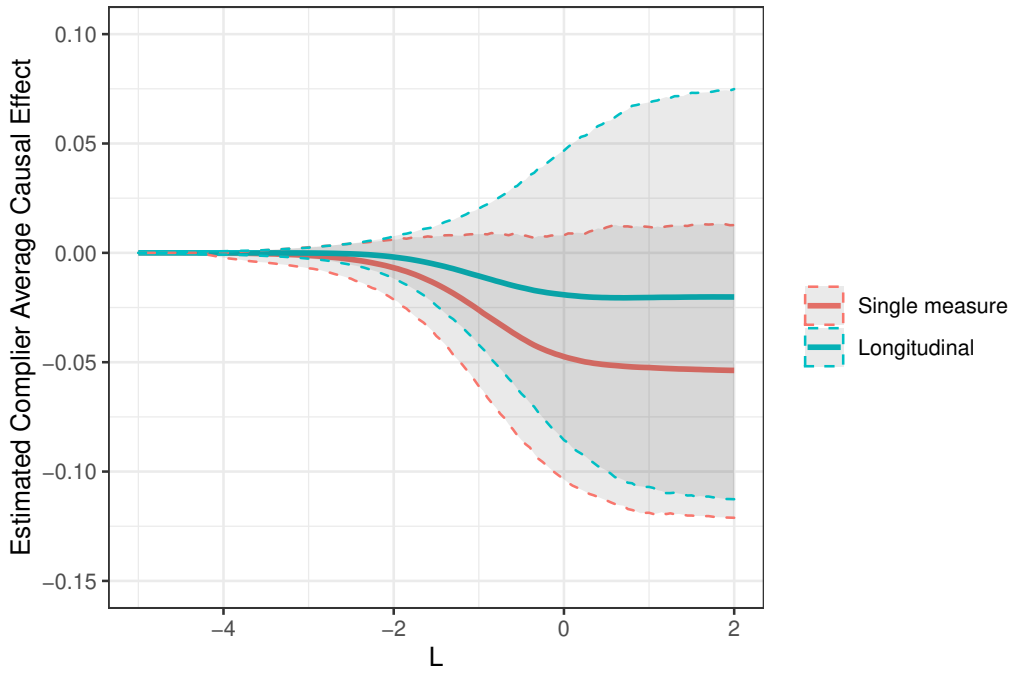

Figure 4.5: Estimated complier average causal effect ($\text{CACE}_L = E\{Y_{it}(1) - Y_{it}(0) \mid \text{PM}_{it}(1) - \text{PM}_{it}(0) < L\}$) on the Honduran cookstove intervention data using the Bayesian approach from Section 4.4 (Longitudinal) and the unmodified approach from Hackstadt et al. (2014) (Single Measure).

4.7 Discussion

In an effort to study the relationship between indoor air pollution and health outcomes within a longitudinal controlled trial, we have explored the applications and extensions of several methods. We were again motivated by the Honduran cookstove intervention that we analyzed in Section 3.6 (Young et al., 2019). Instead of use of treatment, we were interested in the actual ($PM_{2.5}$) exposures that the subject experienced and how the exposure impacted health outcomes. To estimate this

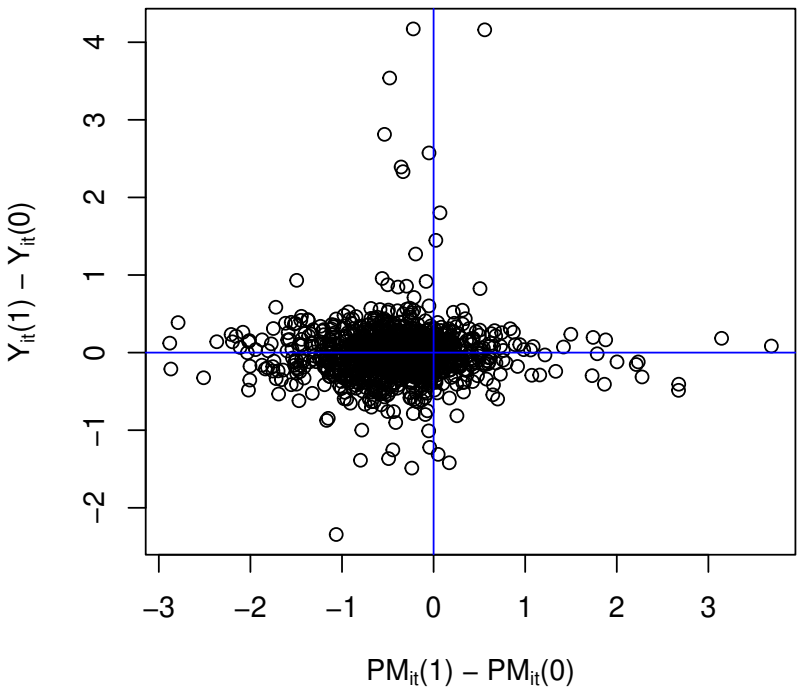

Figure 4.6: Scatter plot of estimated differences in exposure potential values $\mathrm{PM}_{it}(1) - \mathrm{PM}_{it}(0)$ (μg/m^3) by differences in potential outcome values $Y_{it}(1) - Y_{it}(0)$ (percent glycated hemoglobin). The unobserved counterfactuals are estimated as posterior means from the Bayesian model in Section 4.4. We include lines for differences of 0 in potential outcomes and potential exposures.

effect in a principal stratification framework, we used the potential exposure values as the post-treatment variable to define the latent strata. To apply the PS-WLM model developed in Section 3.3.1 and the PS estimands in Equation (3.2), we must first convert the continuous exposures to a binary variable defined by some numerical threshold L. We also extend the Bayesian approach from Hackstadt et al. (2014) to a longitudinal setting by including two random intercepts and a fixed effect by time point in the assumed outcome and exposure models. In the simulation study of Section 4.5, we find that the longitudinal and single measure (without random intercepts or α_t) Bayesian models outperform the PS-WLM and PS methods in every metric and across all thresholds L, as the principal scores models both exhibit negative bias. We posit that this negative bias is due to overlapping distributions of causally lowered and not casually lowered

exposures, leading to truncation in both distributions by the binary exposure variable. Finally, the two Bayesian model variants, when applied to the motivating cookstove intervention study data did not yield a significant effect of any kind across all thresholds L.

The simulations make clear that the Bayesian approach should be chosen over the negatively biased principal scores methods. In Table 4.1 we report the results of the PS-WLM model fit on simulated samples with actual treatment use defining the principal strata, as in Chapter 3. These results show that the PS-WLM model will perform better and without strong bias when using actual treatment use (if it is available), although the Bayesian approaches are still superior at certain thresholds L.

The Bayesian approaches are reliant on the choice of covariates $\mathbf{x}_i$ and $\mathbf{q}_i$ to predict well the exposure and outcome counterfactuals. Poor predictors of the exposure and outcome could diminish the accuracy of the estimated counterfactual exposures and outcomes, impacting the estimated causal effects. Other possible areas of sensitivity in the Bayesian approach include choices of priors and the fixed value ρ. These choices can be altered and tested in an additional simulation study for example, but must be measured while also checking threshold values L, making for a complicated comparison scheme. With their original model, Hackstadt et al. (2014) performed a sensitivity analysis for different fixed values of ρ and found results to be robust to different choices.

In this work we have explored the application of several principal stratification methods to a longitudinal crossover study when using a continous exposure measurement to define the latent strata. These same methods can be applied to parallel group designs, like the Household Air Pollution Intervention Network (HAPIN) and other large randomized controlled cookstove interventions (Clasen et al., 2022).

Chapter 5

Conclusion

We have developed and presented methods in three areas. The first was a penalized regression model to predict PM exposures in space and time that was faster than spatial-only and spatiotemporal universal kriging methods, and more accurate when measurements were missing at regular intervals. The second methodological area was the extension of principal stratification to a longitudinal setting. We developed a weighted linear mixed model (PS-WLM) which relied on principal scores to estimate causal effects by strata, as well as a Bayesian mixture (BM) model to perform the same task. The PS-WLM model outperformed the BM and original principal scores method (CAPS) from Ding and Lu (2017) in every simulation scenario, and was the only model to produce a significant effect estimate from the Honduran cookstove intervention study. The third area of contribution was the implementation of principal stratification methods when using a continuous measure of PM exposure as the post-treatment variable to define principal strata. We found that the PS-WLM model is biased when applied with continuous exposure measurements that have been categorized. We extended a Bayesian principal stratification approach from Hackstadt et al. (2014) to the longitudinal setting, and found that it performed better than the PS-WLM model when using categorized exposure PM_{it}^{L} or actual treatment use S_{it} to define principal strata.

5.1 Estimation of CACE from Chapter 3 vs. CACE_L from Chapter 4

The causal effects targeted in Chapters 3 and 4 are similar, but distinct, quantities. In Chapter 3, we estimate the at-the-time complier average causal effect (CACE) of treatment assignment on the outcome. We define the principal stratum of compliers at-the-time as subjects at time points when they would use their assigned stove ($S_{it}(1) = 1$ and $S_{it}(0) = 0$). In Chapter 4, we estimate at-the-time complier average causal effects for a threshold value L ($\mathrm{CACE}_{\mathrm{L}}$), where compliers at-the-time

are subjects at times when $\mathrm{PM}_{it}(1) - \mathrm{PM}_{it}(0) < L$. With these two definitions of compliers, the estimates for CACE and CACE_L may be calculated from differing groups of subjects and time points. For example, a subject i and time t may have $\mathrm{PM}_{it}(1) - \mathrm{PM}_{it}(0) < L$, but $S_{it}(0) = S_{it}(1) = 0$.

An additional cause for disparity between CACE and CACE_L is that the Bayesian model from Section 4.4 takes a fundamentally different approach to estimating a complier effect than the principal scores methods from Sections 3.2.4 and 3.3.1. Instead of applying weights to observations to create a pseudo population for estimating average effects, the Bayesian approach estimates all missing potential exposures and outcomes and is able to estimate differences in potential values at an individual level. This is an advantage when using a continuous exposure variable to define principal strata, since the individual differences in potential exposures contain more information than the observed exposures that have been converted to a binary variable.

5.2 Future Work

Each of the methods that we developed were tested on simulated samples. Simulation allows us to measure the models' performance with respect to the true effects built in to the generative process. There are many choices that we make in the generation of simulated samples, balancing the complication of the process. If we add too many changing parameters, interpretation of the model fits for the samples may become difficult, but if we over-simplify the generative process, the model fits are perfectly specified and uninteresting. In Chapter 3, we generated simulated samples which met the assumptions listed in Section 3.2.3. In addition to producing the method of principal scores for principal stratification, Ding and Lu (2017) developed methods of testing the sensitivity to certain violations of the general principal ignorability and monotonicity assumptions (3 and 4 in Section 3.2.3). An additional simulation study, involving samples with violated assumptions, could explore the importance of these assumptions and robustness of the models to their violation. In Chapter 4, the inclusion of an intermediary exposure variable (PM_{it}) in the simulated samples meant that we did not guarantee all the assumptions from Section 3.2.3 to be satisified. In addition,

the Bayesian approach from Section 4.4 relied on fewer assumptions than the PS-WLM model from Section 3.3.1. Thus, an investigation of the consequences in these models from violations in their assumptions is warranted.

We performed an additional simulation study in Section 3.5 where the latent principal strata U_{it} were dependent on $\mathbf{X}$, which each of the models (CAPS, PS-WLM, and BM) were built to account for. The results of this simulation study confirmed the models' ability to handle such dependency, but only for one construction of the relationship between $\mathbf{X}$ and U_{it}. Different forms of a dependency between $\mathbf{X}$ and U_{it}, when added to the simulation studies in both Chapters 3 and 4, would be of interest.

In Chapter 4, we simulated samples without covariates $\mathbf{q}_{it}$ by subject or time that influenced the exposure value PM_{it}. The PS-WLM and Bayesian methods from this chapter are built to handle covariates for the exposure, so the simulated samples were over-simplified in this respect. The addition of covariates related to the exposure PM_{it} to the simulation study is a possible future work.

Bibliography

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434):444–455.

Baumgartner, J., Schauer, J. J., Ezzati, M., Lu, L., Cheng, C., Patz, J. A., and Bautista, L. E. (2011). Indoor air pollution and blood pressure in adult women living in rural China. *Environmental Health Perspectives*, 119(10):1390–1395.

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., Tsai, M.-Y., Künzli, N., Schikowski, T., Marcon, A., et al. (2013). Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe – The ESCAPE project. *Atmospheric Environment*, 72:10–23.

Bergen, S., Sheppard, L., Kaufman, J. D., and Szpiro, A. A. (2016). Multipollutant measurement error in air pollution epidemiology studies arising from predicting exposures with penalized regression splines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(5):731–753.

Bergen, S. and Szpiro, A. A. (2015). Mitigating the impact of measurement error when using penalized regression to model exposure in two-stage air pollution epidemiology studies. *Environmental and Ecological Statistics*, 22(3):601–631.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010). A Spatio-Temporal Downscaler for Output From Numerical Models. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(2):176–197.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2012). Space-Time Data fusion Under Error in Computer Model Output: An Application to Modeling Air Quality. *Biometrics*, 68(3):837–848.

Berrocal, V. J., Guan, Y., Muyskens, A., Wang, H., Reich, B. J., Mulholland, J. A., and Chang, H. H. (2020). A comparison of statistical and machine learning methods for creating national daily maps of ambient PM2.5 concentration. *Atmospheric Environment*, 222:117130.

Bonjour, S., Adair-Rohani, H., Wolf, J., Bruce, N. G., Mehta, S., Prüss-Ustün, A., Lahiff, M., Rehfuess, E. A., Mishra, V., and Smith, K. R. (2013). Solid Fuel Use for Household Cooking: Country and Regional Estimates for 1980–2010. *Environmental Health Perspectives*, 121(7):784–790.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2010). Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

Brown, C. A. and Lilford, R. J. (2006). The stepped wedge trial design: a systematic review. *BMC Medical Research Methodology*, 6(1):54.

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, 97(2):109–131.

Chuang, K.-J., Yan, Y.-H., Chiu, S.-Y., and Cheng, T.-J. (2011). Long-term air pollution exposure and risk factors for cardiovascular diseases among the elderly in Taiwan. *Occupational and Environmental Medicine*, 68(1):64–68.

Clark, M. L., Bachand, A. M., Heiderscheidt, J. M., Yoder, S. A., Luna, B., Volckens, J., Koehler, K. A., Conway, S., Reynolds, S. J., and Peel, J. L. (2013). Impact of a cleaner-burning cookstove intervention on blood pressure in Nicaraguan women. *Indoor Air*, 23(2):105–114.

Clasen, T. F., Chang, H. H., Thompson, L. M., Kirby, M. A., and et. al. (2022). Liquefied Petroleum Gas or Biomass for Cooking and Effects on Birth Weight. *New England Journal of Medicine*, 387(19):1735–1746.

Dai, J. Y., Gilbert, P. B., and Mâsse, B. R. (2012). Partially hidden Markov model for time-varying principal stratification in HIV prevention trials. *Journal of the American Statistical Association*, 107(497):52–65.

Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A. S., and Schaap, M. (2016). Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *The Annals of Applied Statistics*, 10(3):1286–1316.

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., Lyapustin, A., et al. (2019). An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment International*, 130:104909.

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., Lyapustin, A., et al. (2020). Assessing NO2 Concentration and Model Uncertainty with High Spatiotemporal Resolution across the Contiguous United States Using Ensemble Model Averaging. *Environmental Science & Technology*, 54(3):1372–1384.

Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., and Schwartz, J. (2016). Assessing PM2.5 Exposures with High Spatiotemporal Resolution across the Continental United States. *Environmental science & technology*, 50(9):4712–4721.

Ding, P. and Lu, J. (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):757–777.

Frangakis, C. E., Brookmeyer, R. S., Varadhan, R., Safaeian, M., Vlahov, D., and Strathdee, S. A. (2004). Methodology for Evaluating a Partially Controlled Longitudinal Treatment Using Principal Stratification, With Application to a Needle Exchange Program. *Journal of the American Statistical Association*.

Frangakis, C. E. and Rubin, D. B. (2002). Principal Stratification in Causal Inference. *Biometrics*, 58(1):21–29.

Gallop, R., Small, D. S., Lin, J. Y., Elliott, M. R., Joffe, M., and Ten Have, T. R. (2009). Mediation analysis with principal stratification. *Statistics in Medicine*, 28(7):1108–1130.

Gneiting, T. (2002). Nonseparable, Stationary Covariance Functions for Space–Time Data. *Journal of the American Statistical Association*, 97(458):590–600. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214502760047113.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729.

Hackstadt, A. J., Matsui, E. C., Williams, D. L., Diette, G. B., Breysse, P. N., Butz, A. M., and Peng, R. D. (2014). Inference for environmental intervention studies using principal stratification. *Statistics in Medicine*, 33(28):4919–4933.

He, M. Z., Kloog, I., Just, A. C., Gutéerrez-Avila, I., Colicino, E., Téllez-Rojo, M. M., Pizano-Zárate, M. L., Tamayo-Ortiz, M., Cantoral, A., Soria-Contreras, D. C., et al. (2022). Intermediate- and long-term associations between air pollution and ambient temperature and glycated hemoglobin levels in women of child bearing age. *Environment international*, 165:107298.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., and Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, 42(33):7561–7578.

Jin, H. and Rubin, D. B. (2008). Principal Stratification for Causal Inference With Extended Partial Compliance. *Journal of the American Statistical Association*, 103(481):101–111.

Keet, C. A., Keller, J. P., and Peng, R. D. (2018). Long-Term Coarse Particulate Matter Exposure Is Associated with Asthma among Children in Medicaid. *American Journal of Respiratory and Critical Care Medicine*, 197(6):737–746.

Keller, J. P. and Peng, R. D. (2019). Error in estimating area-level air pollution exposures for epidemiology. *Environmetrics*, 30(8).

Lin, J. Y., Ten Have, T. R., and Elliott, M. R. (2008). Longitudinal Nested Compliance Class Model in the Presence of Time-Varying Noncompliance. *Journal of the American Statistical Association*, 103(482):462–473.

Lin, J. Y., Ten Have, T. R., and Elliott, M. R. (2009). Nested Markov Compliance Class Model in the Presence of Time-Varying Noncompliance. *Biometrics*, 65(2):505–513.

Lindström, J., Szpiro, A. A., Sampson, P. D., Oron, A. P., Richards, M., Larson, T. V., and Sheppard, L. (2014). A Flexible Spatio-Temporal Model for Air Pollution with Spatial and Spatio-Temporal Covariates. *Environmental and Ecological Statistics*, 21(3):411–433.

Malm, W. C., Sisler, J. F., Huffman, D., Eldred, R. A., and Cahill, T. A. (1994). Spatial and seasonal trends in particle concentration and optical extinction in the United States. *Journal of Geophysical Research: Atmospheres*, 99(D1):1347–1370.

Mann, J. K., Lutzker, L., Holm, S. M., Margolis, H. G., Neophytou, A. M., Eisen, E. A., Costello, S., Tyner, T., Holland, N., Tindula, G., et al. (2021). Traffic-related air pollution is associated with glucose dysregulation, blood pressure, and oxidative stress in children. *Environmental Research*, 195:110870.

McCracken, J. P., Smith, K. R., Díaz, A., Mittleman, M. A., and Schwartz, J. (2007). Chimney stove intervention to reduce long-term wood smoke exposure lowers blood pressure among Guatemalan women. *Environmental Health Perspectives*, 115(7):996–1001.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., Jović, D., Woollen, J., Rogers, E., Berbery, E. H., et al. (2006). North American Regional Reanalysis. *Bulletin of the American Meteorological Society*, 87(3):343–360.

Murray, C. J. L., Aravkin, A. Y., Zheng, P., Brauer, M., Afshin, A., Lim, S. S., and et. al. (2020). Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, 396(10258):1223–1249.

Paciorek, C. J., Yanosky, J. D., Puett, R. C., Laden, F., and Suh, H. H. (2009). Practical Large-Scale Spatio-Temporal Modeling of Particulate Matter Concentrations. *The Annals of Applied Statistics*, 3(1):370–397.

Pati, D., Reich, B. J., and Dunson, D. B. (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, 98(1):35–48.

Peng, R. D., Butz, A. M., Hackstadt, A. J., Williams, D. L., Diette, G. B., Breysse, P. N., and Matsui, E. C. (2015). Estimating the health benefit of reducing indoor air pollution in a randomized environmental intervention. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(2):425–443.

Petersen, A. and Witten, D. (2019). Data-adaptive additive modeling. *Statistics in Medicine*, 38(4):583–600.

Reff, A., Phillips, S., Eyth, A., and Mintz, D. (2020). Bayesian Space-time Downscaling Fusion Model (Downscaler) - Derived Estimates of Air Quality for 2017. Technical Report EPA-454/R-20-005, United States Environmental Protection Agency, Office of Air Quality Planning and Standards Air Quality Assessment Division Research Triangle Park, NC.

Reich, B. J., Chang, H. H., and Foley, K. M. (2014). A spectral method for spatial downscaling. *Biometrics*, 70(4):932–942.

Reich, B. J., Fuentes, M., and Dunson, D. B. (2011). Bayesian Spatial Quantile Regression. *Journal of the American Statistical Association*, 106(493):6–20.

Robins, J. M. and Greenland, S. (1992). Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology*, 3(2):143.

Romieu, I., Riojas-Rodríguez, H., Marrón-Mares, A. T., Schilmann, A., Perez-Padilla, R., and Masera, O. (2009). Improved Biomass Stove Intervention in Rural Mexico. *American Journal of Respiratory and Critical Care Medicine*, 180(7):649–656.

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association*, 75(371):591–593.

Ryder, N. A. and Keller, J. P. (2023). Spatiotemporal Exposure Prediction with Penalized Regression. *Journal of Agricultural, Biological and Environmental Statistics*, 28(2):260–278.

Sampson, P. D., Richards, M., Szpiro, A. A., Bergen, S., Sheppard, L., Larson, T. V., and Kaufman, J. D. (2013). A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2.5 concentrations in epidemiology. *Atmospheric Environment*, 75:383–392.

Schabenberger, O. and Gotway, C. A. (2004). *Statistical Methods for Spatial Data Analysis*. CRC Press.

Schlather, M., Malinowski, A., Menck, P., Oesting, M., and Strokorb, K. (2015). Analysis, Simulation and Prediction of Multivariate Random Fields with Package Random Fields. *Journal of statistical software*, 63:1–25.

Smith, K. R., McCracken, J. P., Weber, M. W., Hubbard, A., Jenny, A., Thompson, L. M., Balmes, J., Diaz, A., Arana, B., and Bruce, N. (2011). Effect of reduction in household air pollution on childhood pneumonia in Guatemala (RESPIRE): a randomised controlled trial. *The Lancet*, 378(9804):1717–1726.

Stan Development Team (2023a). RStan: the R interface to Stan. R package version 2.21.8.

Stan Development Team (2023b). Stan modeling language users guide and reference manual. version 2.31.

Szpiro, A. A. and Paciorek, C. J. (2013). Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics*, 24(8):501–517.

Tielsch, J. M., Katz, J., Khatry, S. K., Shrestha, L., Breysse, P., Zeger, S., Checkley, W., Mullany, L. C., Kozuki, N., LeClerq, S. C., and Adhikari, R. (2016). Effect of an improved biomass stove

on acute lower respiratory infections in young children in rural Nepal: a cluster-randomised, step-wedge trial. *The Lancet Global Health*, 4:S19.

U.S. Environmental Protection Agency (2019). *Integrated science assessment (ISA) for particulate matter (final report, Dec 2019)*. U.S. Environmental Protection Agency, Washington, DC.

Wang, M., Sampson, P. D., Hu, J., Kleeman, M., Keller, J. P., Olives, C., Szpiro, A. A., Vedal, S., and Kaufman, J. D. (2016). Combining Land-Use Regression and Chemical Transport Modeling in a Spatiotemporal Geostatistical Model for Ozone and PM2.5. *Environmental Science & Technology*, 50(10):5111–5118.

Witinok-Huber, R., Clark, M. L., Volckens, J., Young, B. N., Benka-Coker, M. L., Walker, E., Peel, J. L., Quinn, C., and Keller, J. P. (2022). Effects of household and participant characteristics on personal exposure and kitchen concentration of fine particulate matter and black carbon in rural Honduras. *Environmental Research*, 214:113869.

Woo, H., Koehler, K., Putcha, N., Lorizio, W., McCormack, M., Peng, R., and Hansel, N. N. (2023). Principal stratification analysis to determine health benefit of indoor air pollution reduction in a randomized environmental intervention in COPD: Results from the CLEAN AIR study. *Science of The Total Environment*, 868:161573.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Xu, H., Bechle, M. J., Wang, M., Szpiro, A. A., Vedal, S., Bai, Y., and Marshall, J. D. (2019). National PM2.5 and NO2 exposure models for China based on land use regression, satellite measurements, and universal kriging. *Science of The Total Environment*, 655:423–433.

Young, B. N., Peel, J. L., Benka-Coker, M. L., Rajkumar, S., Walker, E. S., Brook, R. D., Nelson, T. L., Volckens, J., L'Orange, C., Good, N., et al. (2019). Study protocol for a stepped-wedge randomized cookstove intervention in rural Honduras: household air pollution and cardiometabolic health. *BMC Public Health*, 19(1):903.

Young, M. T., Bechle, M. J., Sampson, P. D., Szpiro, A. A., Marshall, J. D., Sheppard, L., and Kaufman, J. D. (2016). Satellite-Based NO2 and Model Validation in a National Prediction Model Based on Universal Kriging and Land-Use Regression. *Environmental Science & Technology*, 50(7):3686–3694.

Appendix A

Data and Computing Acknowledgements

A.1 For Chapter 2

The Air Quality System data used in this work is publicly available from the Environmental Protection Agency website: https://www.epa.gov/outdoor-air-quality-data/download-daily-data. Similarly, the IMPROVE observations can be accessed at the website: http://vista.cira.colostate.edu/Improve/. `R` code for the simulations and ambient air pollution analysis in Chapter 2 is available through the publisher, The Journal of Agricultural, Biological, and Environmental Statistics.

IMPROVE is a collaborative association of state, tribal, and federal agencies, and international partners. The U.S. Environmental Protection Agency is the primary funding source, with contracting and research support from the National Park Service. The Air Quality Group at the University of California, Davis is the central analytical laboratory, with ion analysis provided by Research Triangle Institute, and carbon analysis provided by Desert Research Institute.

A.2 For Honduran Cookstove Study from Young et al. (2019)

The research in Honduras was funded by the National Institute of Environmental Health Sciences of the National Institutes of Health under award number ES022269.

A.3 Overall

This work utilized resources from the University of Colorado Boulder Research Computing Group, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University.

Appendix B

Supplementary Material for Chapter 2

B.1 Objects

The observed data, indexed by site i and date t are stored in the vector $\mathbf{x}$. The model coefficients β_{kt}, indexed by covariate k and time t are stored in vectors $\boldsymbol{\beta}_t$ and collated into the stacked vector $\boldsymbol{\beta}$. These vectors are written as

$$\mathbf{x} = \begin{bmatrix} x_{11} \\ \vdots \\ x_{1T} \\ x_{21} \\ \vdots \\ x_{nT} \end{bmatrix}, \qquad \boldsymbol{\beta}_t = \begin{bmatrix} \beta_{0t} \\ \vdots \\ \beta_{(p-1)t} \end{bmatrix}, \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \vdots \\ \boldsymbol{\beta_T} \end{bmatrix}.$$

The penalty matrix $\boldsymbol{\Lambda}_1$ applies the scalar penalty λ_1 to all model coefficients in $\boldsymbol{\beta}$ except those of the intercept and any other variables we choose not to penalize. We have penalty matrix blocks (excluding the intercept β_0) written as

$$\underset{p\times p}{\boldsymbol{\Gamma}_1} = \begin{bmatrix} 0 & 0 & 0 & \cdots \\ 0 & \lambda_1 & 0 & \cdots \\ 0 & 0 & \lambda_1 & \\ \vdots & & \ddots & \ddots \end{bmatrix}, \text{ so that } \underset{pT\times pT}{\boldsymbol{\Lambda_1}} = \begin{bmatrix} \underset{p\times p}{\boldsymbol{\Gamma_1}} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \underset{p\times p}{\boldsymbol{\Gamma_1}} & \mathbf{0} & \cdots \\ \vdots & & \ddots & \ddots \end{bmatrix}.$$

The covariates used to fit the model, indexed by site i and time t are stored in vectors $\mathbf{r}_{it}$ and collated into the matrix $\mathbf{R}$ along with $p \times 1$ zero vectors $\mathbf{0}$ as shown below. To produce the matrix $\boldsymbol{R}_{obs}$ we need only remove the rows containing $\mathbf{r}_{it}$ for combinations of date t and site i where no

measurement was observed.

$$\mathbf{r}_{it} = \begin{bmatrix} r_{0it} \\ \vdots \\ r_{(p-1)it} \end{bmatrix} \qquad \text{and} \qquad \underset{nT\times pT}{\mathbf{R}} = \begin{bmatrix} \mathbf{r}_{11}^\top & \mathbf{0}^\top & \mathbf{0}^\top & \cdots & \\ \mathbf{0}^\top & \mathbf{r}_{12}^\top & \mathbf{0}^\top & \cdots & \\ \mathbf{0}^\top & \mathbf{0}^\top & \ddots & & \\ \vdots & \vdots & & & \\ & & & & \mathbf{r}_{1T}^\top \\ \mathbf{r}_{21}^\top & \mathbf{0}^\top & \mathbf{0}^\top & \cdots & \\ \mathbf{0}^\top & \mathbf{r}_{22}^\top & \mathbf{0}^\top & & \\ \mathbf{0}^\top & \mathbf{0}^\top & \ddots & & \\ \vdots & \vdots & & & \\ & & & & \mathbf{r}_{nT}^\top \end{bmatrix}.$$

The difference matrix $\boldsymbol{D}$ shown below treats each sequential pair of predictions as adjacent in time. If observations are missing at timepoints, sequential predictions in the from the model may not be adjacent in time, so we may adjust $\boldsymbol{D}$ by removing rows which would compare non-adjacent dates. Furthermore, we can define "adjacency" by any number of days between observations. To make $\boldsymbol{D}$ we have difference matrix blocks

$$\underset{(T-1)\times T}{\mathbf{d}} = \begin{bmatrix} -1 & 1 & 0 & \cdots \\ 0 & -1 & 1 & \cdots \\ \vdots & & \ddots & \ddots \end{bmatrix}, \quad \text{so that} \quad \underset{n(T-1)\times nT}{\mathbf{D}} = \begin{bmatrix} \underset{(T-1)\times T}{\mathbf{d}} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \underset{(T-1)\times T}{\mathbf{d}} & \mathbf{0} & \cdots \\ \vdots & & \ddots & \ddots \end{bmatrix}.$$

Table B.1: Median RMSE and R^2 values from 100 simulated test samples for our penalized regression model (Pen. Reg.), our model without its time smoothing penalty ($\lambda_2 = 0$), universal kriging (UK), and the `SpatioTemporal` model (ST). Note that H is the Hurst Index for $\boldsymbol{Z}_2(\boldsymbol{s}, t)$ and that the proportions of locations observed only every third day or only every sixth day (instead of daily) are p_{third} and p_{sixth}, respectively. Lastly, σ is the standard deviation of the non-spatial error ϵ_t.

				Pen. Reg.		$\lambda_2 = 0$		UK		ST	
σ	H	p_{third}	p_{sixth}	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
0.5	0.05	0.333	0.167	0.381	0.895	0.383	0.895	0.605	0.786	0.347	0.911
0.5	0.05	0.667	0.167	0.463	0.843	0.467	0.838	0.906	0.481	0.383	0.892
0.5	0.05	0.667	0.333	0.360	0.912	0.360	0.911	0.399	0.883	0.699	0.665
0.5	0.50	0.333	0.167	0.371	0.874	0.380	0.869	0.415	0.844	0.319	0.906
0.5	0.50	0.667	0.167	0.436	0.846	0.463	0.824	0.477	0.821	0.339	0.902
0.5	0.50	0.667	0.333	0.354	0.901	0.360	0.898	0.401	0.875	0.462	0.822
1.5	0.05	0.333	0.167	0.519	0.811	0.523	0.810	0.665	0.640	0.453	0.856
1.5	0.05	0.667	0.167	0.605	0.744	0.615	0.732	1.004	0.400	0.510	0.815
1.5	0.05	0.667	0.333	0.493	0.836	0.496	0.834	0.666	0.695	0.724	0.626
1.5	0.50	0.333	0.167	0.499	0.783	0.517	0.761	0.642	0.660	0.397	0.857
1.5	0.50	0.667	0.167	0.550	0.747	0.610	0.683	0.693	0.605	0.433	0.840
1.5	0.50	0.667	0.333	0.478	0.809	0.491	0.793	0.659	0.662	0.509	0.774

Finally, we obtain the time-smoothing penalty term from Equation (3) via

$$\mathbf{R}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{r}_{11}^\top \boldsymbol{\beta}_1 \\ \mathbf{r}_{12}^\top \boldsymbol{\beta}_2 \\ \vdots \\ \mathbf{r}_{1T}^\top \boldsymbol{\beta}_T \\ \mathbf{r}_{21}^\top \boldsymbol{\beta}_1 \\ \vdots \\ \mathbf{r}_{nT}^\top \boldsymbol{\beta}_T \end{bmatrix} \quad \text{and} \quad \mathbf{DR}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{r}_{12}^\top \boldsymbol{\beta}_2 - \mathbf{r}_{11}^\top \boldsymbol{\beta}_1 \\ \mathbf{r}_{13}^\top \boldsymbol{\beta}_3 - \mathbf{r}_{12}^\top \boldsymbol{\beta}_2 \\ \vdots \\ \mathbf{r}_{1T}^\top \boldsymbol{\beta}_T - \mathbf{r}_{1(T-1)}^\top \boldsymbol{\beta}_{T-1} \\ \mathbf{r}_{22}^\top \boldsymbol{\beta}_2 - \mathbf{r}_{21}^\top \boldsymbol{\beta}_1 \\ \vdots \\ \mathbf{r}_{nT}^\top \boldsymbol{\beta}_T - \mathbf{r}_{n(T-1)}^\top \boldsymbol{\beta}_{T-1} \end{bmatrix}.$$

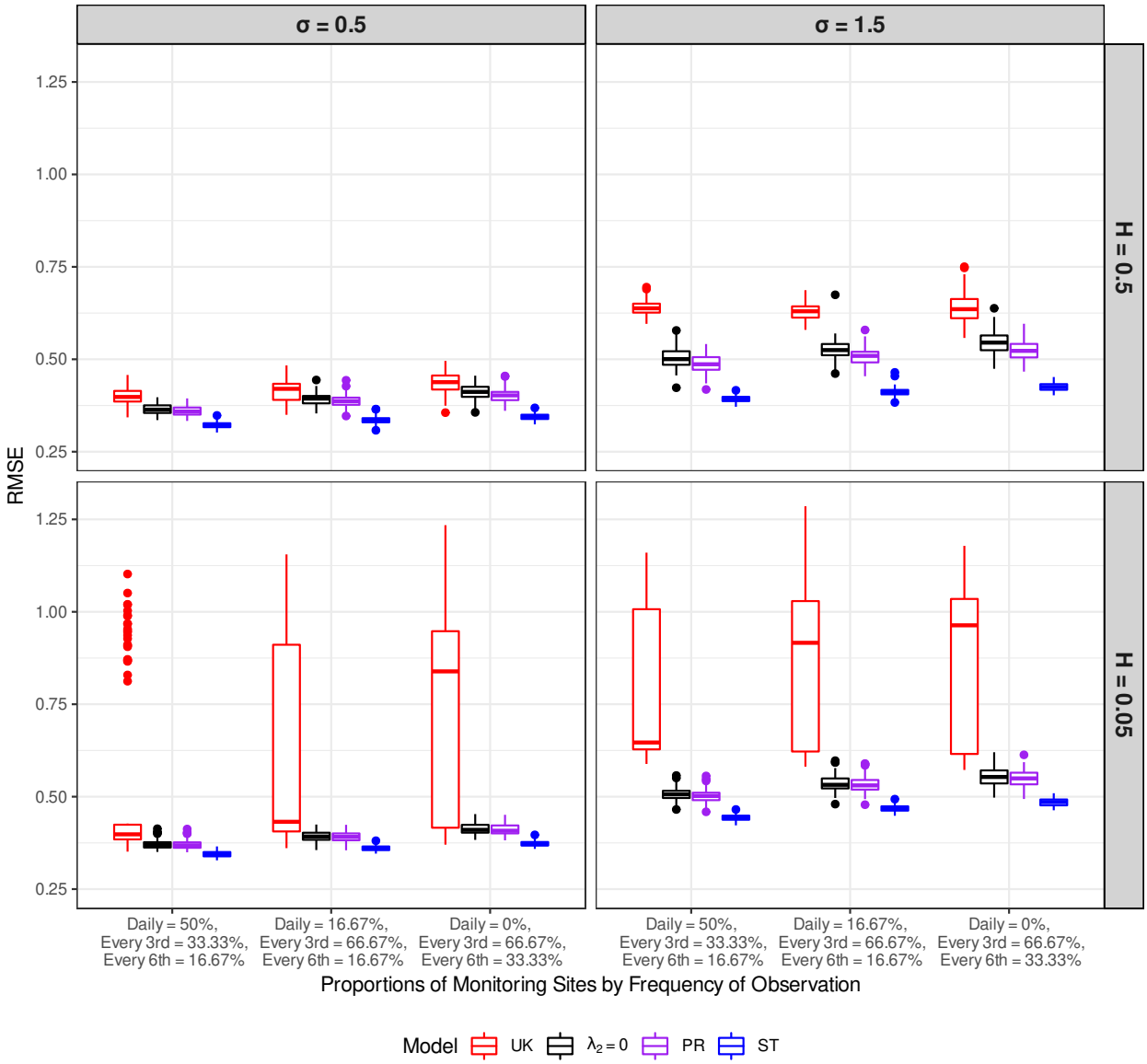

Figure B.1: Performance of spatiotemporal predictive models under staggered infrequent measurement structure. Selected monitoring locations are observed daily, every third day, or every sixth day, according to the proportions listed on the x-axis. If a location is observed infrequently, the dates of observation are not synchronized, i.e. on any given date some locations of each schedule (daily, every third, and every sixth) are observed. Shown are boxplots of RMSE values from each model on 100 replicate samples for each simulation scenario. In order, the boxplots correspond to universal kriging (UK), our penalized regression model without its temporal smoothing penalty ($\lambda_2 = 0$), our model with the penalty (PR), and the `SpatioTemporal` (ST) model. Note that H is the Hurst index for $Z_2(s,t)$, which results in a more variable temporal trend as H approaches zero, and that σ is the standard deviation of non-spatial error added to the training data.

Table B.2: Median computation times from 10 replicate simulated data sets with N_{train} monitoring locations. For the "CV + Fit" time, our penalized regression model (Pen. Reg.) selects over penalty values and number of TPRS basis functions, then fits the resulting best model ("Fit"). The universal kriging (UK) and `SpatioTemporal` (ST) model fitting times include estimation of covariance parameters ("Cov + Fit").

	Pen. Reg.		UK	ST
N_{train}	CV + Fit	Fit	Cov + Fit	Cov + Fit
500	4.2 mins	0.2 mins	7.1 mins	181.1 mins
1000	8.0 mins	0.6 mins	19.1 mins	833.2 mins
1500	13.9 mins	0.9 mins	32.5 mins	2608.6 mins
2000	22.9 mins	1.8 mins	53.6 mins	5486.5 mins

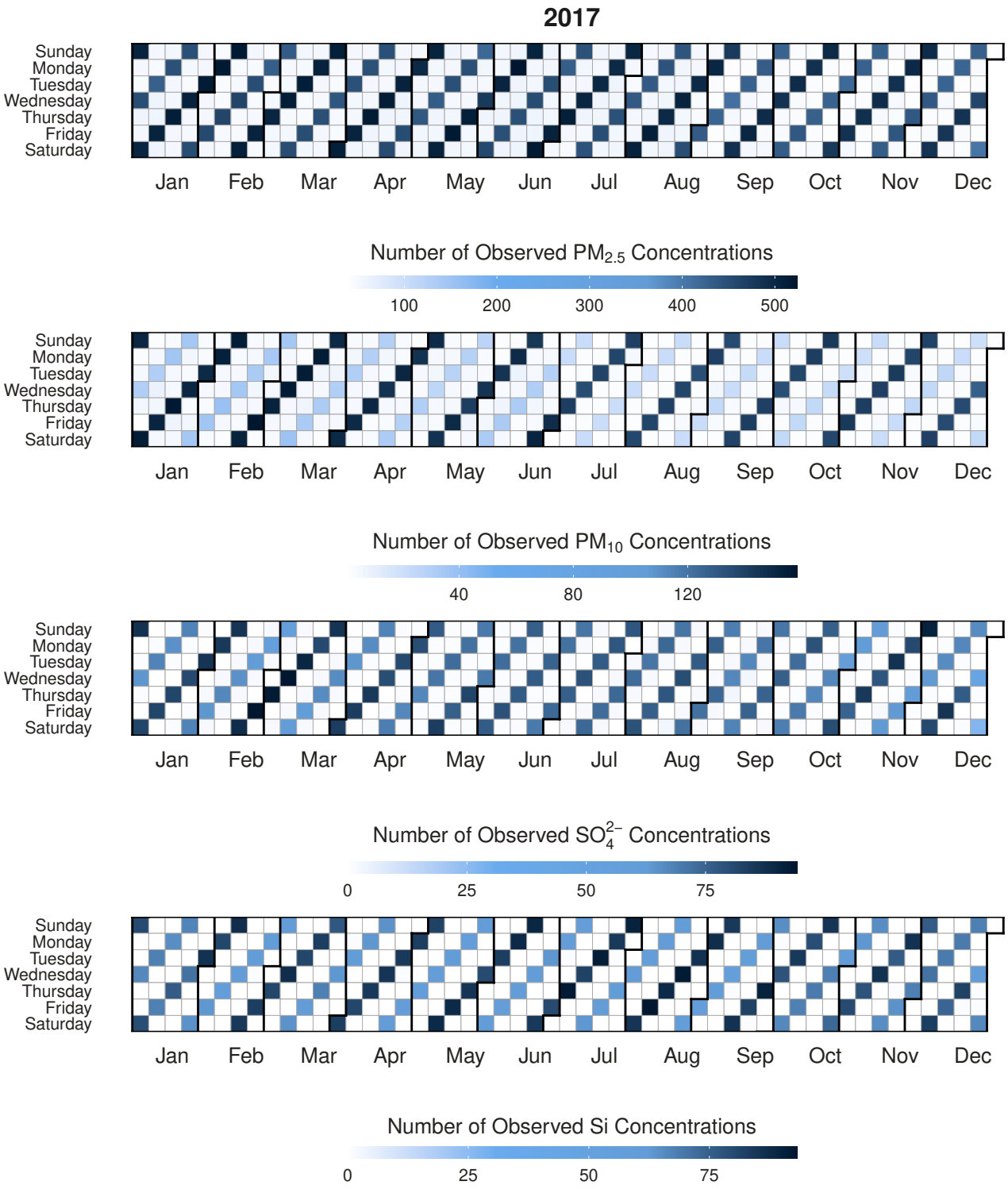

Figure B.2: Calendar heatmaps of the number of uniquely situated $PM_{2.5}$, PM_{10}, sulfate, and silicon measurements recorded per day in 2017 across the eastern United States.

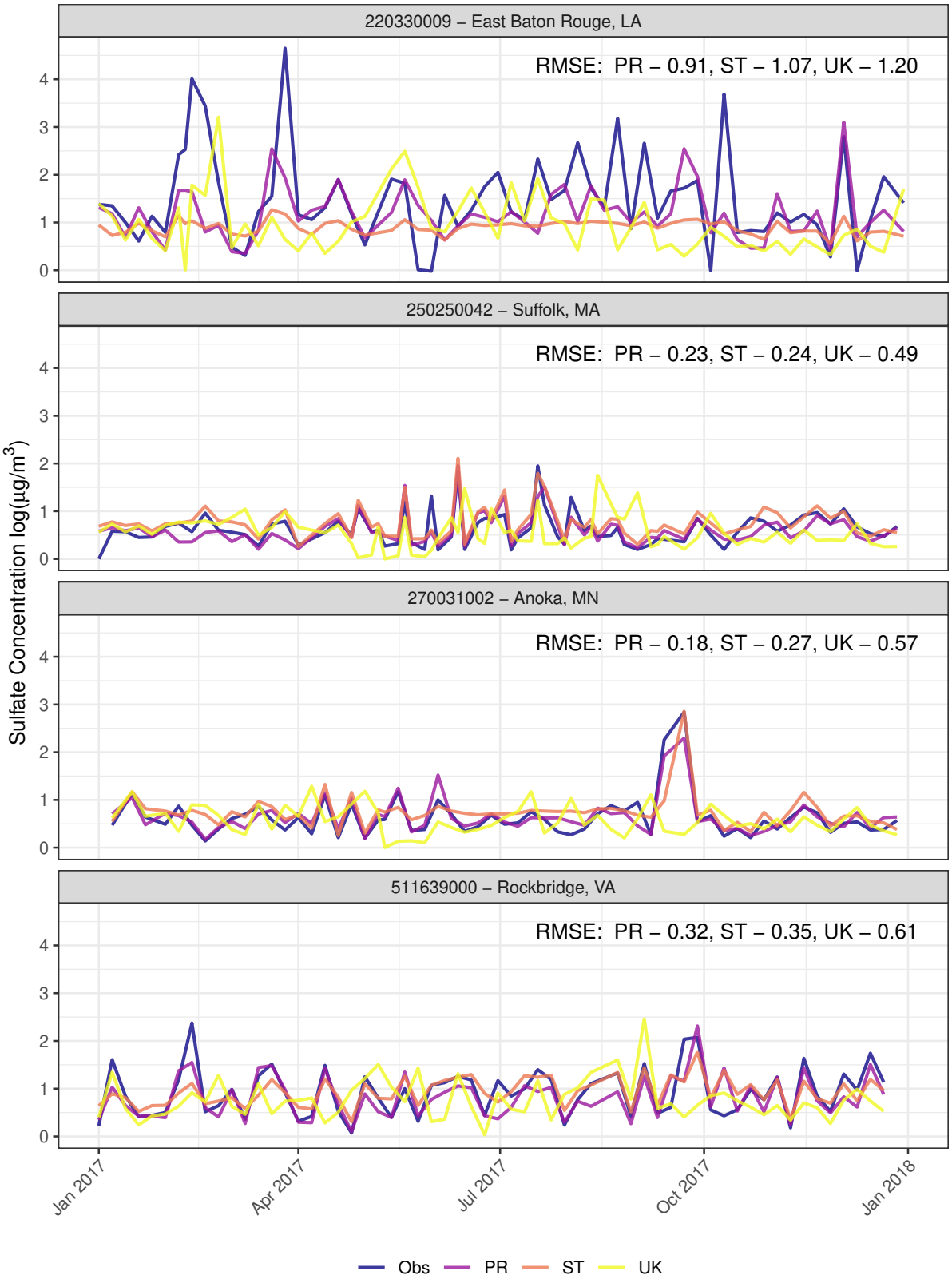

Figure B.3: Observed sulfate concentrations (Obs) over time at four randomly selected monitoring locations. We include predictions from our penalized regression model (PR), the `SpatioTemporal` model (ST), and universal kriging (UK). We also report the RMSE values of all models at each site.

Appendix C

Supplementary Material for Chapter 3

C.1 Never-Taker and Always-Taker Covariate-Adjusted Estimands

We present the never-taker and always-taker average causal effect (NACE and AACE) estimators with covariate-adjustment. The complier average causal effect (CACE) covariate-adjusted estimator is written in Equation (3) of the main text.

$$\begin{aligned}\text{NACE} = E\big\{\big(Y_{it} - \boldsymbol{\beta}_{1,n}^{\mathrm{T}}\mathbf{x}_i\big) \mid Z_{it}=1, S_{it}=0\big\} \\ - E\left\{w_{0,n}(\mathbf{x}_i)\left(Y_{it} - \boldsymbol{\beta}_{0,n}^{\mathrm{T}}\mathbf{x}_i\right) \mid Z_{it}=0, S_{it}=0\right\} \\ + \left(\boldsymbol{\beta}_{1,n} - \boldsymbol{\beta}_{0,n}\right)^{\mathrm{T}} E\left\{w_{z,n}(\mathbf{x}_i)\mathbf{x}_i \mid Z_{it}=z, S_{it}=0\right\}\end{aligned} \tag{C.1}$$

$$\begin{aligned}\text{AACE} = E\big\{w_{1,a}(\mathbf{x}_i)\big(Y_{it} - \boldsymbol{\beta}_{1,a}^{\mathrm{T}}\mathbf{x}_i\big) \mid Z_{it}=1, S_{it}=1\big\} \\ - E\left\{\left(Y_{it} - \boldsymbol{\beta}_{0,a}^{\mathrm{T}}\mathbf{x}_i\right) \mid Z_{it}=0, S_{it}=1\right\} \\ + \left(\boldsymbol{\beta}_{1,a} - \boldsymbol{\beta}_{0,a}\right)^{\mathrm{T}} E\left\{w_{z,a}(\mathbf{x}_i)\mathbf{x}_i \mid Z_{it}=z, S_{it}=1\right\}.\end{aligned} \tag{C.2}$$

Here $z = 0$ and 1 in the third terms of both Equations (C.1) and (C.2). Then $w_{z,n}(\mathbf{x}_i) = w_{0,n}(\mathbf{x}_i)$ when $Z_{it} = 0$ and $w_{z,n}(\mathbf{x}_i) = 1$ when $Z_{it} = 1$. Similarly, $w_{z,a}(\mathbf{x}_i) = 1$ when $Z_{it} = 0$ or $w_{z,a}(\mathbf{x}_i) = w_{1,a}(\mathbf{x}_i)$ when $Z_{it} = 1$

Table C.1: The prior values of the Bayesian model used for all simulation fits. Each prior followed a normal distribution.

Parameter	Prior Mean	Prior Standard Deviation
μ_{zu}	0	1
σ_u	4	1
$\boldsymbol{\beta}$	$\mathbf{0}$	$\mathbf{1}$
$\boldsymbol{\gamma}_u$	$\mathbf{0}$	$\mathbf{1}$
σ_ν	1	1

Table C.2: The "t-statistics" calculated from the estimated differences described in Equation 5 with 200 bootstrap samples to approximate standard error (rounded to three decimal places). Each covariate is binary except peripheral systolic blood pressure which we natural log-transform.

Covariate	CACE	NACE	AACE
Age > 40	0.433	0.433	0.126
Electricity	-0.505	-0.167	0.059
log(Peripheral Systolic Blood Pressure)	-0.352	0.198	0.052
Metabolic Syndrome	0.000	0.271	0.113
BMI > 25	0.251	0.740	0.079