Distribution Agreement

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

Signature:

_____ _____

Behzad Kianian Date

Statistical Methods for Spatial Data in Public Health

By

Behzad Kianian
Doctor of Philosophy

Biostatistics and Bioinformatics

Lance A. Waller, Ph.D.
Advisor

Howard H. Chang, Ph.D.
Committee Member

Adam N. Glynn, Ph.D.
Committee Member

Rachel E. Patzer, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Statistical Methods for Spatial Data in Public Health

By

Behzad Kianian
B.A., University of Pennsylvania, PA, 2009
M.S., Emory University, GA, 2017

Advisor: Lance A. Waller, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics and Bioinformatics
2020

Abstract

Statistical Methods for Spatial Data in Public Health
By Behzad Kianian

Data in public health often contain a spatial component relevant to understanding underlying relationships of interest. Accounting for different manifestations of spatial components in statistical analyses is frequently challenged by a dearth of developed methodology or high computational costs. First, we consider the problem of estimating treatment effects from observational data with propensity score matching allowing for the presence of spatial and multi-level confounding. We build on recently developed distance-adjusted propensity score matching (DAPSm) and propose a two-stage approach that first matches within clusters (WC), and then uses the DAPSm approach to match remaining subjects (WC+DAPsm). We demonstrate the benefits and robustness of our approach through an extensive simulation study. We apply our method to a population of patients in Georgia who have recently started dialysis, where both the treatment (informed of transplant options) and outcome (1-year referral for transplant) may be plausibly affected by individual, facility, and area-level factors.

Next, we consider the task of using satellite-derived aerosol optical depth (AOD) as a predictor for particulate matter ($PM_{2.5}$) concentrations, allowing broader coverage than the network of air pollution monitors. However, AOD contains large contiguous areas of missing data due to cloud cover. We propose imputing missing AOD data using lattice kriging, a large-scale spatial statistical method, and random forest, a regression tree-based machine learning method, as well as a distance-based ensemble for combining the two methods. Throughout our application, we construct cross-validation folds and testing data based on spatially clustered holdouts more closely mimicking observed data patterns than traditional random holdouts. Our results show that the proposed distance-based ensemble outperforms individual methods.

For the third topic, we discuss on-going work assessing the equity of COVID-19 testing site access in the Atlanta area. We adapt methods from the environmental justice literature using empirical cumulative distribution functions to compare demographic subgroup access to testing sites. We consider different measures of access, and we conduct Monte Carlo simulations of test site placements under different sampling schemes to assess factors associated with site placement.

Statistical Methods for Spatial Data in Public Health

By

Behzad Kianian
B.A., University of Pennsylvania, PA, 2009
M.S., Emory University, GA, 2017

Advisor: Lance A. Waller, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics and Bioinformatics
2020

Acknowledgments

The work presented here would not have been possible without the support of Lance Waller, who always maintained a positive attitude, regardless of setbacks and obstacles. I am indebted to Howard H. Chang for his constant support, as well as to my other committee members, Rachel Patzer and Adam Glynn, for providing invaluable feedback and guidance that has made my research possible. My development as a graduate student would not be possible without the help of other faculty members in the Department of Biostatistics and Bioinformatics, particularly Limin Peng. I also thank Amita Manatunga, Robert Lyles, Ying Guo, Michael Haber, Yijian Huang, Yijuan Hu, Steve Qin, Reneé Moore, David Benkeser, Benjamin Risk, Rebecca Zhang, John Hanfelt, Paul Weiss, Kirk Easley, Qi Long, Jian Kang, and the rest of the faculty for outstanding instruction and support. My work as a student would not be possible without the outstanding support staff in our department: Mary Abosi, Angela Guinyard, Joy Hearn, Melissa Sherrer, and Bob Waggoner. Finally, I thank my peers for their camaraderie and friendship.

Contents

List of Figures

List of Tables

Chapter 1

Propensity score matching for multi-level and spatial data

1.1 Introduction

In observational studies, estimating treatment effects often is complicated by large observed and unobserved differences between treated and control groups. Within the potential outcomes framework, propensity score matching has gained popularity over the last several decades for its potential to adjust for the observed differences between treated and control groups to generate more credible estimates of treatment effects under certain assumptions (Rosenbaum and Rubin, 1983; Stuart, 2010). Validity of propensity score matching results depends crucially on the assumption of no unmeasured confounding, such that all pre-treatment variables relevant for both the outcome and treatment assignment are assumed to be measured and included in the propensity score estimation procedure. In many settings, however, unobserved confounders may exist at the local geographic (spatial-) and facility-level.

Disclaimer: A portion of the data reported here have been supplied by the United States Renal Data System (USRDS). The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the U.S. government.

Recent work has extended propensity score matching techniques to the setting of spatially-indexed or areal data, where unmeasured spatial confounders may impact the credibility of estimates under usual propensity score matching techniques (Papadogeorgou et al., 2019a; Davis, 2018). Other work has attempted to address the issue of matching in the presence of multi-level data where there may be unmeasured confounders at the cluster-level (also referred to as facility-level for the remainder) (Stuart and Rubin, 2008; Arpino and Mealli, 2011; Arpino and Cannas, 2016; Rickles and Seltzer, 2014). Recent research has also addressed similar issues in the context of weighting with propensity scores using inverse probability of treatment weighting (IPTW) (Davis et al., 2019; Li et al., 2013a). To our knowledge, existing research has not considered the setting with both spatial and multi-level confounding present in a matching framework.

Our motivating application examines patients with End Stage Renal Disease (ESRD) who have recently initiated regular dialysis at a dialysis facility, following work by Johansen et al. (2012). The patient's supervising physician at the facility is required to fill out a medical evidence form (CMS-2728) with important information on the patient's characteristics. Additionally, the supervising physician should indicate whether the patient was informed of their kidney transplant options or not, and if not, why not. A substantial proportion of patients are not informed of their transplant options because they were "not assessed". We note that this metric is not validated and there is evidence to suggest that it contains substantial measurement issues (see Salter et al. (2014) and the Discussion below).

Our interest lies in understanding the impact of being "not assessed" on 1-year referral for transplant. In estimating the treatment effect in the treated group, several important issues arise. First, the patients who are being informed and the patients who are not assessed may be different in substantively important ways: for example, initial health status may be important both for whether the patient was informed

soon after diagnosis as well as whether they received a referral for transplant within a year. Second, dialysis facilities can vary substantially from one another in terms of 1-year patient referral (Patzer et al., 2015) and on patients being informed of transplant options. Third, there may be geographic variation in patients being informed and their referral rates. Existing research has shown geographic and neighborhood variation on a number of measures such as pre-ESRD nephrology care (Yan et al., 2013; Hao et al., 2015; McClellan et al., 2009) and arteriovenous fistula (AVF) use (Hopson et al., 2008; McClellan et al., 2010). Patients may also face different underlying environmental exposures based on where they live (related to renal function) (Xu et al., 2018), and comparisons between patients may be strengthened by adjusting for these kinds of differences.

This article adapts and synthesizes recent work to the data setting where subjects are both clustered within facilities *and* spatially-indexed, with the purpose of estimating the average treatment effect on the treated (ATT). In these settings, the researcher may not have access to detailed confounders at the facility- and spatial-level, despite their perceived importance for treatment assignment and health outcomes. These unmeasured components present a challenge in generating credible estimates of treatment effects.

The methodology we propose here builds on recent work by Arpino and Cannas (2016) and Papadogeorgou et al. (2019a). First, we estimate a propensity score using observed covariates and spatial coordinates only. This model attempts to adjusts for potential spatial confounders solely through the inclusion of fixed effect terms for the projected latitude and longitude coordinates. Treated patients are then matched to control patients *only* within the same facility using this propensity score. This procedure supposes that the facility-level confounder is of primary importance as compared to the spatial-level confounder. Many treated subjects may not have viable matches within the same facility; for example, a facility may have had few patients,

or the vast majority of patients may have received the same treatment.

The second step of the proposed procedure uses the recently developed distance-adjusted propensity score matching (DAPSm) method (Papadogeorgou et al., 2019a) while adding random effects (intercepts) or fixed effects for the facility to the propensity score model to account for the facility-level unmeasured confounders (Arpino and Mealli, 2011). The algorithm then generates matches based on weighting propensity score differences and spatial distances between treated and controls. In this way, the method aims to match as many treated subjects as possible while attempting to adjust for both facility and spatial confounders at both stages. Even if a treated subject cannot be matched to a control within the same facility in the first stage, the procedure aims to match treated and controls if their estimated propensity scores are similar and their spatial distance is not too great. The novelty of this proposed approach lies in combining recent methods separately developed in the multi-level setting and the spatial setting for the data setting where both features are likely to be present.

We compare the statistical performance of our approach to other viable methods: for example, matching on the estimated propensity score which includes facility random or fixed effects along with spatial coordinates in a single step; using the DAPSm method in a single step; or using a two-stage method where the second stage matches allows for matches outside of the subject's facility with standard propensity score matching, as in the preferential within-cluster method proposed in Arpino and Cannas (2016). We compare approaches, including the proposed combined approach, in a simulation study under several different scenarios. Ultimately, we demonstrate that accounting for facility- and spatial-level confounding in a setting motivated by our data application is an important step towards better understanding treatment effect estimation.

1.2 Potential outcomes framework and setting

For subjects $i = 1, .., n$, let the binary treatment $Z_i = 1$ for the treatment group (not informed of transplant options because not assessed), and $Z_i = 0$ for the control group (informed of transplant options). Let $Y_i(Z)$ represent the potential outcomes under different treatment assignments, such that $Y_i(1)$ and $Y_i(0)$ represent the potential outcomes under treatment and control, respectively. For notational convenience, certain subscripts are dropped for the subsequent development (i.e. $Y_i(Z) = Y(Z)$) when there is no ambiguity. In the data application's primary outcome, Y represents the binary 1-year referral for transplantation measure.

The treatment effect is defined as $Y(1) - Y(0)$, but only one potential outcome is observed for any individual. Analyses of data, both randomized and observational, instead focus on *average* treatment effects. The estimand of interest for the motivating application is the average treatment effect on the treated: $ATT = E[Y(1) - Y(0)|Z = 1]$. This estimand requires the estimation of two pieces: $E[Y(1)|Z = 1]$ and $E[Y(0)|Z = 1]$. The former quantity can be estimated from the data directly following some basic assumptions, while the latter quantity requires additional assumptions and imputation, as we cannot know what the outcomes would be had treated individuals not been treated.

We highlight several key assumptions in propensity score studies. First, we assume the Stable Unit Treatment Value Assumption (SUTVA) holds, which states that a subject's potential outcomes do not depend on other subjects' treatment allocation (Stuart, 2010; Rubin, 1980). In multi-level and network applications, we may question whether the SUTVA actually holds. The purpose of this article is not to examine the SUTVA in detail, but we briefly discuss the issue in our discussion section with respect to the data application.

Following the notation of Papadogeorgou et al. (2019a), we consider a minimal set of confounders $\mathbf{C}$, which may contain individual confounders such as age and health

insurance status, facility confounders like quality of care, and spatial confounders such as neighborhood socioeconomic status. The assumption of ignorability states that:

$$Y(1), Y(0) \perp Z|\mathbf{C}. \tag{1.1}$$

This assumption is also referred to as the conditional independence assumption, selection on observables, or no unmeasured confounding (if $\boldsymbol{C}$ is observed). Rosenbaum and Rubin (1983) show that if there is no unmeasured confounding conditional on $\mathbf{C}$ (Equation 1.1), then there is also no unmeasured confounding conditional on the propensity score, defined as $P(Z = 1|\mathbf{C})$:

$$Y(1), Y(0) \perp Z|P(Z = 1|\mathbf{C}). \tag{1.2}$$

In order to proceed with propensity score approaches, we additionally assume overlap:

$$0 < P(Z = 1|\mathbf{C}) < 1. \tag{1.3}$$

The overlap assumption ensures that in propensity score matching approaches, every treated subject has a control subject with a similar propensity score. When interest is solely in estimating the ATT (as is the case here), the above assumptions can be weakened to be $Y(0) \perp Z|\mathbf{C}$ and $P(Z = 1|\mathbf{C}) < 1$ (Heckman et al., 1997; Caliendo and Kopeinig, 2008; Imbens, 2004). Under these assumptions, different approaches like propensity score weighting, matching, and sub-classification can be used to balance treated and control units on the observed characteristics to closely resemble a randomized control trial.

In many settings, we may not have the full set of minimum confounders $\mathbf{C}$. Suppose that $\mathbf{C} = (\mathbf{X}, V, U)$, where V and U are unmeasured scalars representing facility and spatial confounding, respectively, and $\mathbf{X}$ are the observed covariates. Then matching based on an estimate of $P(Z = 1|\mathbf{X})$ would no longer identify the ATT.

Although U and V are not measured directly, the spatial location and the facility assignment are often available to the researcher. The method we propose uses this information to reduce potential unmeasured confounding in a two-stage matching approach. This method synthesizes methods recently developed in the literature, which we detail in the next section.

In propensity score matching, treated and control pairs are formed based on the difference in the estimated (logit) propensity score. In this study, we restrict our attention to caliper 1:1 nearest-neighbor matching without replacement. For each treated subject, a greedy algorithm searches for the nearest matching control subject in a set of controls where the logit propensity score difference is less than some threshold value (i.e. the "caliper"), defined in terms of standard deviations of the logit propensity score in the entire sample (Austin, 2011b). As controls are selected for treated subjects, they are no longer considered for future treated subjects (i.e. matching is done "without replacement"). If successful, the resulting matched dataset from this process will have similar distributions of the propensity score and observed confounders. Other variants of this approach include optimal matching (as opposed to greedy nearest-neighbor matching) or matching with replacement, where controls can be matches for multiple treated subjects. We restrict our attention in order to maintain focus on the other properties of the proposed estimator in comparison with other methods and to simplify the number of varying factors. Austin (2014) found that nearest-neighbor matching did no worse than optimal matching methods, and that caliper methods without replacement did not do worse than methods with replacement.

1.3 Methodology for multi-level and spatial data

We propose a two-stage matching algorithm in order to adjust for potential confounders at the spatial- and facility-level. This method assumes that the facility confounder is of greater importance than the spatial confounder based on a prior belief about the data application. To motivate the two-stage approach, recall that $\mathbf{C} = (\mathbf{X}, V, U)$. In this setting, we assume that there each subject belongs to one of $j = 1, ..., J$ facilities, and that each individual has a pair of spatial coordinates $\mathbf{s} = (s_1, s_2)$.

For treated subject k and control k', if one could match exactly on observed covariates $\mathbf{X}$ *and* the facility assignment, this would ensure that $V_k = V_{k'}$ even though V is unmeasured. Furthermore, if the researcher could ensure that the pair of treated and control subjects had the same location of exposure to U such that $\mathbf{s}_k = \mathbf{s}_{k'}$, then we could also ensure that $U_k = U_{k'}$, despite U being unmeasured. Following Papadogeorgou et al. (2019a), since individuals will not be exactly in the same spatial location, we instead assume that there exists some δ_ϵ for all $\epsilon > 0$ such that if $|s_k - s_{k'}| < \delta_\epsilon$ then $|U(s_k) - U(s_{k'})| < \epsilon$. Matching on $P(Z = 1|\mathbf{X})$, facility assignment, and spatial location simultaneously would likely result in a large portion of the treated sample being unmatched, due to the curse of dimensionality and due to some facilities having few patients or having patients predominantly in one treatment group or another. For this reason, we introduce a staged matching approach that attempts to somewhat adjust for facility and spatial-level confounding while retaining more of the treated patients. We examine the resulting number of unmatched treated patients within our approach in detail in the simulation study and application below.

1.3.1 Stage 1: Matching within facility

At the first stage, we propose estimating the propensity score with observed covariates $\mathbf{X}$ and the spatial location $\mathbf{s} = (s_1, s_2)$, using logistic regression for all of the available data:

$$logit[P(Z = 1|\mathbf{X}, \mathbf{s})] = \alpha_0 + \alpha_{\mathbf{X}}{}^T\mathbf{X} + \alpha_{s_1}s_1 + \alpha_{s_2}s_2 + \alpha_{s_{12}}s_1s_2 + \alpha_{s_1^2}s_1^2 + \alpha_{s_2^2}s_2^2. \quad (1.4)$$

Matching then proceeds using 1:1 nearest-neighbor caliper matching on the logit propensity score, using 0.2 standard deviations in the sample as the caliper width (Austin, 2011b), and restricting matches to occur in the same facility. Note that we do not include any facility indicators in the propensity score in this first stage, as matching will only occur within facilities. In this framework, the result is that V will be balanced despite not being measured.

Spatial coordinates (s_1, s_2) together with their interactions are included as fixed effect terms in the logistic regression model in order to capture basic spatial patterns in treatment assignment in the observed data. Other functional forms based on spatial coordinates could also be substituted. Thus, U may be balanced if the true spatial pattern of treatment assignment is captured by this basic formulation. If the true spatial pattern cannot be captured in this way, the method will balance V (more important) but do less well in balancing U (less important) across treated and controls. The matching thus reflects the researcher's belief in the perceived importance of different kinds of confounding.

1.3.2 Stage 2: Distance adjusted propensity score matching with facility effects

The first stage may leave many treated patients unmatched, depending on the study. Analyses based on sub-samples may not be generalizable, thus we proceed with a second stage of matching. As with the first stage, the goal is still to generate balance on observed covariates $\mathbf{X}$ while also attempting to balance for unobserved confounders V and U at the facility- and spatial-level.

Papadogeorgou et al. (2019a) propose a method called distance adjusted propensity score matching (DAPSm) that matches treated units with control units based on a combination of estimated propensity score difference and spatial distance. Define the estimated propensity score for unit i as $\text{PS}_i = P(Z_i = 1|\mathbf{X}_i)$, and define the Euclidean distance between two units i and k as Dist_{ik}. The distances between treated and controls are then standardized to range from 0 to 1 in the sample to match the range of propensity score differences. The distance-adjusted propensity score (DAPS) is then defined as:

$$\text{DAPS}_{ik} = w \times |\text{PS}_i - \text{PS}_k| + (1 - w) \times \text{Dist}_{ik}, \tag{1.5}$$

where the weight $w \in [0, 1]$ controls how much to weight spatial distance relative to the propensity score difference. Setting the weight $w = 1$ will mean that the DAPS for a pair of treated and control patients is entirely equal to the propensity score difference, and $w = 0$ implies that the DAPS is equal to the spatial distance only. Papadogeorgou et al. (2019a) discuss details on selecting a weight and an algorithm for selecting matches. In their simulation studies, they find their method has good performance compared to alternative methods in terms of MSE compared to the gold-standard propensity score matching approach. Although their method is developed to use point-referenced data, we use areal-level centroid coordinates in our

data application.

We use the DAPS for matching with particular propensity score formulations: we estimate propensity scores with either **random effects** (random intercept for the facility):

$$logit[P(Z_i = 1|\mathbf{X}_i)] = \alpha_0 + \alpha_{\mathbf{X}}^{\mathbf{T}}\mathbf{X}_i + \gamma_{j[i]} \tag{1.6}$$

where $\gamma_{j[i]} \sim N(0, \sigma^2)$, or **fixed effects**:

$$logit[P(Z_i = 1|\mathbf{X}_i)] = \alpha_0 + \alpha_{\mathbf{X}}^{\mathbf{T}}\mathbf{X}_i + \sum_{l=1}^{J-1} \alpha_{fac,l} I(l = j[i]) \tag{1.7}$$

where $I(\cdot)$ is an indicator function and $j[i]$ indicates the facility assignment for unit i.

These propensity score models are estimated based on the full data in order to most accurately estimate random or fixed effects and covariate effects on the propensity of receiving treatment. Matching is done by using the DAPSm technique described above on the remaining treated and control subjects not matched in stage 1 by finding matches sequentially, starting from the smallest DAPS score. For this approach, a caliper could be specified for either the DAPS score itself, or solely on the propensity score (PS) difference, as in stage 1. If a caliper is specified on the PS difference, then regardless of the weight w specified, matches will still not exceed the caliper-specified difference on the estimated PS.

The final output of these two stages is a combined set of matches between treated and controls, where either facility is matched exactly and space is adjusted for through fixed effect terms for $\mathbf{s}$ in the propensity score model, or the facility is adjusted for through a random/fixed effect in the propensity score and spatial distance is accounted for using the DAPS, a weighted sum of spatial distance and propensity score difference. Further adjustments can also follow if the researcher believes there are interactions between V and $\mathbf{X}$ in the propensity score model; for example, if one

believes facilities systematically treat insured patients differently from uninsured patients, matches between treated and control can be forced to have the same insurance status in stage 1, or random slopes that vary by facility can be introduced in stage 2.

We implement this procedure in R (R Core Team, 2020), building upon the `DAPSm` package (Papadogeorgou et al., 2019b).

1.3.3 Other methods

An earlier two stage approach to matching was proposed in Arpino and Cannas (2016) which considered a single-level propensity score for both steps, and found superior performance compared to alternatives. Arpino and Mealli (2011) and Arpino and Cannas (2016) also evaluated fixed effect and random effect propensity score models and found they both improved upon the propensity score matching approach without any facility adjustment.

Stuart and Rubin (2008) consider another approach in the context of a study of schools, with a treated and control group, where only observed subject characteristics need to be controlled for. Because of incomplete matches provided by the original control group, additional matches are drawn from a secondary control group, which may differ in cluster-level covariates. Their method then adjusts for potential cluster-level biases introduced from matches made with the secondary control group. Rickles and Seltzer (2014) extend this approach to a multisite study and consider a two-stage matching approach, where treated units that are not matched to controls within the same cluster in the first stage are then matched to controls in a different but similar cluster, as determined by some baseline cluster-level covariates. They adjust any systematic difference in outcomes that results from matching to a different cluster, and they estimate the ATT both within and across all clusters.

Li et al. (2013a) considers multi-level data in the context of a weighting approach, where random intercepts and fixed effects for facilities are considered in the propensity

score and outcome models in a doubly-robust approach. They generally found that ignoring the cluster in both the propensity score and outcome model led to much higher bias, but that the outcome model specification was more impactful than the propensity score model specification. Davis et al. (2019) consider a similar approach with conditional autoregressive (CAR) random effects for area-level data to adjust for potential geographic confounding, and Davis (2018) similarly considers a matching approach in the presence of geographic confounding. The approaches of Davis (2018), Davis et al. (2019), and Li et al. (2013a) could potentially be extended to the setting where patients are in small areas and facilities by including both spatial and facility random effects in the propensity score model.

1.4 Simulation study

1.4.1 Data generation

To illustrate and assess performance of our proposed two-stage approach, we consider the following simulation study. Our data generation emphasizes a stronger facility-level association with the treatment assignment mechanism and outcome, with an unobserved spatial covariate being present but of secondary importance. We aim to generate a reproducible set of simulations that display facility- and spatial-level confounders resembling potential data applications. We proceed by first generating the patient and facilities and the assignment of patients to facilities in two separate settings (distance-based vs. random); second, we specify the propensity score and outcome models based on observed and unobserved covariates in two scenarios (continuous or binary outcome).

For all of the simulations, one of two background datasets is used to determine where patients and facilities are located, and how patients are assigned to facilities. We used 2010 Census data on block groups in the state of Georgia for the simula-

tion; N = 1000 patients were sampled from these block groups using a multinomial distribution with probabilities equal to the block group share of the state population. J = 80 facilities similarly were sampled, but only from block groups with above average populations. Both patients and facilities were assigned to population-weighted centroids (projected coordinates) of their respective block groups, and the coordinates were then randomly jittered so that locations were unique for all patients and facilities.

Two background datasets were generated that assigned patients to facilities differently:

- **Setting 1: Distance-based**. Distance is strongly related to facility assignment (but random), to resemble many data applications. An exponential function is used to generate probabilities of facility assignment for each patient, and then a multinomial function is used to randomly assign the facility to the patient.

- **Setting 2: Random**. For each patient, facility assignment is based on a multinomial probability distribution where each facility is given equal probability. This setting is considered so that the spatial confounder and facility confounder are independent of each other.

Figure A.1 in the Appendix demonstrates visually the patient assignment to facility in these two settings. Data generation then proceeds as follows:

1. Four standard normal covariates are generated for each patient: $X_1, X_2, X_3, X_4 \sim N(0, 1)$.

2. A spatial covariate U is generated using a Normal distribution with a Matern covariance function, where smoothness and range parameters follow one of 4 pairs: $(1.46, 1)$, $(1.46, 0.1)$, $(0.1, 1)$, $(0.1, 0.1)$. U is then standardized. Figure A.2 in the Appendix illustrates the impact of these smoothness/range parameters on the spatial covariate's realization.

3. A facility covariate is generated from a normal distribution: $V \sim N(0, 2)$.

4. Treatment assignment is based on probabilities of treatment generated by the following (true) propensity score model:

$$logit[P(Z = 1|\mathbf{X}, V, U)] = -2.3 + 0.5V + 0.3U + 0.2X_1 + 0.3X_2 - 0.2X_3 - 0.3X_4 \quad (1.8)$$

The steps here closely resemble the setting of Papadogeorgou et al. (2019a) with some deviation in the precise parameters chosen in the propensity score model. The model above generates roughly 11% treated patients in the sample, comparable to what we see in our data application. The models emphasize the importance of the facility over the spatial unmeasured covariate, but both are important for both the treatment assignment and outcome.

Our simulations show results for both the continuous and binary outcome case. For a continuous outcome, we use the following model similar to Papadogeorgou et al. (2019a), where the ATT is 1:

$$Y = Z + 2V + 0.5U + 0.55X_1 + 0.21X_2 + 1.17X_3 - 0.11X_4 + \epsilon,$$

where $\epsilon \sim N(0, 1)$.

For the binary outcome, incidence is sampled from a Bernoulli distribution where the probability of incidence is specified through a logistic regression model:

$$logit[P(Y = 1|\mathbf{X}, U, V)] = -2 + 0.29Z + 0.5V + 0.3U + 0.2X_1 + 0.3X_2 - 0.2X_3 - 0.3X_4.$$

Parameters were chosen such that the ATT is approximately 0.05, and if no one in the sample were treated, the outcome incidence would be approximately 15%, using the methodology of Austin (2010, 2014). $K = 500$ datasets are generated for each case. The settings and shorthand are summarized in Table 1.1.

Abbreviation	Outcome Type	Facility Assignment	True ATT
S1-A through S1-D	Continuous	Distance-Based	1.0
S2-A through S2-D	Continuous	Random	1.0
S3-A through S3-D	Binary	Distance-Based	≈ 0.05
S4-A through S4-D	Binary	Random	≈ 0.05

Table 1.1: Simulation Settings
A through D reflect the different Matern smoothness/range combinations for U
A: (1.46, 1), B: (1.46, 0.1), C: (0.1, 1), D: (0.1, 0.1)

1.4.2 Methods compared

We consider separately the propensity score estimation and the matching algorithm. The propensity score models considered in simulations are as follows:

1. **Single-level (S)**: A propensity score model with only observed covariates and spatial coordinates, with no fixed effect or random effect for the facility (single-level), as in (1.4).

2. **Random Effects (RE)**: A propensity score model that adds random intercept terms for the facilities to the single-level model.

3. **Fixed Effects (FE)**: A propensity score model that adds fixed effects for the facilities to the single-level model.

Depending on the use case, the above formulations may include or exclude the spatial coordinate terms at the estimation stage. In particular, when using DAPSm method, the method itself takes into account distance between treated and control units, so the propensity score method used will exclude the spatial coordinates. The matching methods considered (in conjunction with the propensity score models above) are as follows:

1. **Propensity Score Matching (PS)**: Matching on the propensity score model in a single step. We consider all 3 propensity score models (**S, RE, FE**) above with spatial coordinate terms included for this method. Arpino and Mealli

(2011) and Arpino and Cannas (2016) include these approaches in their simulation study.

2. **Within-cluster (WC)**: This method limits matches between treated and control to the same facility. This method may discard and ignore entire facilities with small numbers of patients. This method will be based on the single-level propensity score model with spatial coordinates included. Arpino and Mealli (2011) assessed this approach.

3. **Preferential within-cluster (WC+)**: Broadly based on the method from Arpino and Cannas (2016) which generates matches within the same facility when possible (subject to caliper), and otherwise pulls matches outside of the facility. Their method considers matching with replacement, but we consider matching without replacement here as it is more typical in biostatistical applications and based on the recommendation of simulation studies (Austin, 2014). The first stage uses the single-level propensity score (S), but the second stage may be any of the 3 propensity score methods (all with spatial coordinates included): **WC+S, WC+RE, WC+FE**.

4. **DAPSm**: Using the DAPSm approach in Papadogeorgou et al. (2019a), which matches based on balancing distance and the estimated propensity score. All 3 propensity score models are considered here, with the spatial coordinates excluded from the estimation of the propensity score: **DAPSm-S, DAPSm-RE, DAPSm-FE**.

5. **Preferential within-cluster + DAPSm (WC+DAPSm)**: A new approach described above combining the two previous methods that proceeds in two steps: (a) Use the single-level propensity score with spatial coordinates to generate matched pairs within the same facility; (b) For the remaining unmatched treated subjects, use the DAPSm method in conjunction with each of the 3

propensity score estimates (excluding spatial coordinates): **WC+DAPSm-S, WC+DAPSm-RE, WC+DAPSm-FE**.

6. **Gold Standard**: Matching using the propensity score that includes unobserved covariates U and V.

Greedy nearest-neighbor matching without replacement is used for all estimation methods. A caliper of 0.2 standard deviations of the logit propensity score is used for all methods, except for DAPSm, where the caliper is based on the propensity score instead of the logit propensity score. Details on fitting parameters are found in the Appendix. Results for the DAPSm method presented in the main text are based on choosing the smallest weight that balances the observed covariates. Alternative approaches, including a constant weight approach and a weight-search method using the DAPS caliper (rather than the PS caliper), are presented in the Appendix.

1.4.3 Results

We highlight select results from the simulation study. Figure 1.1 summarizes standardized differences between treated and controls across simulations for select covariates and simulation settings in the first scenario. An absolute value of 0.1 is often cited in judging there to be no meaningful difference between two groups (Austin, 2011a). Figure 1.1(a) and 1.1(b) show balance for the unobserved spatial covariate U in the high-smoothness, long-range setting (S1-A) and the low-smoothness, short-range setting (S1-D), respectively. Similar to Papadogeorgou et al. (2019a), in the latter setting, few methods are able to balance U well, including the DAPSm method.

In the high-smoothness, long-range setting, most methods appear to do a fair job of balancing U, especially when compared to the case with no adjustments made. In part, this is likely because the spatial trend can be mostly captured by including some simple fixed effect terms for the projected longitude/latitude in the propensity score

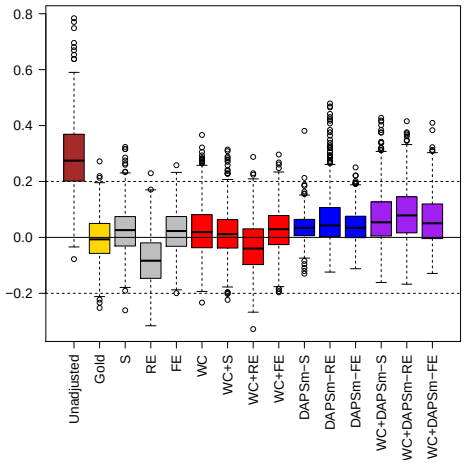
model. The proposed two-stage approach actually does slightly worse in balancing U on average as compared to the preferential within-cluster methods without the use of DAPSm (WC+S, WC+RE, WC+FE). This result stems from prioritizing balance on the facility covariate V over U through the use of the propensity score caliper in the second stage rather than imposing a caliper on the DAPS or distance. In the Appendix, Figure A.4(a) and A.4(b) demonstrate that the two-stage approach that uses the DAPS caliper for the second stage instead of the propensity score caliper does substantially better in balancing U, at some cost to balance on V. For covariate X_4, all methods offer substantial improvements compared to the unadjusted sample; the proposed two-stage approaches do about as well as any other approach tested.

Table 1.2 summarizes all simulation results in terms of average estimates, relative mean-squared error (compared to the gold-standard propensity score matching model), and the proportion of treated subjects successfully matched with each method. For the continuous outcome settings, average estimates in the two-stage proposed method with the fixed effect propensity score model (WC+DAPSm-FE) are the least biased (after the gold-standard approach) in a majority of the 8 settings. The within-cluster only method also does well in terms of bias, but at the cost of matching a portion of the treated sample (59% to 62% of treated subjects depending on the simulation setting). The preferential within-cluster method with the fixed effects propensity score model without the use of the DAPSm approach (WC+FE) also does consistently well across settings in terms of bias. Relative MSE results point to similar conclusions, favoring WC+DAPSm-FE in several settings while matching around 94% to 95% of the treated subjects, resulting in more efficient and generalizable estimates.

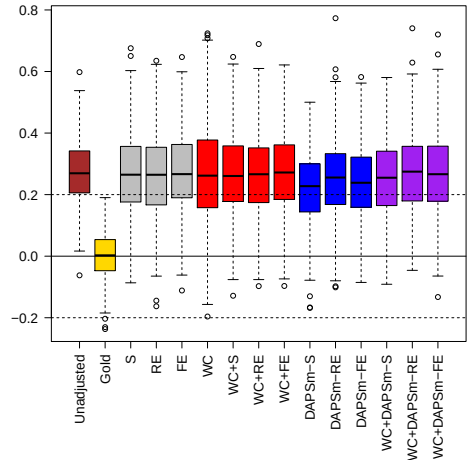
The binary outcome results in the lower half of Table 1.2 demonstrate more mixed results with no clear best method. Most methods with either random effects or fixed effects, in either a single-stage or two-stage model, perform similarly in terms of bias.

The proposed two-stage approach with fixed effects or random effects does consistently well across settings, although it is not the clear top performer. The within-cluster only method does well in terms of bias but has a substantially higher relative MSE as compared to some of the other approaches tested.

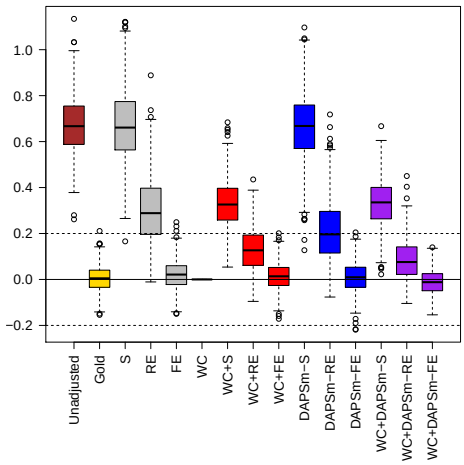
Additional DAPSm and WC+DAPSm methods with different parameter choices results appear in the Appendix. In the Appendix, we also consider the setting with fewer facilities ($J = 40$ vs. $J = 80$) to assess the potential impact on the results. We find that a larger proportion of subjects can be matched in the within-cluster only method, but that otherwise the same conclusions hold regarding the performance of the proposed two-stage estimator.



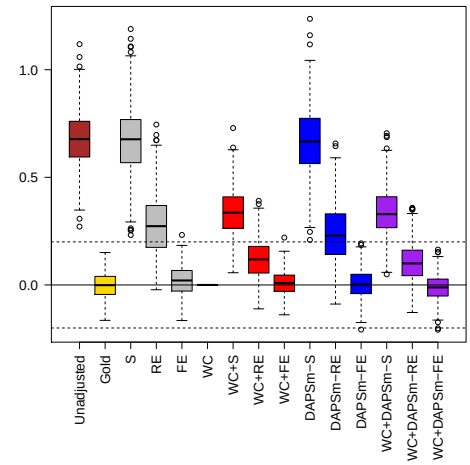
(a) U: S1-A (Smoothness 1.46, Range 1)



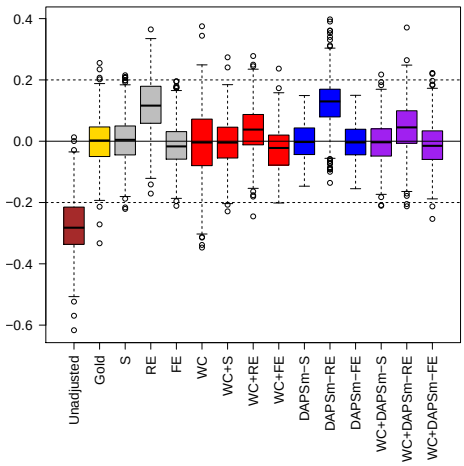
(b) U: S1-D (Smoothness 0.1, Range 0.1)



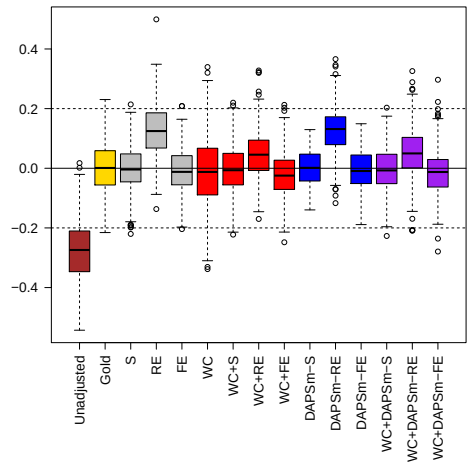
(c) V: S1-A



(d) V: S1-D



(e) X_4: S1-A



(f) X_4: S1-D

Figure 1.1: Standardized difference for U, V, and X_4 in distance-based setting (S1-A and S1-D)

Table 1.2: Simulation results: mean ATT estimate, relative MSE, and average proportion of treated matched

Continuous	Mean ATT Estimate (True = 1)								Relative MSE								Proportion treated subjects matched							
Method	1-A	1-B	1-C	1-D	2-A	2-B	2-C	2-D	1-A	1-B	1-C	1-D	2-A	2-B	2-C	2-D	1-A	1-B	1-C	1-D	2-A	2-B	2-C	2-D
Gold	1.01	1.00	1.01	1.00	1.01	1.02	1.00	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.98	0.97	0.98	0.98	0.98	0.98	0.97
S	2.81	2.85	2.91	2.92	2.85	2.91	2.99	3.01	75.87	79.43	70.32	80.41	76.64	80.35	81.21	81.37	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
RE	1.77	1.82	1.87	1.88	1.76	1.83	1.87	1.87	16.59	17.98	16.38	19.08	15.10	17.35	18.16	17.38	0.95	0.96	0.96	0.96	0.95	0.96	0.95	0.95
FE	1.07	1.11	1.17	1.19	1.04	1.13	1.16	1.19	1.27	1.41	1.71	2.03	1.10	1.60	1.64	1.94	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC	1.02	1.06	1.14	1.13	1.00	1.09	1.12	1.13	1.24	1.33	1.63	1.56	1.28	1.58	1.57	1.57	0.62	0.62	0.61	0.61	0.59	0.59	0.59	0.59
WC+S	1.90	1.93	2.02	2.03	1.99	2.03	2.09	2.11	20.36	21.68	20.95	24.32	23.34	24.71	25.91	26.13	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC+RE	1.33	1.40	1.46	1.45	1.35	1.43	1.45	1.45	4.24	5.03	5.31	5.97	4.21	5.59	5.62	5.53	0.96	0.96	0.96	0.96	0.95	0.96	0.96	0.95
WC+FE	1.05	1.08	1.15	1.15	1.02	1.10	1.13	1.15	1.05	1.23	1.49	1.53	0.88	1.40	1.33	1.52	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
DAPSm-S	2.82	2.77	2.87	2.88	2.88	2.86	2.98	2.98	76.07	73.47	67.06	76.73	79.53	76.13	81.35	79.05	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
DAPSm-RE	1.61	1.68	1.75	1.78	1.76	1.76	1.82	1.80	11.23	13.83	13.17	15.26	15.65	15.12	16.27	15.15	0.94	0.95	0.95	0.95	0.95	0.95	0.94	0.94
DAPSm-FE	1.05	1.06	1.13	1.14	1.05	1.08	1.11	1.14	1.43	1.15	1.56	1.57	1.24	1.24	1.44	1.64	0.94	0.94	0.94	0.94	0.95	0.95	0.94	0.94
WC+DAPSm-S	1.94	1.95	2.03	2.03	2.01	2.01	2.09	2.11	21.42	22.50	21.44	24.32	23.98	23.56	26.02	25.91	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC+DAPSm-RE	1.28	1.34	1.41	1.41	1.36	1.40	1.44	1.43	3.47	4.41	4.71	5.10	4.70	5.14	5.24	5.28	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
WC+DAPSm-FE	1.00	1.05	1.11	1.10	1.02	1.08	1.10	1.10	1.12	1.19	1.27	1.26	1.11	1.31	1.27	1.31	0.94	0.95	0.94	0.95	0.95	0.95	0.95	0.95
Binary	Mean ATT Estimate (x100) (True ≈ 5)								Relative MSE								Proportion treated subjects matched							
Method	3-A	3-B	3-C	3-D	4-A	4-B	4-C	4-D	3-A	3-B	3-C	3-D	4-A	4-B	4-C	4-D	3-A	3-B	3-C	3-D	4-A	4-B	4-C	4-D
Gold	5.0	5.0	5.0	4.9	5.3	5.2	4.7	5.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.98	0.97	0.98	0.97	0.98
S	10.9	12.2	12.3	12.4	12.0	12.3	12.5	12.6	2.25	2.66	2.75	3.05	2.37	2.85	2.90	3.15	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
RE	6.2	6.3	7.4	7.1	6.0	6.7	6.9	6.9	1.06	1.00	1.24	1.22	1.06	1.13	1.12	1.22	0.96	0.96	0.96	0.96	0.95	0.96	0.96	0.96
FE	5.6	6.0	6.4	6.7	6.0	6.3	6.4	6.7	1.01	0.94	1.05	1.16	0.95	1.01	1.03	1.12	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC	5.1	5.4	5.9	5.6	5.5	5.6	5.4	5.9	1.56	1.61	1.45	1.61	1.37	1.55	1.51	1.77	0.62	0.61	0.61	0.61	0.58	0.59	0.59	0.59
WC+S	8.7	9.3	9.3	9.4	9.5	9.7	9.4	9.8	1.43	1.62	1.53	1.73	1.51	1.75	1.68	1.89	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC+RE	5.7	6.0	6.6	6.6	6.0	6.4	6.2	6.5	0.99	1.07	1.07	1.15	0.96	1.08	0.99	1.24	0.96	0.96	0.96	0.96	0.95	0.96	0.96	0.96
WC+FE	5.7	6.2	6.5	6.5	6.3	6.4	6.2	6.8	0.99	1.06	1.04	1.16	0.91	1.05	1.04	1.17	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
DAPSm-S	11.9	11.8	12.1	12.0	12.4	12.2	12.3	12.3	2.56	2.60	2.57	2.95	2.49	2.77	2.74	3.03	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
DAPSm-RE	5.6	6.2	6.4	6.5	6.5	6.5	6.6	6.6	1.08	1.00	1.10	1.21	1.03	1.08	1.08	1.10	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95
DAPSm-FE	5.3	5.5	5.8	6.3	6.2	5.6	5.8	5.9	1.01	1.04	0.94	1.07	0.93	1.02	1.04	0.95	0.94	0.94	0.94	0.94	0.94	0.95	0.94	0.94
WC+DAPSm-S	9.1	9.4	9.5	9.5	9.8	9.7	9.6	9.7	1.54	1.74	1.60	1.77	1.61	1.78	1.67	1.87	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC+DAPSm-RE	5.5	6.1	6.3	6.2	6.7	6.5	6.3	6.4	1.00	1.03	1.09	1.09	1.01	1.13	1.01	1.17	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
WC+DAPSm-FE	5.3	5.7	5.9	6.0	6.3	6.1	5.8	6.2	1.01	1.05	0.96	1.12	0.91	1.08	1.05	1.18	0.94	0.95	0.95	0.94	0.95	0.95	0.95	0.95

1.5 Application

1.5.1 Analysis

We apply the methods considered in the simulation study along with the proposed method of combining within-cluster matching with the DAPSm method to a study of the 2017 US Renal Data System (USRDS) (United States Renal Data System, 2017) together with data on referrals for kidney transplants collected as part of the *R*educing *D*isparities *I*n *A*ccess to kid*N*ey *T*ransplantation (RaDIANT) Community Study (Patzer et al., 2014, 2017). The focus of our application is to better understand the relationship between informing incident End Stage Renal Disease (ESRD) patients starting dialysis about transplant options and referrals for transplants within one year. In this setting, the facility and (to a lesser extent) the area of residence may play important roles in patient's being informed and the patient's referral status after one year. Comparisons between patients who were not assessed (treated) and those who were informed (controls) should take these factors into account, along with individual patient characteristics.

At the time of being diagnosed with ESRD, a Medical Evidence Report (CMS-2728) is filled out by the physician at the dialysis facility. Among the information included is whether the patient was informed of kidney transplant options. The question asked is: "Has the patient been informed of kidney transplant options?". If the answer is no, the follow-up question is "If patient NOT informed of transplant options, please check all that apply." Options include "Medically unfit", "Patient declines information", "Unsuitable due to age", "Patient has not been assessed", "Psychologically unfit", and "Other".

Johansen et al. (2012) considers the association of race and insurance type with delayed assessment for transplantation among patients starting dialysis using this question. The study of Johansen et al. (2012) is focused on "Patient has not been

assessed." In their study, comparisons were made between "Not assessed" and all other patients: those who were informed and those who were not informed for other reasons. For this application, we focus on the comparison of patients who were "Not assessed" (and in no other category) and "Informed"; we focus on a single state (Georgia) in a 2-year period from 2012-2013.

The primary outcome of interest is 1-year referral, but we additionally examine 1-year waitlisting and 1-year mortality to aid in conclusions and understanding the quality of the matching approaches used. Patients are considered referred within 1 year if a referral was recorded at a transplant center within a year of the first ESRD service date and prior to death. Additionally, we include the additional cases as referred within 1 year: (1) waitlisting date within 1 year but with no recorded referral date, (2) waitlisting date within 1 year but prior to the recorded referral date, and (3) no referral or waitlisting date but a transplant date within 1 year of first ESRD service date. To summarize, a patient is considered referred within 1 year if they were referred, waitlisted, or received a transplant within 1 year of the first ESRD date and the event occurred prior to death.

We note that the relevant question from the medical evidence form (CMS-2728) used to determine "treatment" for this study is not validated and may be prone to serious misclassification. Salter et al. (2014) provide some evidence based on a sample of $n = 388$, finding a large portion of patients who were reported as informed did not themselves report being informed, and a large portion of patients who were not reported as informed *did* report being informed. In considering the relationship between the CMS-2728 question and 1-year referral for transplant in our study, we can consider our study as an additional assessment of measurement error in this measure. If the recorded question does not signify anything about actual 1-year referral, one possible interpretation is that the question may not accurately capture whether patients were actually informed or not. We note this measurement error issue

as a serious limitation for the interpretability of the results; our primary purpose in presenting these results are to demonstrate the use of the proposed methods in conjunction with existing methods for better understanding the data. We caution against strong causal conclusions from this analysis.

The 2017 USRDS contains information on 3,081,768 patients in the patient standard analytic file. The final analytical sample then consists of 4,906 patients who resided in and initiated dialysis in Georgia from 2012 to 2013. Details on the exact inclusion criteria for the study are provided in the Appendix.

In the analytical sample, 36.9% had a one-year referral in this sample, and 9.7% were not assessed for their transplant options by the dialysis facility. Covariates are largely taken from the 2728 form, including race, ethnicity, age, sex, incidence year, BMI, glomerular filtration rate (CKD-EPI method), co-morbidities, pre-ESRD care, medical coverage, employment status, facility numbers of patients, social workers and registered nurses, and county-level covariates. ZIP code centroids of the patients' addresses at the first ESRD service date are used to capture spatial location.

In the main text, we present results for the DAPSm methods using a propensity score caliper of 0.2 standard deviations. The weight for the DAPSm methods is chosen based on minimizing imbalance for observed covariates, resulting in a weight of 0.05 for the DAPSm-S, DAPSm-RE, and DAPSm-FE methods using a threshold of 0.15 in absolute standardized differences. For the WC+DAPSm methods, a weight of 0.5 that balanced the propensity score difference and distance equally was used. For all other methods, a 0.2 standard deviation of the logit propensity score is used. Additional assessments of sensitivity to various tuning parameters appear in the Appendix, including the use of a DAPS caliper instead of a propensity score caliper.

Table 1.3: Summary of covariates across treated and control patients

Variable	Informed (n = 4428) Mean (SD)	Not Assessed (n = 478) Mean (SD)	Standardized difference
Incidence Year = 2013	0.49 (0.50)	0.52 (0.50)	0.06
Incidence Age	58.19 (13.53)	57.46 (13.20)	-0.06
Female	0.46 (0.50)	0.43 (0.50)	-0.05
Race			
White	0.38 (0.48)	0.27 (0.45)	-0.23
Black	0.60 (0.49)	0.72 (0.45)	0.26
Other	0.02 (0.15)	0.01 (0.09)	-0.15
Hispanic	0.03 (0.16)	0.02 (0.14)	-0.03
Log(BMI)	3.38 (0.26)	3.39 (0.26)	0.04
Height	169.63 (11.69)	169.70 (11.27)	0.01
Log(Weight)	4.43 (0.28)	4.44 (0.27)	0.04
Log(GFR-EPI)	2.09 (0.56)	2.00 (0.61)	-0.14
Hemodialysis	0.90 (0.30)	0.97 (0.16)	0.44
Access type			
AVF	0.12 (0.33)	0.11 (0.32)	-0.03
Graft	0.02 (0.15)	0.04 (0.20)	0.08
Catheter	0.75 (0.43)	0.82 (0.38)	0.18
Other/NA	0.10 (0.30)	0.03 (0.16)	-0.47
Diabetes			
Insulin	0.43 (0.50)	0.42 (0.49)	-0.02
Oral	0.10 (0.29)	0.11 (0.32)	0.05
No Meds	0.06 (0.23)	0.07 (0.26)	0.07

Table 1.3: Summary of covariates across treated and control patients

Variable	Informed (n = 4428) Mean (SD)	Not Assessed (n = 478) Mean (SD)	Standardized difference
Retinopathy	0.06 (0.24)	0.04 (0.19)	-0.14
Hypertension	0.90 (0.30)	0.91 (0.29)	0.03
ASHD	0.08 (0.27)	0.06 (0.25)	-0.06
Congestive heart failure	0.26 (0.44)	0.27 (0.44)	0.03
Other Cardiac	0.16 (0.37)	0.12 (0.32)	-0.12
Peripheral vascular disease	0.07 (0.26)	0.08 (0.27)	0.02
Amputation	0.03 (0.17)	0.04 (0.20)	0.06
CVA, TIA	0.08 (0.27)	0.10 (0.29)	0.06
Inability to ambulate	0.04 (0.20)	0.06 (0.23)	0.06
Inability to transfer	0.02 (0.14)	0.03 (0.16)	0.03
Need assistance	0.09 (0.29)	0.09 (0.29)	-0.01
Institutionalized	0.04 (0.20)	0.05 (0.23)	0.06
Drug and Alcohol Dependence			
Alcohol	0.02 (0.12)	0.03 (0.17)	0.09
Drug	0.01 (0.11)	0.02 (0.14)	0.07
Tobacco	0.09 (0.29)	0.07 (0.26)	-0.08
COPD	0.08 (0.27)	0.06 (0.24)	-0.06
Cancer	0.05 (0.22)	0.05 (0.21)	-0.03
Toxic nephropathy	0.01 (0.07)	0.00 (0.05)	-0.08
No co-morbidities	0.02 (0.14)	0.02 (0.14)	0.02
Employment status			
Unemployed/Med LOA/Other	0.36 (0.48)	0.51 (0.50)	0.30
Ret-age/Ret-disability	0.52 (0.50)	0.41 (0.49)	-0.24

Table 1.3: Summary of covariates across treated and control patients

Variable	Informed (n = 4428) Mean (SD)	Not Assessed (n = 478) Mean (SD)	Standardized difference
Employed/Student/Homemaker	0.12 (0.32)	0.09 (0.28)	-0.12
Insurance			
Group	0.22 (0.41)	0.14 (0.35)	-0.21
Medicaid	0.24 (0.43)	0.30 (0.46)	0.13
Medicare	0.57 (0.50)	0.50 (0.50)	-0.13
Other	0.10 (0.30)	0.12 (0.33)	0.07
VA	0.02 (0.14)	0.02 (0.13)	-0.03
Medicare Adv.	0.08 (0.28)	0.07 (0.25)	-0.07
Primary Cause			
Diabetes	0.43 (0.50)	0.41 (0.49)	-0.04
Glomerulonephritis	0.06 (0.24)	0.03 (0.18)	-0.15
Hypertension	0.41 (0.49)	0.46 (0.50)	0.09
Other	0.08 (0.27)	0.08 (0.26)	-0.01
Unknown	0.02 (0.14)	0.02 (0.15)	0.02
Pre-ESRD Nephrology Care			
No	0.31 (0.46)	0.32 (0.47)	0.03
Unknown	0.09 (0.29)	0.24 (0.42)	0.34
Yes	0.60 (0.49)	0.44 (0.50)	-0.32
EPO			
No	0.62 (0.48)	0.53 (0.50)	-0.19
Unknown	0.26 (0.44)	0.41 (0.49)	0.31
Yes	0.12 (0.32)	0.06 (0.24)	-0.24
Pre-ESRD Dietary Care			

Table 1.3: Summary of covariates across treated and control patients

Variable	Informed (n = 4428) Mean (SD)	Not Assessed (n = 478) Mean (SD)	Standardized difference
No	0.77 (0.42)	0.62 (0.48)	-0.31
Unknown	0.18 (0.38)	0.33 (0.47)	0.33
Yes	0.05 (0.22)	0.05 (0.21)	-0.02
Hemoglobin			
< 10 g/dL	0.55 (0.50)	0.60 (0.49)	0.09
>= 10 g/dL	0.29 (0.46)	0.27 (0.44)	-0.05
(Missing)	0.15 (0.36)	0.13 (0.34)	-0.07
Serum Albumin			
< 3.5 g/dL	0.52 (0.50)	0.56 (0.50)	0.08
>= 3.5 g/dL	0.25 (0.43)	0.23 (0.42)	-0.05
(Missing)	0.24 (0.43)	0.22 (0.41)	-0.05
County			
Pct. white non-Hispanic	53.26 (17.67)	51.16 (17.66)	-0.12
Household poverty	18.23 (5.54)	20.11 (5.82)	0.32
Log(Pop.)	11.80 (1.39)	11.53 (1.59)	-0.17
Facility			
Facility - Hospital Based	0.04 (0.19)	0.02 (0.13)	-0.18
For-profit	0.88 (0.33)	0.83 (0.38)	-0.13
Non-profit	0.12 (0.33)	0.17 (0.37)	0.12
Profit status unknown	0.00 (0.05)	0.01 (0.08)	0.05
FTE RN	3.72 (2.56)	3.43 (2.25)	-0.13
FTE SW	0.83 (0.48)	0.84 (0.51)	0.02
FTE RN / Patients	0.05 (0.02)	0.05 (0.02)	0.04

Table 1.3: Summary of covariates across treated and control patients

Variable	Informed (n = 4428) Mean (SD)	Not Assessed (n = 478) Mean (SD)	Standardized difference
FTE SW / Patients	0.01 (0.01)	0.01 (0.01)	0.07
# patients end of year	79.03 (42.85)	72.76 (38.13)	-0.16
# patients start of year	74.53 (40.36)	70.78 (36.89)	-0.10

1.5.2 Results

Overlap

Table 1.3 summarizes the covariates used in the propensity score model for the analytical sample for treated and controls. Notably, the "not assessed" group is more likely to be Black and more likely to be on hemodialysis. Other notable differences include insurance, employment status, and pre-ESRD care. The propensity score overlap in the full, un-matched sample is shown in Figure 1.3 (a) through (c). There is largely overlap between the Not Assessed (treated) and Informed (control) groups in the single-level logit propensity score (a), which includes individual covariates, ZIP code coordinates, county, and facility covariates. However, when fixed effects for the facility are included in the propensity score model, the results change markedly in Figure 1.3(b). There are clear areas where there is no overlap – these are facilities with all patients being informed of their transplant options, or none being informed. The propensity score model that includes a random intercept for the facility yields a distribution of scores in between the single-level and fixed effects model, but there is still a substantial area of non-overlap, suggesting there will be treated units who

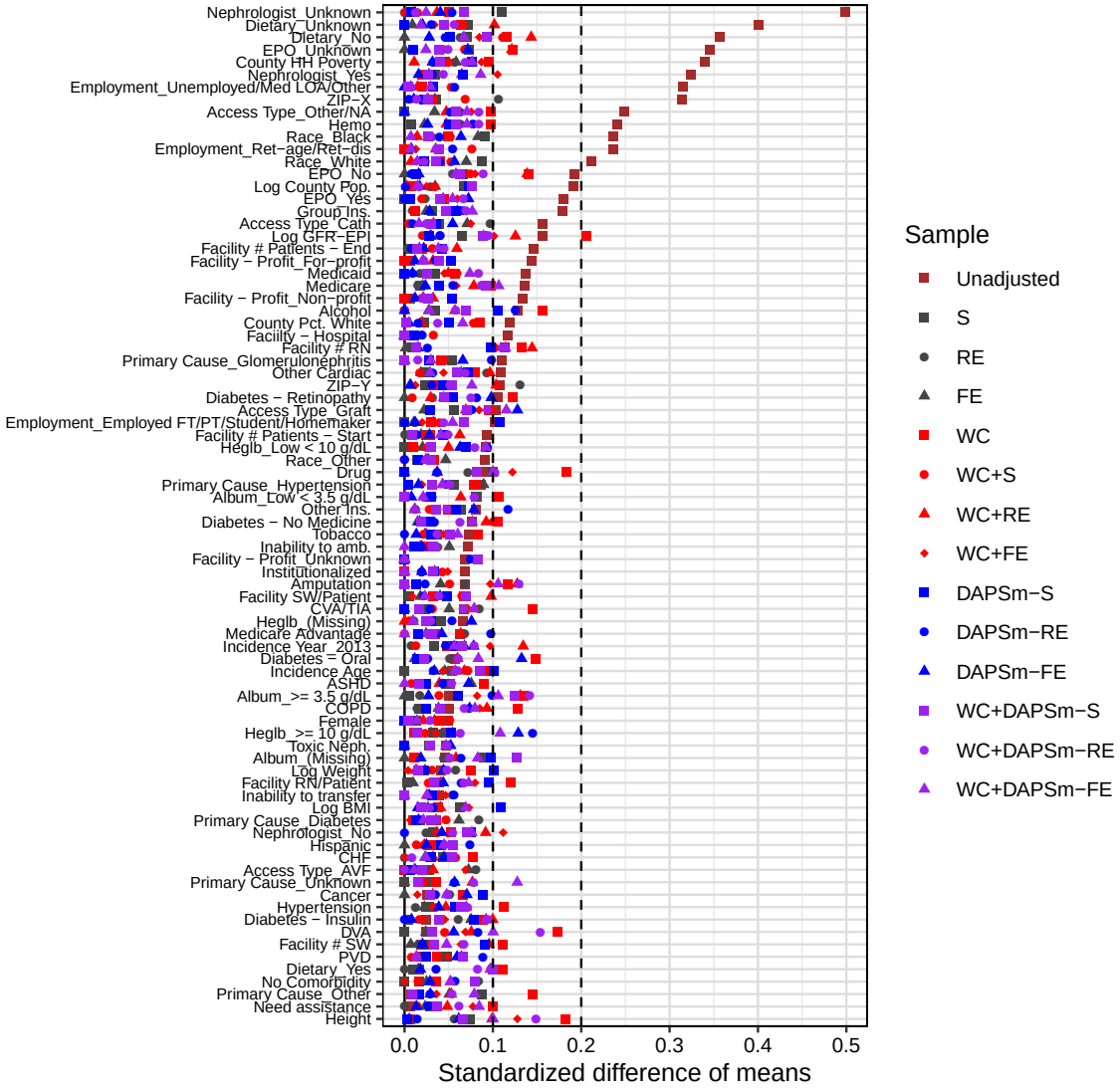
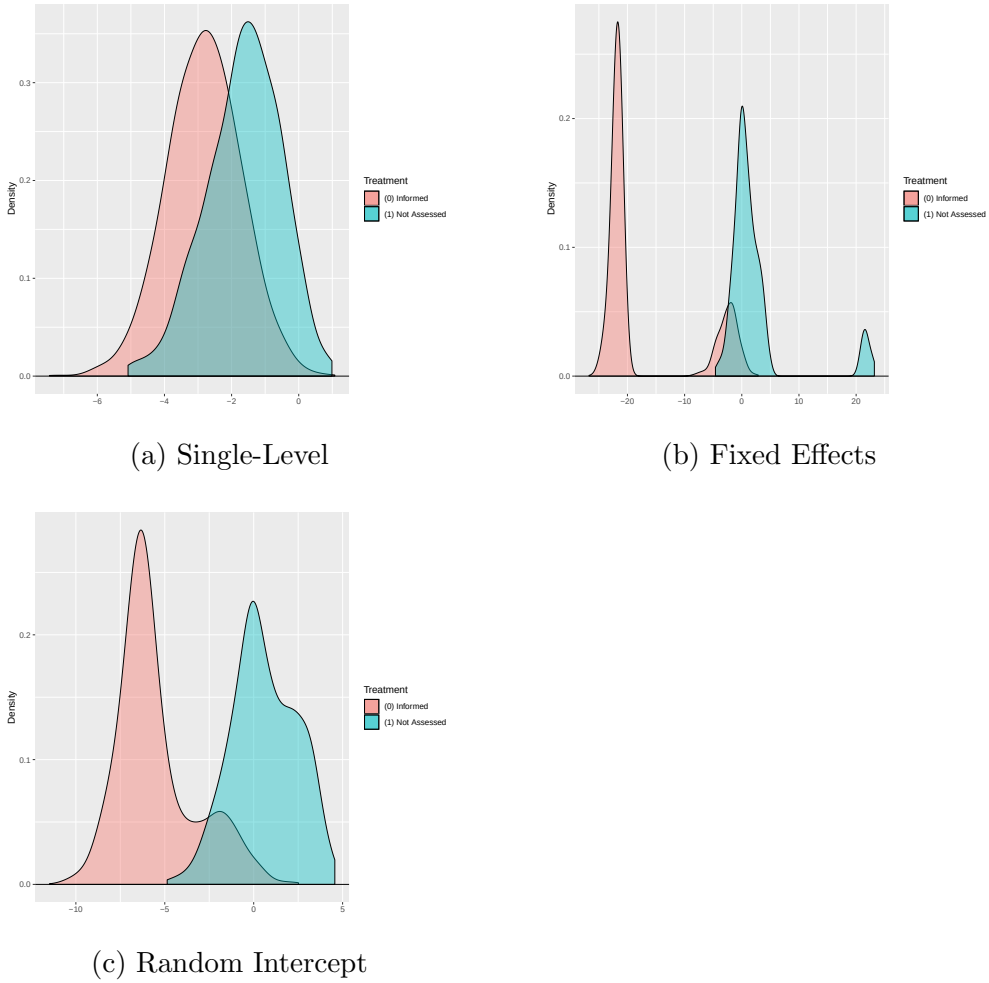


Figure 1.2: Absolute standardized differences for all methods

Standardized differences are absolute differences in means divided by standard deviation of the covariate among the treated.

cannot be matched to any controls.

For patients in facilities with only treated or only control subjects, the estimated propensity scores are very close to 0 and 1, particularly when fixed effects are included. Because a caliper is used in matching, the matched samples with random and fixed effects only include the middle area of overlap between treated and control patient propensity scores. This changes the meaning of the estimand given that the matched



(a) Single-Level

(b) Fixed Effects

(c) Random Intercept

Figure 1.3: Distributional balance for logit propensity scores
Models displayed include individual, county and facility covariates along with ZIP x/y projected coordinates. Blue density curve represents treated/"Not Assessed". Fixed effects model excludes facility indicator for Hospital vs. Free-Standing due to convergence issues.

sample does not include all of the treated subjects.

These results suggest that facility is very strongly related with treatment assignment. Matching that uses the random effects or fixed effects model will be limited to the area of common support. If facility assignment is believed to be important for referrals, it is then difficult to distinguish between the effects of the treatment of interest and the facility for a substantial portion of the sample. Ignoring facility indicators in either the fixed effects or random effects model would only be justifiable if one believed the facility was unassociated with the outcome (one-year referral)

conditional on all of the other individual and facility-level covariates included in the propensity score model (Stuart, 2010). Thus we are left with two options that show the limits of propensity score matching in this scenario: (1) match most of the sample while excluding the facility ID (either through fixed indicators or random effects), or (2) matching on a substantially smaller fraction of the sample where there is overlap while taking into account the facility ID. We present various methods in this vein to assess how balance and outcome estimates are sensitive to this choice, but we recognize immediately that propensity score matching will limit the conclusions we can draw given this lack of overlap.

Balance

Figure 1.2 demonstrates balance across the covariates included in the propensity score model for all methods, sorted by imbalance in the unadjusted sample, emphasizing absolute standardized differences greater than 0.2 and 0.1. In total, 10 covariates are greater than the 0.2 threshold in absolute standardized differences. All methods reduce this number to 0, but the within-cluster only method, perhaps due to matching only 43% of the sample, still leaves a substantial number of covariates imbalanced using the 0.1 threshold (Table 1.4). The proposed two-stage approach with either a random effect or fixed effect included in the propensity score model (WC+DAPSm-RE and WC+DAPSm-FE) gets the proportion of treated subjects matched to 58% and 59%, respectively, and they reduce the number of imbalanced covariates compared to the within-cluster only method. However, these methods are still intrinsically limited by the lack of overlap of the propensity score distributions between the treated and control patients. The approaches using the single-level propensity score match nearly the entire sample and produce good balance on the observed covariates – but there is clearly a questionable assumption being made about the role of dialysis facilities. Normally, propensity score models would be adjusted if there were still

lingering imbalance on observed covariates, but we wish to compare methods based upon broadly similar propensity score models in this exploratory analysis.

Outcome and Interpretation

Figure 1.4(a) and Table 1.4 show estimates and 95% confidence intervals from differences in proportions in the matched samples for 1-year referral. Confidence intervals are constructed from two-sample difference in proportions with no adjustment for the matched nature of the data. The unadjusted result shows essentially no effect from being recorded as not assessed vs. informed of transplant options, receiving referrals within 1-year at roughly the same rate. These estimates vary across methods but are substantively small, ranging from -1.52 to 2.34 – the variability across methods and size of the uncertainty is too great to make any conclusions about the impact of not being assessed for transplant referrals within 1-year. As mentioned previously, the estimated effects may point to the recorded treatment being an unreliable metric for whether patients are actually informed of their options, thus providing little signal for one-year referral. Figure 1.4(b) summarizes estimates for 1-year waitlisting. Among all patients in the analytical sample, 6.7% were waitlisted within a year. In contrast to 1-year referral, where the effects across matching methods are generally small and close to zero, the 1-year waitlisting results suggest mostly small and consistently negative results. Again, these results may suggest that CMS-2728 is not accurately recording whether patients are truly being informed of their transplant options or not.

Finally, 1.4(c) suggests that there may be lingering confounding present after matching with the various methods. In particular, we would not expect that being informed of transplant options at the start of ESRD would have an effect on 1-year mortality – the mechanism is likely to operate much more slowly, by allowing people to obtain transplants over the course of several years more quickly than persons who

were not informed when starting dialysis. Although 1-year mortality is not a *pre-treatment* measure, substantial deviations between the treatment and control groups in this measure may suggest that there are unobserved differences in the two treatment groups, where one group is healthier than the other at the start of ESRD after matching. In our analytical sample, there is an overall 1-year mortality proportion of 13.3%, with the "not assessed" group having a 1-year mortality proportion that is 1.56% lower than the "informed" group before matching. Several of the approaches result in a matched samples where the not assessed (treatment) group has a higher 1-year mortality proportion than the informed (control) group.

In addition to assessing differences in proportions between treated and control patients, we look at linear probability models that adjust for any imbalanced covariates over 0.1 absolute standardized differences after matching in Figure A.6(b) for 1-year referral. While linear probability models are not well suited for modeling binary outcome data, the coefficient on the treatment effect is more readily interpretable as a risk difference. Including covariates that may remain imbalanced after matching follows the suggestions of Ho et al. (2007), where matching is treated as a pre-processing approach before using a parametric model. Figure A.6(c) shows similar model adjustments in a logistic regression; a drawback of this approach is that conditional and marginal odds ratios need not be the same (Austin, 2007). These additional results largely show the same patterns in effect estimates. Finally, in the Appendix in Figure A.7 and Table A.8, we also present results from Cox proportional hazards models for time-to-waitlisting and time-to-referral for 1-year of follow-up. Estimates of the cause-specific hazard ratio are presented, where death acts as a competing risk for our outcome of interest (Austin et al., 2016). Again, substantive conclusions here do not change.

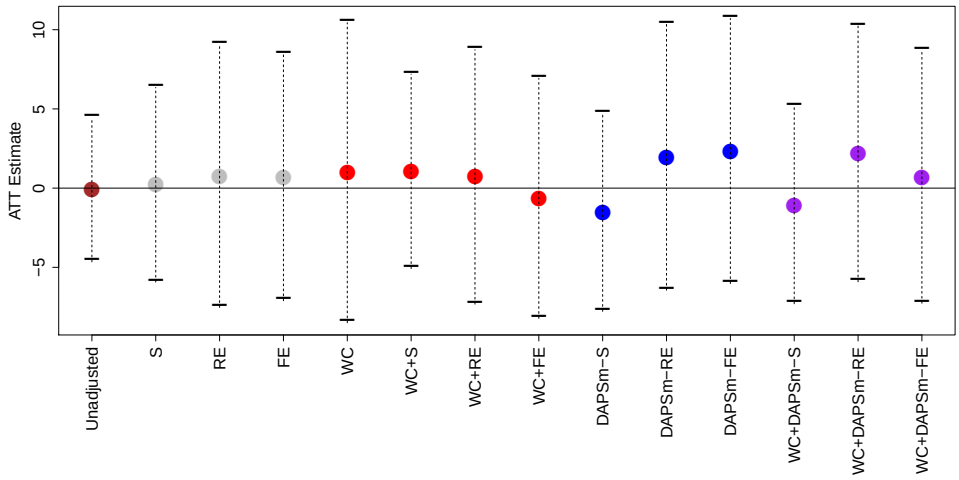
Because the DAPSm methods may be sensitive to the choice of weight or the type of caliper chosen, the Appendix presents results for both the propensity score caliper

as well as the DAPS caliper for many weights ranging from 0 to 1, demonstrating some variability in the effect estimates (Figures A.8-A.11). In particular, when using a propensity score caliper on the second DAPSm stage of the proposed two-stage approach (Figure A.10), the estimates did not vary greatly based on the DAPSM weight chosen. However, using the DAPS caliper (Figure A.11) produced a wider range of estimates, with a small negative effect estimated when the weight was close to 0 (strongly favoring spatial distance vs. the propensity score difference in calculating DAPS). The single-stage DAPSm methods with the DAPS caliper similarly showed small negative effect estimates for very small weights. Differences in estimates suggest some sensitivity to the parameters chosen, but in all cases, the uncertainty of the estimates is large, and our substantive conclusions do not change.

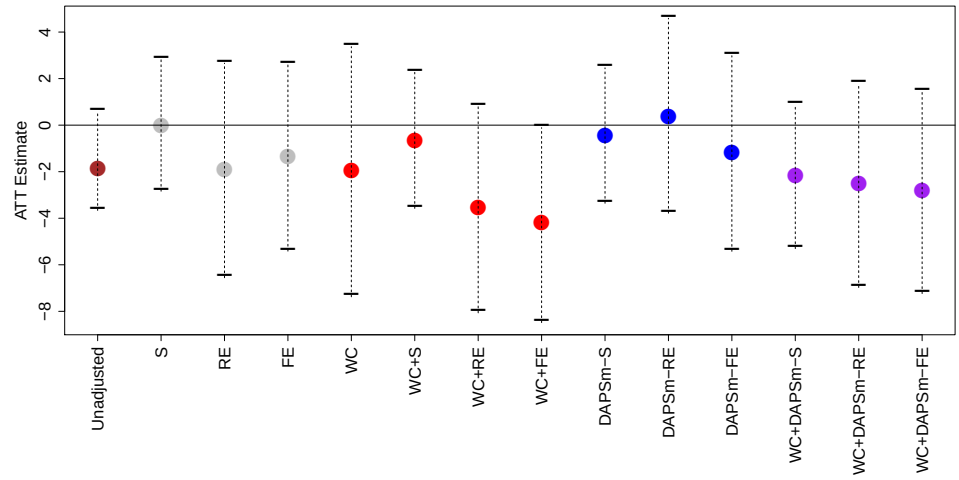
	Estimate of 1-year outcome (95% CI)			Pct. treated	ASD Summary		
Method	Referral	Waitlisting	Death	matched	> 0.10	> 0.15	> 0.20
Unadjusted	-0.10 (-4.66, 4.45)	-1.84 (-3.66, 0.59)	-1.56 (-4.37, 1.79)	100.00	24	13	10
S	0.21 (-5.93, 6.36)	0.00 (-2.84, 2.84)	1.69 (-2.36, 5.78)	98.74	0	0	0
RE	0.76 (-7.53, 9.05)	-1.89 (-6.56, 2.67)	-0.38 (-5.78, 5.01)	55.23	5	0	0
FE	0.68 (-7.11, 8.47)	-1.36 (-5.41, 2.61)	1.69 (-3.55, 6.98)	61.72	0	0	0
WC	0.97 (-8.50, 10.44)	-1.94 (-7.38, 3.37)	0.97 (-4.89, 6.88)	43.10	20	7	0
WC+S	1.06 (-5.08, 7.20)	-0.64 (-3.58, 2.27)	1.06 (-3.02, 5.16)	98.54	3	0	0
WC+RE	0.71 (-7.32, 8.74)	-3.53 (-8.07, 0.82)	2.12 (-3.02, 7.31)	59.21	12	1	0
WC+FE	-0.65 (-8.22, 6.93)	-4.19 (-8.48, -0.11)	3.55 (-1.37, 8.55)	64.85	6	0	0
DAPSm-S	-1.52 (-7.78, 4.73)	-0.43 (-3.36, 2.46)	-1.09 (-5.37, 3.19)	96.23	8	0	0
DAPSm-RE	1.95 (-6.44, 10.33)	0.39 (-3.78, 4.59)	4.28 (-0.92, 9.63)	53.77	7	0	0
DAPSm-FE	2.34 (-6.02, 10.71)	-1.17 (-5.42, 2.98)	1.56 (-3.93, 7.10)	53.56	3	0	0
WC+DAPSm-S	-1.08 (-7.30, 5.14)	-2.16 (-5.31, 0.89)	-0.22 (-4.44, 4.01)	96.86	4	0	0
WC+DAPSm-RE	2.16 (-5.91, 10.23)	-2.52 (-6.98, 1.80)	2.52 (-2.73, 7.84)	58.16	7	4	0
WC+DAPSm-FE	0.70 (-7.27, 8.67)	-2.82 (-7.24, 1.45)	2.82 (-2.29, 8.00)	59.41	10	0	0

Table 1.4: Estimates for 1-year outcomes, percent of treated patients matched, and summary of imbalance across various methods.

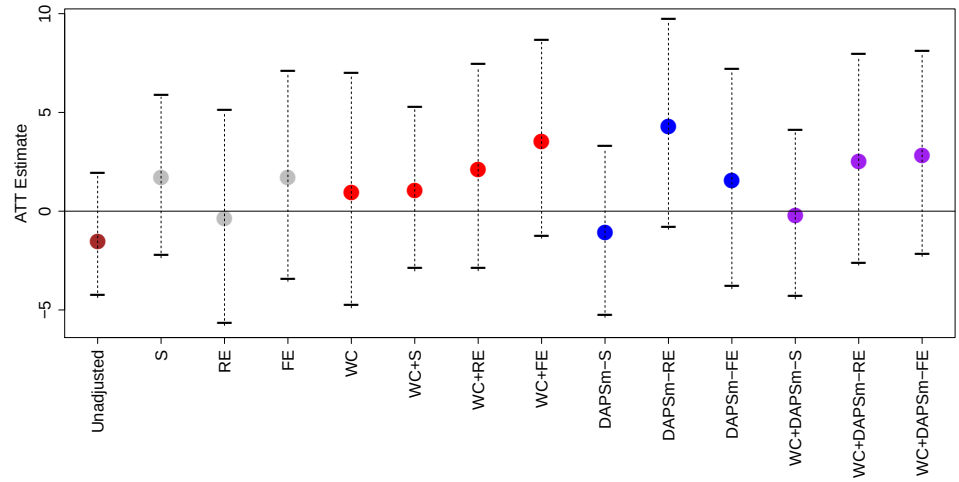
Variable imbalance is summarized by number of variables with absolute standardized difference (ASD) between treated and control greater than various thresholds. Confidence intervals for 1-year waitlisting and 1-year mortality are based on the score interval (Mee and Anbar, 1984; Miettinen and Nurminen, 1985), as implemented in the PropCIs R package (Scherer, 2018).



(a) Outcome: 1-year referral



(b) Outcome: 1-year waitlisting



(c) Outcome: 1-year mortality

Figure 1.4: Estimates of ATT: proportion difference between not assessed (treated) and informed (control)

1.6 Discussion

Applications with individual-level health data with unmeasured potential confounders at the facility- and spatial-level pose a challenge for standard propensity score matching techniques. We propose a two-stage approach that synthesizes recently developed methods that leverage matching within facilities in the first stage and spatial distance together with the propensity score in the second stage to obtain matches that may more plausibly adjust for unobserved confounders at the facility- and spatial-level. Comparing estimates across a variety of approaches, using propensity score models with and without fixed or random effects in the proposed two-stage approach, along with existing methods, can provide a useful assessment of the variability of results.

Our simulation study demonstrated that the two-stage approach generally did as well as competing methods by attempting to adjust for potential confounders at the spatial- and facility-level. With a continuous outcome, the proposed method with fixed effects often had superior performance in terms of bias and MSE. With a binary outcome, most methods that adjusted for the facility through a fixed or random effect appeared to do well on various metrics, and our proposed method did not do substantially better or worse than these competing methods. We also demonstrated that in the setting where the facility confounder is likely to be of primary importance and the spatial confounder of secondary importance, most methods that adjust for the facility, either through within-cluster matching or through the inclusion of a random or fixed effect, greatly reduce bias and MSE relative to methods that use a single-level propensity score with no matching on the facility.

In the data application to incident dialysis patients in 2012-2013 in Georgia, plausible estimation of the causal effect is limited by the following factors: (1) the facility and question about informing patients about transplant options are difficult to separate; (2) due to a substantial lack of overlap, taking into account the facility through random or fixed effects results in a substantially smaller sample size that is no longer

generalizable to the full treated population; (3) the study does not likely have sufficient power to estimate a meaningful difference between the treated and control populations; (4) differences in 1-year mortality suggest that matching may result in treated and controls who may still differ in important ways; (5) for reasons laid out in Salter et al. (2014), the treatment (not assessed vs. informed) is likely subject to substantial reporting error, and this reporting error may vary systematically by facility, further complicating the validity of the analysis. In a sample of 388 patients, Salter et al. (2014) find that in 27.8% of the patients, the provider reported informing the patient about options when the patient did not report being informed. Furthermore, patient-reported informed status and not provider-reported informed status was significantly associated with an increased likelihood of waitlisting. These issues offer several opportunities for further analyses and new research strategies. An additional limitation is that our analysis can only include patients who have reached ESRD and started dialysis; our analysis cannot include those who have chronic kidney disease but who have not yet started dialysis, nor can we account for those who may have died prior to ESRD.

Numerous issues are present with using CMS-2728 data as an accurate comparison of patients. Eggers (2010) notes that the physician may not have access to a patient's medical history when filling out the form; alternatively, the form may be filled out by an administrative assistant. Our analysis can be interpreted in this context: if the recorded question accurately measured whether patients were being informed about their transplant options, then we would expect there to be some statistical signal and a meaningful estimated effect on referral for transplant after adjusting for patient, facility, and spatial confounders. Since we see no effect across all methods we attempted, we might conclude either that the measure has considerable measurement error, or that the measure is not an important factor for 1-year referral. We also may doubt whether any method here was able to appropriately fully adjust for con-

founding. Finally, our results only hold for patients in Georgia who meet our various inclusion criteria in 2012-2013. More research should be conducted to understand the CMS-2728 measure and better measuring patient health at the start of ESRD for comparison purposes.

Nonetheless, the use of a variety of different matching techniques, and including the fixed or random effects in the propensity score, provides insight into the lack of overlap and the difficulty in separating treatment effect from facility assignment that might otherwise be ignored. We additionally acknowledge that there may be substantial unobserved *individual* confounders not included in the analysis due to the limited number of variables on the medical evidence form and the potential for mismeasurement. In particular, few laboratory measurements of patients that would indicate health status are available for the full sample, and socio-economic status is not determined for individuals directly. Reported ZIP code centroids may also not capture the relevant spatial scale for the unobserved spatial confounders. We also focused on area-level covariates related to the patients' residence, but it may also be important to consider the area-level covariates of the dialysis facility.

Our study also assumes SUTVA holds. In applications where spatial and facility-level factors are at play, there may be reason to expect that treatments assigned to one patient impact another patient in the same area or facility (interference). Future research should examine how to incorporate interference into such analyses; examples of spatial causal analyses that consider interference include Keele and Titiunik (2018), Zigler et al. (2012), and Verbitsky-Savitz and Raudenbush (2012). Papadogeorgou et al. (2019b) considers interference and clusters with an application to air pollution data.

There are other considerations for considering facility-level random or fixed effects. Suppose that that we match individuals within the same facility, but that the facility distinguishes between patients based on some individual covariate (unobserved or

observed). We may not fully capture the nature of facility confounding through the use of a fixed or random effect in this case. For example, a particular facility may systematically treat uninsured patients differently from insured patients, or it may base its decision for informing patients on some unobserved health status that is not well captured by the observed covariates. Matching two individuals who have different insurance statuses or health statuses within a facility may lead to an erroneous conclusion in this case. Nevertheless, one may still believe on average the facility impact on treatment assignment can be captured by the inclusion of a random or fixed effect. Other approaches for considering facility-level confounding include one proposed by Zubizarreta et al. (2012).

While our simulations provide insight into performance of several propensity score matching approaches, more research is needed for developing these methods further while explicitly incorporating both facility and spatial factors. Future research should also address measurement error in the treatment variable, with or without validation data, and consider facility-level variation in measurement error. Future work can also consider instances where the spatial confounding may be considered to be of greater importance than the facility.

1.7 Acknowledgments and disclaimer

The data utilized in this study was funded in part by the National Institute on Minority Health and Health Disparities award U01MD010611. Support for the preparation of this document was provided by contract number HHSM-500-2013-NW006C from ESRD Network 6, funded by the Centers for Medicare & Medicaid Services (CMS) — an agency of the US Department of Health and Human Services. The conclusions presented are solely those of the authors and do not represent those of Southeastern Kidney Transplant Coalition or the CMS. The content of this publication does not

necessarily reflect the policies or positions of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. The authors assume full responsibility for the accuracy and completeness of the ideas presented.

This chapter fulfills USRDS privacy requirements (manuscript #MS2020-54). **USRDS Disclaimer**: A portion of the data reported here have been supplied by the United States Renal Data System (USRDS). The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the U.S. government.

Chapter 2

Imputing satellite-derived aerosol optical depth using a multi-resolution spatial model and random forest for $PM_{2.5}$ prediction

2.1 Introduction

Ambient outdoor air pollution, particularly particulate matter less than 2.5 micrometers in aerodynamic diameter ($PM_{2.5}$), poses a substantial risk to human health (WHO, 2018; Lim et al., 2012; Lelieveld et al., 2015). Air pollution monitors that can directly measure pollution concentrations are placed at a limited set of locations, resulting in large areas without direct measurements of ground-level pollution exposure. Aerosol optical depth (AOD) measures the amount of aerosol in the atmosphere and can be remotely sensed by satellite instruments at various spatial resolutions (Levy et al., 2013). A growing literature has developed for using satellite-derived AOD as a proxy and predictor for $PM_{2.5}$ concentrations, often in conjunction with land-use and

meteorological variables, using a range of model types such as geographically weighted regression, linear mixed effect models, and machine learning methods (Sorek-Hamer et al., 2016; Chu et al., 2016; Shin et al., 2020).

However, AOD itself has substantial missingness, complicating the process of predicting $PM_{2.5}$ concentrations. Gaps in AOD coverage are a result of cloud cover, snow cover, and surface brightness; for the Moderate Resolution Imaging Spectroradiometer (MODIS) 10km product, on average each grid cell has no AOD available on approximately 70% of days, with substantial variation across regions (Belle and Liu, 2016; Belle et al., 2017). AOD's patterns of missingness are also not random for the purpose of $PM_{2.5}$ prediction; cloud and snow cover may plausibly be related to $PM_{2.5}$ concentrations (Bi et al., 2019; Belle et al., 2017). At the scale of the continental United States, research suggests that missingness as a result of cloud cover is not likely to greatly bias monthly and yearly $PM_{2.5}$, although there is regional and seasonal variation (Christopher and Gupta, 2010). However, Liang et al. (2020) show that long-term $PM_{2.5}$ estimates in China are substantially biased as result of missing AOD observations. Furthermore, for health effects research, the relevant geographic scale is small and more impacted by missingness. When using $PM_{2.5}$ estimates based on satellite-derived AOD with substantial missingness, time series studies will miss many days and lose statistical power, and cohort studies will use potentially biased exposure estimates, resulting in a loss of statistical power. A number of approaches have been proposed for handling missing AOD observations when estimating $PM_{2.5}$ (Shin et al., 2020). One approach has been to combine different AOD retrievals, although this will still result in incomplete coverage; e.g., Geng et al. (2018) combines AOD measurements from Terra and Aqua satellite using linear regression. Other approaches have used AOD where available, but otherwise bypassed the need for gap-filling AOD (Kloog et al., 2011, 2012; Lee et al., 2016).

Many recent studies use multi-stage approaches, where AOD is gap-filled, and

then a model relating $PM_{2.5}$ to the gap-filled AOD and other land-use and meteorological variables is fit. These gap-filling models may use land-use and meteorological terms, as well as chemical transport model (CTM) estimates. Hu et al. (2017) forego a statistical modeling procedure for gap-filling AOD, saving computational time, and they replace missing AOD values with CTM (GEOS-Chem) estimates of AOD. Xiao et al. (2017) and Huang et al. (2018) use linear models that include cloud fraction estimates, meteorological and land-use data together with smoothing splines to account for spatial correlation for imputing AOD. Lv et al. (2016) gap-fill AOD using a model that relates the ratio of daily and seasonal averages of $PM_{2.5}$ to seasonal AOD values for a grid cell for each city under study; a second stage then uses ordinary Kriging to fill in remaining gaps. Because of the computational costs of ordinary Kriging, this method will not scale well to large datasets, but previous studies suggest that smoothing splines may not perform as well as Kriging in some settings (Laslett, 1994). Chen et al. (2019) use a mixed effect model to first combine Terra and Aqua AOD measurements, and interpolate missing AOD values using inverse-distance weighting (IDW). IDW with a maximum distance will not be able to provide full coverage for AOD, however, as there are large missing areas with no observed data. Random forest (RF) is arguably the most popular machine learning method used for gap-filling, due to the fast implementations available and its ability to account for complex non-linear interactions of features (Breiman, 2001; Shin et al., 2020). Bi et al. (2019) uses a two-stage model with RF being used to impute AOD using a number of relevant variables, including MODIS cloud and snow fractions. Stafoggia et al. (2019) and Zhang et al. (2018) also impute AOD as part of a multi-stage process using RF. However, judging performance based on "out-of-bag" measures or random holdouts of observed data may be misleading in spatial prediction problems with large contiguous areas of missing data. Furthermore, when a strong spatial pattern is present as in AOD, it is unclear how RF performs compared to spatial statistical models.

Importantly, models for gap-filling AOD are generally more costly to fit than models for estimating $PM_{2.5}$ due to the much larger number of daily observations. For example, in our case study using a modeling grid of 12km spatial resolution over the contiguous United States entails over 50,000 daily cells. While several studies have used machine learning methods for AOD gap-filling to overcome the computational costs, traditional spatial statistical methods like Kriging are not well-suited to handle large datasets due to the need to invert the spatial covariance matrix. Over the course of the last decade or so, several spatial statistical methods have been developed to handle big data (Heaton et al., 2019; Bradley et al., 2016). Although considerable attention has been given to using ensemble and hybrid approaches for estimating $PM_{2.5}$ (e.g., Shao et al. (2020); Xiao et al. (2018); Di et al. (2019); Murray et al. (2019)), AOD gap-filling for large areas has received less focus, possibly due to the greater computational cost.

To our knowledge, studies have not thus far considered ensemble methods for combining large-scale spatial statistical methods with machine learning methods for gap-filling AOD. In this study, we focus on a particular spatial statistical method, lattice kriging (LK) (Nychka et al., 2015), together with RF for AOD gap-filling. Our study considers both RF and LK models for gap-filling MODIS AOD, as well as ensemble methods for combining these predictions following the super learner methodology (Van der Laan et al., 2007; Naimi and Balzer, 2018). Our case study focuses on the contiguous United States using daily data for the month of July 2011. We focus on a single month for computational reasons as the AOD models are fit daily. We assess performance using spatially clustered holdouts for AOD gap-filling models that may more accurately measure performance than more commonly used approaches. Finally, we assess whether the imputed AOD product using ensemble methods improves $PM_{2.5}$ estimation in a random forest model. Broadly, we find that ensemble methods can be effective for AOD gap-filling, but there is less evidence to suggest

an ultimate benefit for $PM_{2.5}$ estimation. In Section 2.2, we describe the motivating data. In Section 2.3, we discuss briefly the lattice kriging and random forest methods, as well as super learner methodology for combining multiple predictors. In Section 2.4, we assess the results from the daily AOD gap-filling experiments using spatially clustered cross-validation folds. In Section 2.5, we assess whether imputing AOD through the super learner method improves $PM_{2.5}$ prediction on these days with a random forest model. We conclude in Section 2.6.

2.2 Data

2.2.1 Study area

The study area of interest is the contiguous United States, consisting of 48 states and Washington DC using daily data for the month of July 2011. Descriptions of the data sources follow the work of Hu et al. (2017).

2.2.2 $PM_{2.5}$ measurements

We obtain measurements of $PM_{2.5}$ from the U.S. Environmental Protection Agency (EPA) Air Quality System (AQS) (`https://www.epa.gov/outdoor-air-quality-data`). We used 24-hour averaged concentrations collected from 1248 federal reference method samplers.

2.2.3 MODIS AOD

For the purpose of this study, we utilize Collection 6 level 2 Aqua MODIS retrievals at 550 nm wavelength using the MYD04_L2 product (Levy et al., 2013, 2015). High-confidence AOD retrievals from the combined deep-blue/dark target parameter were used (Belle and Liu, 2016). Following previous work (Hu et al., 2017), these retrievals

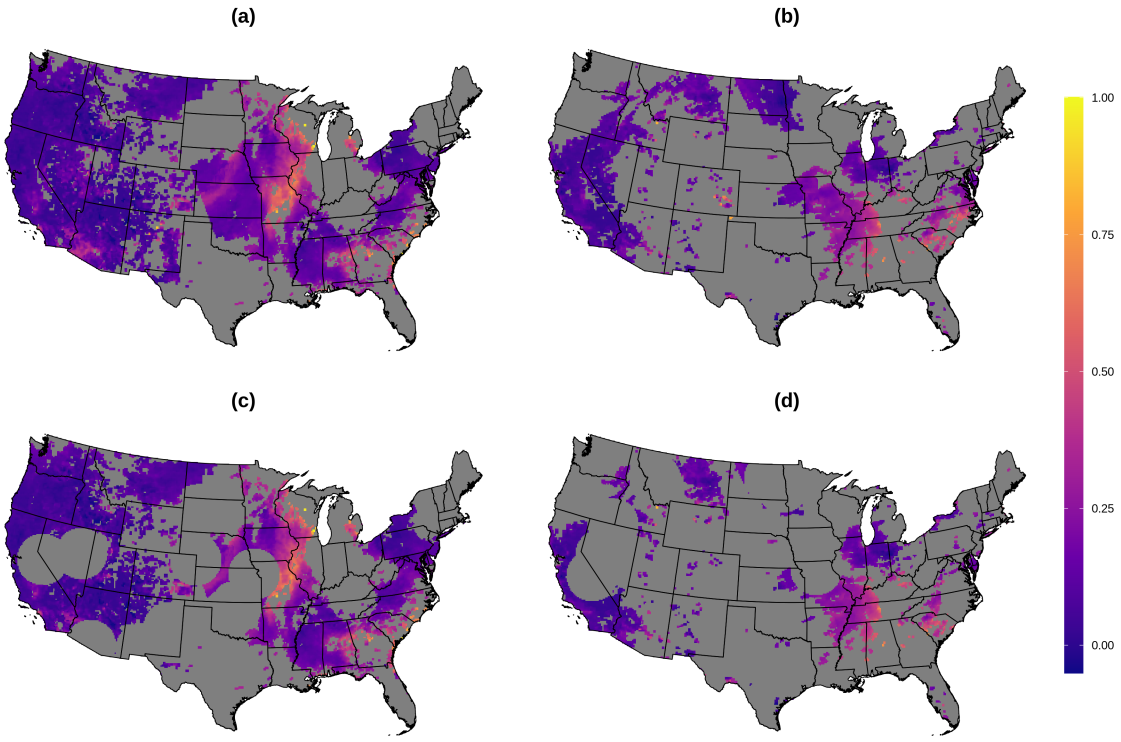


Figure 2.1: MODIS AOD for July 1 and July 12, 2011. Full observed data for **(a)** July 1 and **(b)** July 12, 2011; training data for **(c)** July 1 and **(d)** July 12, 2011. July 1 has the least missingness, and July 12 has the most missingness in July 2011. Grid cells with observed AOD values greater than 1 are excluded from display.

at a resolution of 10km are regridded to 12 km × 12 km Community Multi-Scare Air Quality (CMAQ) grids. We consider daily MODIS AOD data from July 2011, for a total of 53,807 daily cells in the contiguous United States. The proportion of cells in which daily AOD is observed ranges from a minimum of 26.33% to 54.63%, with an average of 41.08%. The top row of Figure 2.1 demonstrates two days with the least and most missing observed AOD points. We used MODIS AOD rather than a finer-scaled product (e.g., 1km^2 products) as our goal was to explore large-scale variation in the national map for AOD.

2.2.4 GEOS-Chem AOD

GEOS-Chem is a "global 3-D model of atmospheric chemistry driven by assimilated meteorological observations from the Goddard Earth Observing System (GEOS) of the NASA Global Modeling Assimilation Office (GMAO)" (`http://acmg.seas.harvard.edu/geos/`) (Bey et al., 2001). We utilize version 10.1 of the model using GEOS-5 meteorological data for 2011, with total column AOD calculated as the sum of 6 AOD parameters (sulfate-nitrate-ammonium, black carbon, organic carbon, accumulation-mode sea-salt, coarse-mode sea-salt, and total dust) over 37 vertical layers (from the surface up to ≈ 20 km) (Hu et al., 2017; Li et al., 2013b).

2.2.5 Meteorological variables

We obtained meteorological data from the North American Land Data Assimilation System phase 2 (NLDAS-2) (`https://ldas.gsfc.nasa.gov/nldas/`) (Cosgrove et al., 2003; Mitchell et al., 2004). These data have a spatial resolution of approximately 13 km and are available hourly. For this analysis, we use pressure at surface (pa), u- and v-direction wind speed (m/sec), temperature (K), relative humidity (%), precipitation (kg/m^2), fraction of total precipitation that is convective (no units), convective available potential energy (J/kg), surface DW shortwave radiation flux (W/m^2), surface DW longwave radiation flux (W/m^2), and potential evaporation (kg/m^2). Measurements are averaged from 10 a.m. to 4 p.m. local time to construct daily daytime observations, roughly coinciding with the Aqua overpass time (about 1:30 pm).

2.2.6 Land use

We include include elevation obtained from the National Elevation Dataset at 30 m spatial resolution (`https://viewer.nationalmap.gov/basic/`). We obtained total

length of highways (m), total length of limited-access road (m), and total length of local road (m) from ESRI StreetMap USA (Environmental Systems Research Institute, Redlands, California, USA). Forest cover (unitless) and impervious surface (%) are derived from the National Land Cover Database (`https://www.mrlc.gov/`). In addition, we include point emissions data for $PM_{2.5}$ and PM_{10} combined (in tons) from the EPA 2011 National Emissions Inventory report (`https://www.epa.gov/air-emissions-inventories`). Population density is obtained from the 2010 Census at the tract level (population/km^2).

2.2.7 Data integration

Data were projected into a common coordinate system using the U.S. Lambert conformal conic projection. For each 12 km × 12 km grid cell, forest cover, impervious surface, and elevation were averaged, while road length and point emission values were summed. Meteorological variables and population density were assigned based on nearest distance. Grid cells containing multiple $PM_{2.5}$ monitors for a day were averaged.

2.3 Statistical methods

We provide a brief description of lattice kriging (with additional details in the Supplementary Materials), random forest, and super learner methods.

2.3.1 Lattice kriging

We follow the model description of lattice kriging (LatticeKrig or LK) laid out by Nychka et al. (2015). LK has been effectively used for spatial prediction in a variety of different applications, such as indoor gamma radiation dose-rates (Chernyavskiy et al., 2016) and satellite-measured land surface temperatures (Heaton et al., 2019).

At a high-level, LK models the spatial process using several levels of two-dimensional basis functions, which are laid out on a grid and approximately double with each successive layer. These basis functions are compact, which means that for a particular point only a small number of basis function are used to make the prediction. The coefficients associated with the basis functions are assumed to be correlated, and this structure can flexibly model observed spatial covariance structures. Estimation proceeds through a likelihood-based approach after specifying various tuning parameters.

Following the notation of Nychka et al. (2015), we observe $\{y_i\}$ at locations $\{\boldsymbol{x}_i\}$ for $i = 1, .., n$. We assume $\{y_i\}$ follow an additive model consisting of a mean function based on covariates, a spatial process, and a measurement error term:

$$y_i = \boldsymbol{Z}_i^T \boldsymbol{d} + g(\boldsymbol{x}_i) + \epsilon_i, \tag{2.1}$$

where $\boldsymbol{d}$ is a $p \times 1$ vector of fixed coefficients associated with the covariates $\boldsymbol{Z}_i$, and $g(\boldsymbol{x_i})$ denotes the spatial process. The mean-zero error terms ϵ_i are presumed to be independent and identically distributed, i.e., $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, where $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T$.

The overall spatial process $g(\boldsymbol{x}_i)$ can be written as a sum of L independent spatial processes $g_l(\boldsymbol{x}_i)$:

$$g(\boldsymbol{x}_i) = \sum_{l=1}^{L} g_l(\boldsymbol{x}_i) = \sum_{l=1}^{L} \sum_{j=1}^{m(l)} c_j^l \phi_{j,l}(\boldsymbol{x}_i), \tag{2.2}$$

where $\phi_{j,l}$ denotes the the lth level of resolution's jth basis function, and c_j^l denotes the coefficient associated with this basis function. Although the basis functions and number of levels are fixed (i.e., chosen), the coefficients for each level l, $\boldsymbol{c}^l = (c_1^l, ..., c_{m(l)}^l)^T$ are assumed to follow a multivariate normal with mean zero and covariance $\rho \boldsymbol{Q}_l^{-1}$:

$$\boldsymbol{c}^l \sim N(\boldsymbol{0}, \rho \boldsymbol{Q}_l^{-1}). \tag{2.3}$$

Each level's spatial process is independent with marginal variance $\rho\alpha_l$ subject to the constraint $\sum_{l=1}^{L} \alpha_l = 1$, so that the marginal variance of the overall spatial process $g(\boldsymbol{x}_i)$ is ρ. We provide a more thorough description of the model in the Supplemental Materials, and we also refer readers to the originating paper (Nychka et al., 2015), the comparison paper by Heaton et al. (2019), and the documentation for the R implementation (Nychka et al., 2016). We note that a disadvantage of LK is that it is not currently implemented to handle spatio-temporal data.

Parameters

Several parameters can impact LK's predictions and inference. We review them here and discuss their impact on the implied spatial covariance, along with the associated parameter name in the `R` package `LatticeKrig` (version 8.4) (Nychka et al., 2016) implementing the method:

- *Number of basis functions and levels*: The number of basis functions follows from: (1) specifying the number of levels of resolutions, denoted by `nlevel` in the package, and (2) specifying the number of basis functions along the longest dimension at the first (coarsest) level of resolution, parameterized by `NC`. Each successive level of resolution has roughly double the basis functions, so this determines the entire grid. Nychka et al. (2015) suggests choosing these so that the coarsest level of resolution can capture the overall correlation range, and so that the finest level of resolution can capture fine scale changes in the spatial process. Holding all else constant (including the levels of resolution), increasing the number of basis functions at the coarsest level decreases the implied covariance for a given distance. A parameter for adding extra basis functions to the edges to reduce artifacts in prediction is determined by `NC.buffer`, which is set to 5 by default.

- *Relative weight of each spatial level's process*: Recall that each level's spatial

process $g_l(\boldsymbol{x}_i)$ has a marginal variance of $\rho\alpha_l$ where $\sum_{l=1}^{L}\alpha_l = 1$. In the package implementation, $\sqrt{\alpha_l}$ multiplies the basis functions (after normalization), such that

$$g(\boldsymbol{x}) = \sum_{l=1}^{L}\sqrt{\alpha_l}g_l(\boldsymbol{x}) = \sum_{l=1}^{L}\sum_{j=1}^{m(l)} c_j^l(\sqrt{\alpha_l}\phi_{j,l}(\boldsymbol{x}_i)).$$

Choosing α parameters (relative weights) can be simplified into a single tuning parameter ν (`nu` in the R package), where $\alpha_l \propto 2^{-2l\nu}$. Small values of ν (e.g., 0.1) weight each level of resolution more equally, while larger values of ν (e.g., 1.25) result in more heavily weighting the coarsest level of resolution.

- *Scale/range parameter*: Briefly, the coefficient vector $\boldsymbol{c}^l$ for level l follows a Gaussian Markov random field, and in particular, a spatial autoregression (SAR). In the `LatticeKrig` package, one specifies $a = 4+\kappa^4$ (or `a.wght`). Holding other parameters constant, large values of a suggest less effective correlation range, i.e., for a given distance the implied correlation of the LK model will be lower as a is increased. A small value of a, e.g. 4.01 (the default setting in `LatticeKrig`) is similar to a thin-plate spline where there is a very large range and strong spatial dependence. Some greater detail in the originating paper and package documentation (Nychka et al., 2015, 2016).

2.3.2 Random forest

Random forest (RF) consists of constructing a large number of regression trees (Breiman, 2001). At a high level, regression trees search for the best (as determined by mean-squared error) binary split among the covariates (also referred to as predictors or features), and then split the data accordingly. This process continues until some condition is met (e.g., there is only 1 observation left, so no further split can be made). Two key components of RF are: (1) bagging, or bootstrap aggregation, wherein each tree is fit to a random sample (with replacement) of the original sample; and (2) at

each node, the algorithm considers only a random subset m of the original p predictors for deciding on the best split. A single decision tree would likely overfit to the data. Averaging many trees that implement these two components results in reducing variance while maintaining low bias from the procedure.

Consider a the vector of $p \times 1$ features $\boldsymbol{z_i} = (z_i^1, ..., z_i^p)^T$, and the response y_i, as in the preceding subsection for LK, with $i = 1, .., n$. In our application, the $\boldsymbol{z}$ vector includes the spatial coordinates $\boldsymbol{x}$ of the observation, as well as land-use and meteorological covariates. We describe the algorithm largely following the descriptions of Zhang et al. (2019) and Chapter 15 of Hastie et al. (2009):

1. Draw a random sample of size n with replacement, denoted as $D_b = (\boldsymbol{z}_i^*, y_i^*)_{i=1}^n$.

2. Fit a single tree to the bootstrapped data D_b through the following steps:

   (a) Start at the root node, $\mathcal{N}$, which consists of all bootstrapped observations.

   (b) Draw a random sample of m features (without replacement) from the original set of p features.

   (c) Find the best split among the m selected features that minimizes the mean-squared error. The mean-squared error is based on using the average of the response values for the two subnodes. More formally, let $j = 1, ..., m$ index the selected features, $z^{(j)}$. For each of the m features, consider all unique values c which are candidates for splitting observations into two subnodes, $\mathcal{N}_1$ and $\mathcal{N}_2$. For each particular quantitative variable $z^{(j)}$ and possible split value c, observations are assigned to $\mathcal{N}_1$ if $z^{(j)} \leq c$, and otherwise they are assigned to $\mathcal{N}_2$. We then select the best split c among the m variables that minimizes the squared error,

$$\sum_{k=1}^{2} \sum_{i \in \mathcal{N}_k} (\bar{y}_k^* - y_i^*)^2,$$

where $\bar{y}_k^\star$ is the average of responses in subnode $\mathcal{N}_k$ (Zhang et al., 2019).

(d) Continue each new daughter node until they have no more than n_{size} observations, no variation in the response, or no variation in the predictors.

(e) Denote the resulting tree as T_b.

3. Repeat steps (1) and (2) for $b = 1, ..., B$ total trees.

4. Produce predictions based on the average of the B trees, $\hat{y} = B^{-1} \sum_{b=1}^{B} T_b(\boldsymbol{z})$.

Because a random sample with replacement is taken for each tree in RF, each tree can generate out-of-sample predictions for those observations not selected. The overall out-of-sample, or out-of-bag (OOB), prediction error can be used to approximate the test data error and tune the parameters in many settings. However, in spatial data settings, the training and test data may reflect very different spatial domains. Using the OOB prediction to tune parameters may be deceiving in this case; we discuss this in greater detail in the Supplemental Materials.

There are several options for RF in R such as `randomForest` (Liaw and Wiener, 2002). We utilize the `ranger` package (version 0.12.1) in R (Wright and Ziegler, 2017) which is built to better handle large data. A benefit of RF is that it may easily be parallelized, as the fitting of each tree is independent. The `ranger` implementation automatically detects the number of CPU cores in an environment and parallelizes accordingly. Random forest, unlike LK as currently implemented, could be fit to spatio-temporal data using appropriate features to denote temporal aspects of the data (e.g., see Just et al. (2018)). However, we take the approach of fitting separate models for each day for the task of AOD gap-filling, treating each day as a separate experiment.

Parameters

The key parameters are the number of trees (B), the number of predictors to randomly select for each split (m) from the original p, and the node size (n_{size}). In general, we choose B to be large for predictions, while we consider different values for n_{size} and m. In our experience, n_{size} is of secondary importance compared to m. In this implementation, B is denoted by `num.trees` (500 by default), n_{size} is denoted by `min.node.size` (5 by default in regression), and m by `mtry` (the integer floor of $\sqrt{p}$ by default). These are the same defaults as in `randomForest`, with the exception of `mtry`, which is $p/3$ for regression problems. Several research articles discuss these parameters in a variety of contexts; see Segal (2004), Biau and Scornet (2016), Probst et al. (2019) and the references therein.

2.3.3 Super learner methods

Super learners (SL), related to stacked generalization and stacked regression methods (Breiman, 1996), use a potentially large and diverse set of algorithms by weighting their predictions optimally according to some risk measure such as squared error loss. Although a large number of algorithms are recommended in practice, we use just RF and LK as our algorithms in order to demonstrate the use of SL and to maintain focus on the cross-validation approach. The process for super learners is as follows (Van der Laan et al., 2007; Polley and Van der Laan, 2010; Naimi and Balzer, 2018):

1. Divide observed data into k folds.

2. For each fold k, let the kth fold be the validation data, and the remainder be the training data. Fit each algorithm or model m to the training data and make predictions on the kth fold.

3. Stack all predictions $\hat{\boldsymbol{y}}_m$ for each algorithm.

4. Estimate the weights α_m for algorithm $m = 1, ..., M$ using the model formulation

$$y_i = \sum_{m=1}^{M} \alpha_m \hat{y}_{i,m} + \epsilon_i, \tag{2.4}$$

where $\alpha_m \geq 0$ and $\sum_{m=1}^{M} \alpha_m = 1$. α_m can be estimated by non-negative least-squares methods and then normalizing the weights to sum to 1.

After these α_m model weights are estimated through the cross-validation process, each algorithm is fit to the full observed data, and test data predictions are made by using these weights for combining predictions. Davies and Van Der Laan (2016) provides a discussion of extending super learner theory to the case of spatial data. Murray et al. (2019) uses a similar stacked regression approach for determining weights in combining separate models for $PM_{2.5}$ prediction.

For each day, we construct 10 cross-validation folds using the `blockCV` R package (version 2.1.1) (Valavi et al., 2019). This constructs spatial blocks for the validation dataset, so that performance more accurately mimics the task of gap-filling AOD. In the Supplemental Materials, we provide a full set of the maps showing these spatial block cross-validation folds. Sarafian et al. (2019), Murray et al. (2019) (for $PM_{2.5}$ prediction) and Young et al. (2016) (for NO2) also consider spatially clustered cross-validation approaches for assessing model performance. Based on the cross-validation folds, we stack LK and RF validation predictions. We assess 4 different methods for combining LK and RF, where we restrict the weights in (2.4) to be between 0.1 and 0.9:

1. **Average**. We construct a simple average of RF and LK predictions. Cross-validation data is not used in this approach.

2. **SL: overall**. After stacking all of the cross-validation predictions for all days together, we produce a single set of optimal weights with (2.4) for making predictions.

3. **SL: daily**. We stack cross-validation predictions for each day separately, and we obtaining a daily set of optimal weights with (2.4).

4. **SL: distance-based**. For each cross-validation fold on each day, we determine the closest distance between each point in the cross-validation fold and the training data. We then stack all of the cross-validation predictions across days together with these nearest-neighbor distances. We then bin these stacked predictions according to distances with bin widths of 25km from 0 to 300km and higher. Using (2.4), we estimate the optimal weights for LK and RF for each distance grouping. We then fit a simple loess model relating interval midpoint distance and optimal weight, and we use these fitted optimal weights for combining LK and RF for predictions. The motivation for this last technique is that the farther the unobserved point is from the observed data, the more different algorithms may be in predictions. If there is strong spatial correlation, then LK may perform better; in contrast, if there is limited range in the spatial correlation and the covariates are more important, then RF may produce better fits based on the relationship between the covariates and response.

2.4 Application to AOD imputation

2.4.1 Experimental setting

From the observed AOD data, we consider a spatially clustered approach for creating a testing dataset on which to evaluate the results. In the proposed method, ten random AOD observations are selected from the observed AOD values for each day. These observations and any other observations within a 250km radius are then held out as the test dataset. Figure 2.1 demonstrates the observed and training data on two particular days. This approach to creating testing data is an attempt to mimic

the actual observed pattern of AOD data, where large contiguous areas are missing and require imputation. In particular, for many missing observations, there are likely no points nearby to aid in prediction. In our analysis, we consider each day separately for model fitting and prediction. We primarily assess model performance on the basis of root mean-squared error (RMSE) and the coefficient of determination (R^2), as well as the intercept and slope from a linear model relating the true left-out AOD observations to the predicted values. Discussion of tuning random forest and lattice kriging is provided in the Appendix.

2.4.2 Results

We highlight several results from our analysis. First, daily results show that any of the average/super learner approaches match or exceed performance from either LK or RF alone on a majority of days (Table 2.1, Figure B.2). While there are some days where either LK or RF does particularly well, there are also days where they perform worse than any of the other methods. The ensemble methods are the best or close to the best in terms of RMSE and R^2 on a majority of days. The distance-based SL method performs best on more days (10 out of 31 days) than any of the other approaches considered. Based on our described approach, the distance-based SL prediction weights LK greater for testing points that are close to training data, while for points farther away from the training data, it weights RF more, relying on a combination of location, land use, and meteorological features (see Figure B.7 in the Supplemental Materials).

Evaluating test predictions across all 31 days together, LK and RF have R^2 values of 0.644 and 0.619, respectively. The average and SL methods improve to R^2 values in the range of 0.657 to 0.659. Compared to LK alone, the RMSE is reduced 2.34% and 2.30% in the distance-based and overall SL models, respectively. A simple average of the LK and RF predictions also provides most of these gains. The super learner

method based on the daily construction of weights performs well but is marginally worse than the other ensemble methods.

Both LK and RF methods diverge substantially from using just the observed AOD data in the July 2011 averages (Figure 2.2). LK and RF predictions are notably different from each other in areas of high elevation in the Appalachian Mountains and in parts of Colorado, where LK predicts higher values relative to RF (Figure 2.3). There are also apparent edge effects in some areas like Florida and Texas, where LK predictions will tend to diverge more substantially from those of RF. These may be partly due to issues with LK's coefficient estimation in areas where there is little data for a particular day (see Figures B.3-B.4 in the Supplemental Materials for plots of daily AOD predictions and daily differences between LK and RF).

Method	R^2	RMSE (x100)	Intercept	Slope	# of days ranked Best	# of days ranked Worst
LatticeKrig	0.644	6.66	-0.01	0.94	7	12
Random Forest	0.619	6.90	-0.01	0.92	3	19
LK-RF Average	0.658	6.52	-0.01	0.97	4	0
SL: Overall	0.659	6.51	-0.01	0.97	4	0
SL: Daily	0.657	6.52	-0.01	0.96	3	0
SL: Distance-based	0.659	6.50	-0.01	0.96	10	0

Table 2.1: Summary statistics on combined test AOD predictions across all days of July 2011.

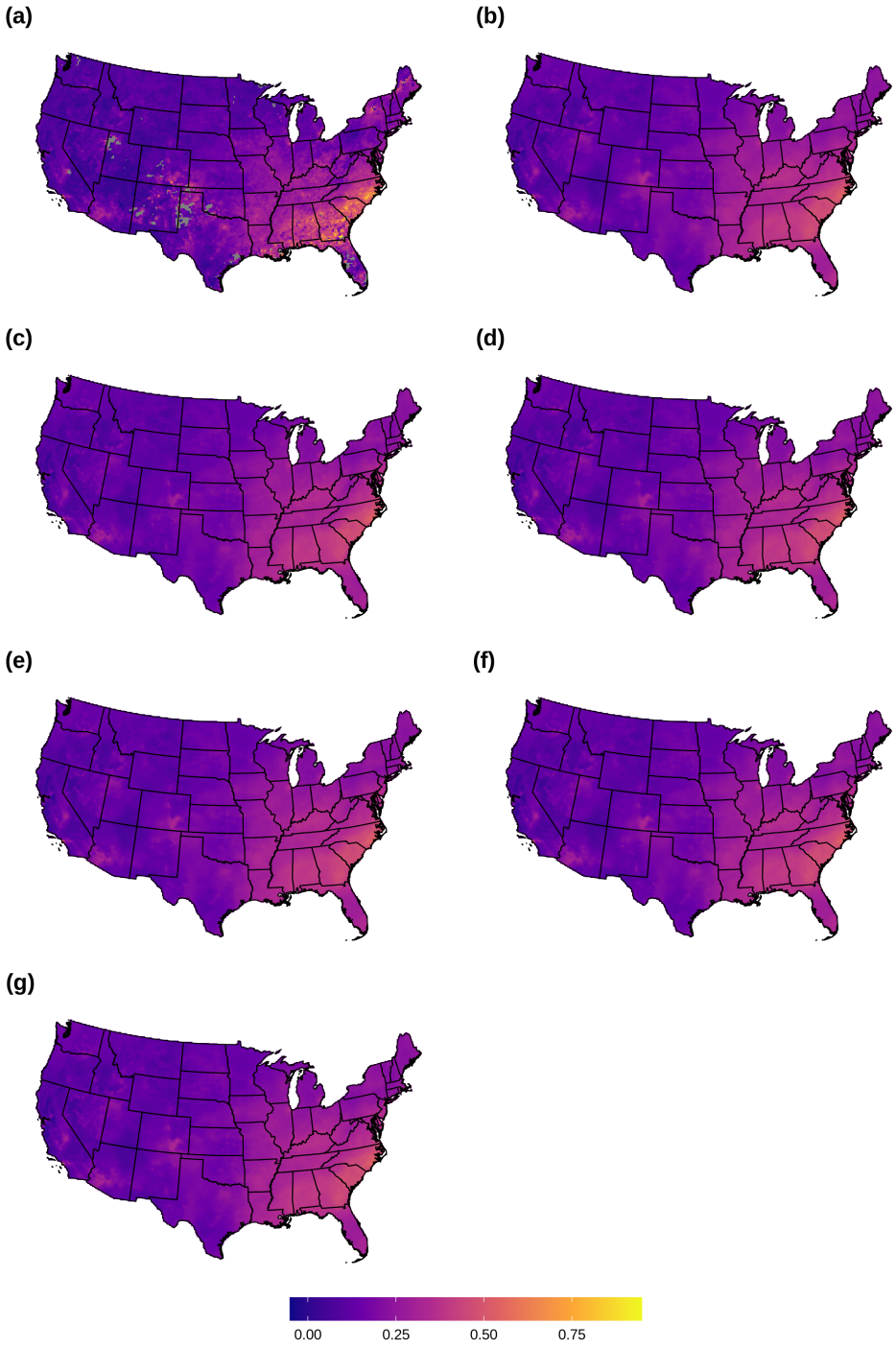


Figure 2.2: July 2011 average of observed and predicted daily AOD: **(a)** Observed AOD; **(b)** LK; **(c)** RF; **(d)** Average of LK and RF; **(e)** SL: Overall; **(f)** SL: Daily; **(g)** SL: Distance-based. Grid cells with observed AOD values greater than 1 are excluded from display.

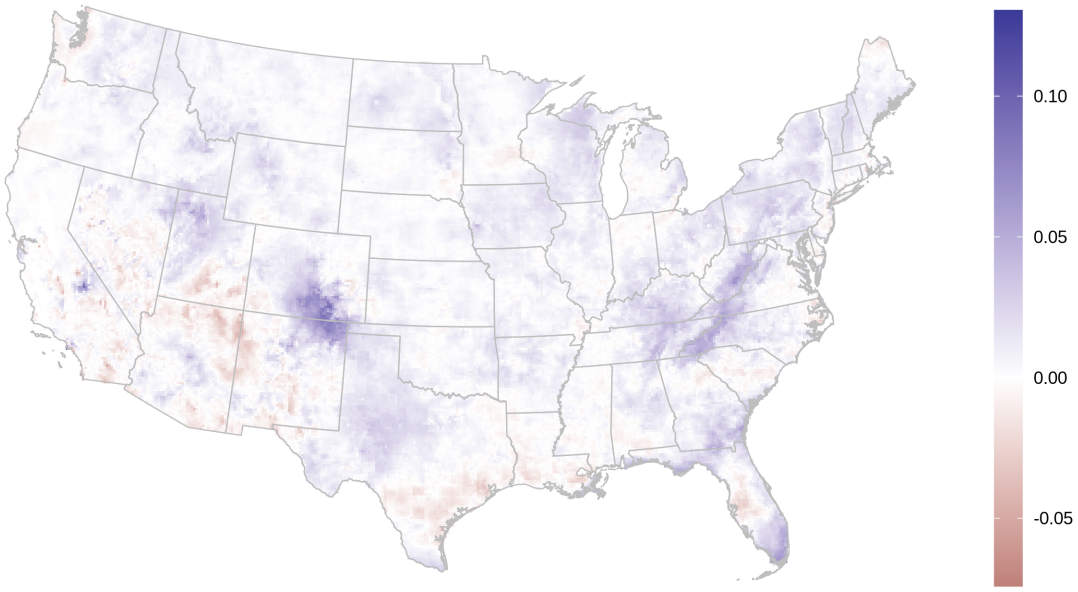


Figure 2.3: Difference between lattice kriging and random forest average AOD predictions.

2.5 Impact of imputed AOD on $PM_{2.5}$ prediction

2.5.1 Experimental setting

Using the super learner distance-based method from the AOD gap-filling analysis, we fit several random forest models for PM2.5 concentration estimation in order to assess whether the inclusion of imputed AOD can improve performance. There are five variations on the features available for random forest:

- **M1**: Include neither AOD nor GEOS-Chem.
- **M2**: Include GEOS-Chem.
- **M3**: Includes gap-filled AOD. This variable is defined to be observed AOD where available, and otherwise the predicted AOD value based on the super-learner distance-based method. GEOS-Chem is also included as a separate feature.
- **M4**: Include AOD by replacing missing values of AOD with GEOS-Chem (as in Hu et al. (2017)).
- **M5**: Includes observed AOD, and training solely on observations where AOD is observed. For predictions, missing values of AOD are replaced with the gap-filled AOD. GEOS-Chem is also included as a separate feature.

We consider three distinct 10-fold cross-validation approaches for assessing performance:

- **Random**: Cross-validation folds are constructed by selecting observed $PM_{2.5}$ monitors at random on a daily basis.
- **Constant spatial clusters**: Cross-validation folds are constructed by creating spatially clustered areas that are *constant* across all days. A particular area of the map will be assigned to the same cross-validation fold for every day.

- **Varying spatial clusters**: Cross-validation folds are constructed by creating spatial clusters at random for each day.

Spatial clusters are constructed using the `blockCV` package (version 2.1.1) in R with block widths of 150km. Figure B.8 in the Supplemental Materials displays the constant spatial construction by color-coding monitor locations.

The other features included in the random forest models are the same as those in the AOD analysis. All models except M5 additionally include an indicator variable for whether AOD was observed at the location, and all models include a so-called *convolution* layer of $PM_{2.5}$. Several analyses (Hu et al., 2017; Di et al., 2019) have demonstrated that a weighted-average of nearby $PM_{2.5}$ observations can aid in model prediction for $PM_{2.5}$. Briefly, for each location, the convolution layer of $PM_{2.5}$ is a weighted average of all other *training* $PM_{2.5}$ observations, not including the location itself. The weights are inversely proportional to the squared distance between locations (less distant observations in the training data are weighted more). The procedure for creating the convolution $PM_{2.5}$ layer must be repeated for each training/validation split for each day.

Models were fit both for each day separately as well as for all of the days in July 2011 together. In the latter spatio-temporal random forest model, day of the year and day of the week are additionally included as integer predictor variables. Our primary metrics of interest are RMSE, R^2, and the full prediction maps, but we also present the intercept/slope estimates from fitting a linear regression model with the true $PM_{2.5}$ values as the dependent variable and the random forest prediction as the independent variable. For all models, we set the number of trees to 2000. We varied m (`mtry`) between values of 4, 8, 12, and 16 and presented the best results for each model and cross-validation fold type. The full maps and feature importance results are based on $m = 4$. The Supplemental Materials include additional figures and tables for $m = 8$.

2.5.2 Results

We highlight a few notable results. First, the daily random forest models suggest that RMSE is improved consistently but only marginally by including the imputed AOD predictor vs. the four alternatives (Table 2.2, M3a). The outlier model is M5, which trains solely on observations where AOD is observed and predicts using the imputed AOD where AOD is missing. The results from model M5 are substantially worse than the other models, with a relatively biased prediction map (Figure 2.5(d)). For the remainder of the results, we omit discussion of this model. Generally the gain in RMSE for M3 ranges from 0.01 to 0.03 $\mu g\, m^{-3}$ against the other models. The results for R^2 are similar, with small gains of approximately 0.003. The gain in performance is less in the random cross-validation case than in the two spatially clustered cross-validation analyses. Second, the daily random forest models tended to have better $PM_{2.5}$ prediction in locations where AOD was not observed, regardless of the features included. Third, cross-validated RMSE is substantially larger in spatial cross-validation settings than in the case with folds consisting of randomly selected locations.

The spatio-temporal random forest results in the second set of columns in Table 2.2 show somewhat different patterns. RMSE and R^2 are generally improved over the daily models for the random and varying spatially clustered cross-validation analyses, but there is no longer a benefit to including imputed AOD. On the contrary, the model predictions tend to do better when neither AOD nor GEOS-Chem are included on the basis of R^2 and RMSE. The exception to these results are in the constant spatially clustered cross-validation setting – here there is some very marginal improvement from including imputed AOD over the other models. We posit that in spatio-temporal models, multiple days' observations in the same area as where we intend to make a prediction on a different day can largely diminish the predictive power of AOD. However, when the same spatial area is consistently missing, the model can no longer

rely on other days' observations for the same area to improve prediction accuracy. Notably, this setting mirrors qualities of producing full maps of $PM_{2.5}$ observations. Given that there is a fixed network of monitors (not all of which operate on every day), $PM_{2.5}$ prediction is primarily focused on areas where there is never a monitor present. We emphasize that the improvement in RMSE from including the imputed AOD in this constant spatially clustered cross-validation setting is small at 1.1%, 0.49%, and 0.57% compared to the models with no AOD or GEOS-Chem, just GEOS-Chem, or the combined AOD/GEOS-Chem variable, respectively. Notably in this case, daily models slightly outperform the performance of the spatio-temporal models. In the other cross-validation settings, RMSE and R^2 both improve substantially from fitting a full spatio-temporal model over a series of daily models.

Results for RMSE and R^2 by region (as defined by NOAA) and cross-validation setting are also provided in Tables 2.3 and 2.4. RMSE results tend to be worst in the West, Southwest, and Central regions across cross-validation settings. Notably, although RMSE is quite low for the Northwest region, the R^2 for this area is comparatively low. The spatio-temporal models improve the RMSE and R^2 except for the constant spatially clustered cross-validation, where there is no improvement and perhaps a slight decrease in performance. Variable importance metrics for the $m = 4$ setting based on the spatio-temporal models are presented in Table B.1 using the permutation-based method (Breiman, 2001). Briefly, this importance metric denotes the increase in mean-squared error on the OOB sample for each tree after permuting the values of the feature. On this basis, the convolution layer of $PM_{2.5}$ is the most important predictor for these models. When imputed AOD is included, the relative importance of several other variables is slightly diminished. While imputed AOD is not the most important feature, it appears substantively important on the basis of mean decrease in accuracy. Additional feature importance tables are provided in the Supplemental Materials for $m = 8$ (Table B.2). In general, for larger m, the convolu-

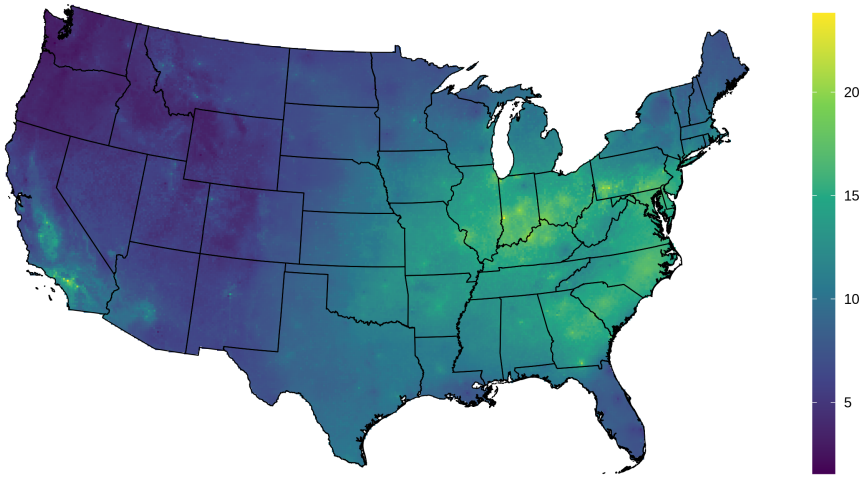


Figure 2.4: Average July 2011 $PM_{2.5}$ predicted map ($\mu g\,m^{-3}$) using imputed AOD spatio-temporal random forest model (M3).

tion layer of $PM_{2.5}$ becomes more important – it is more likely to be selected as the opitimal feature for splitting a node as m increases, and it is a particularly strong predictor.

Figure 2.4 shows the July 2011 averaged values from M3 with the gap-filled AOD as a feature in the spatio-temporal random forest model. When comparing model M3 to models M1, M2, and M4, the average of monthly mean differences are close to 0 $\mu g\,m^{-3}$, but the monthly mean differences are apparently spatially correlated (Figure 2.5). The model trained only on points where AOD is observed (M5) leads to over-estimated average monthly values of $PM_{2.5}$ relative to the model using gap-filled AOD, with an average difference of 0.25 $\mu g\,m^{-3}$. The standard deviation of daily differences for all grid cells for July 2011 is 0.44 $\mu g\,m^{-3}$ for M3 and M1, 0.32 $\mu g\,m^{-3}$ for M3 and M2, and 0.38 $\mu g\,m^{-3}$ for M3 and M4, suggesting small but meaningful variability in the daily model predictions (Figure B.11).

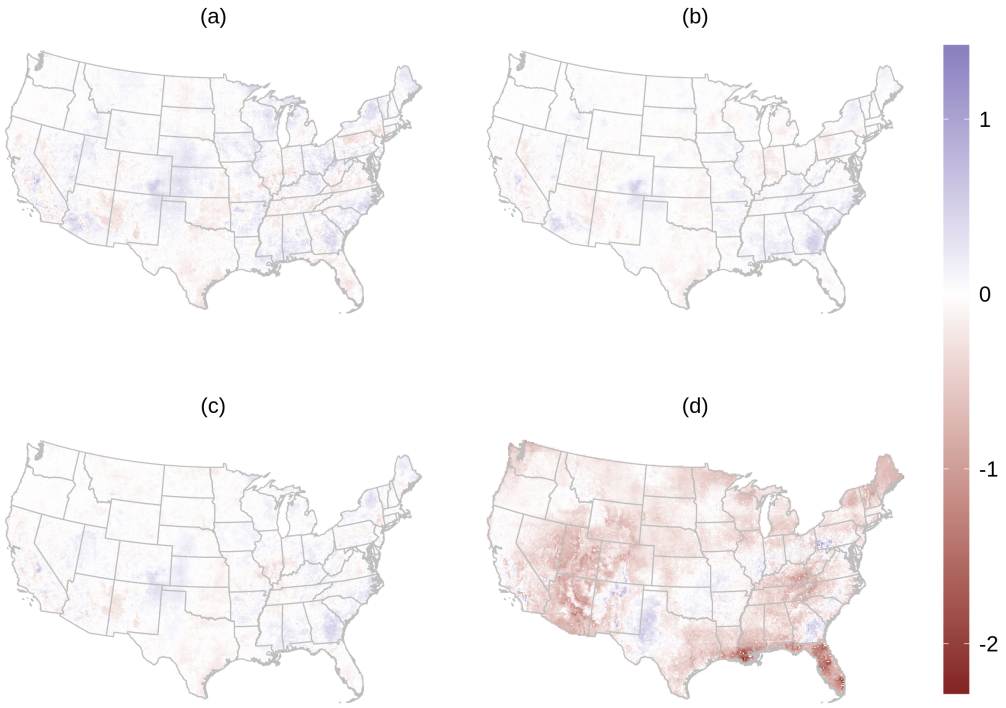


Figure 2.5: Difference between the imputed AOD RF model (M3) and other RF models in average July 2011 $PM_{2.5}$ predictions ($\mu g\, m^{-3}$): **(a)** M1: model with no AOD or GEOS-Chem; **(b)** M2: GEOS-Chem; **(c)** M4: Replacing missing values of AOD with GEOS-Chem; **(d)** M5: Train on observed AOD, and predict by replacing missing AOD values with imputed values.

		Daily					Spatio-temporal				
Setting	AOD Status	M1a	M2a	M3a	M4a	M5a	M1b	M2b	M3b	M4b	M5b
RMSE ($\mu g\, m^{-3}$)											
Random	All	3.30	3.29	3.28	3.30	3.68	2.96	2.98	2.99	2.98	3.20
Random	Missing	3.30	3.28	3.27	3.29	3.84	2.99	3.00	3.02	3.01	3.29
Random	Observed	3.31	3.30	3.29	3.31	3.41	2.92	2.94	2.95	2.94	3.04
Constant cluster	All	3.66	3.64	3.62	3.64	3.99	3.68	3.66	3.63	3.66	3.75
Constant cluster	Missing	3.61	3.58	3.56	3.59	4.09	3.63	3.60	3.57	3.61	3.74
Constant cluster	Observed	3.74	3.73	3.72	3.73	3.82	3.76	3.75	3.72	3.73	3.76
Varying cluster	All	3.66	3.65	3.63	3.65	4.00	3.33	3.34	3.35	3.35	3.56
Varying cluster	Missing	3.61	3.58	3.57	3.60	4.11	3.34	3.33	3.34	3.34	3.61
Varying cluster	Observed	3.75	3.74	3.72	3.74	3.83	3.33	3.35	3.35	3.35	3.47
R^2 (x100)											
Random	All	75.3	75.5	75.7	75.4	69.7	80.4	80.2	79.8	80.0	77.1
Random	Missing	75.9	76.2	76.3	76.1	68.1	80.5	80.4	79.9	80.1	76.5
Random	Observed	73.8	73.9	74.1	73.8	72.3	79.9	79.7	79.3	79.5	78.0
Constant cluster	All	69.9	70.1	70.4	70.1	64.1	69.6	69.9	70.3	69.9	68.6
Constant cluster	Missing	71.5	71.8	72.2	71.7	63.6	71.2	71.6	72.1	71.5	69.8
Constant cluster	Observed	66.8	66.9	67.2	66.9	65.2	66.4	66.7	67.1	66.8	66.6
Varying cluster	All	69.7	70.0	70.3	70.0	63.9	75.6	75.6	75.4	75.5	72.1
Varying cluster	Missing	71.4	71.8	72.1	71.7	63.2	76.3	76.4	76.2	76.3	72.2
Varying cluster	Observed	66.5	66.6	67.1	66.7	65.1	74.1	73.9	73.9	73.7	71.8

Table 2.2: R^2 and RMSE results from daily and spatio-temporal random forest model for different 10-fold cross-validation settings.

		Daily					Spatio-temporal				
Setting	AOD Status	M1a	M2a	M3a	M4a	M5a	M1b	M2b	M3b	M4b	M5b
Random	Central	3.70	3.67	3.65	3.68	4.15	3.56	3.56	3.58	3.56	3.78
	East North Central	3.17	3.17	3.16	3.18	3.52	3.04	3.03	3.01	3.02	3.15
	Northeast	3.31	3.27	3.26	3.29	3.74	2.97	2.96	3.00	2.97	3.23
	Northwest	1.48	1.49	1.49	1.49	1.82	1.23	1.24	1.24	1.24	1.35
	South	2.35	2.35	2.33	2.35	2.77	2.11	2.13	2.15	2.15	2.39
	Southeast	3.55	3.54	3.53	3.56	4.22	3.39	3.41	3.40	3.43	3.72
	Southwest	4.13	4.11	4.09	4.10	4.43	3.58	3.61	3.62	3.54	3.94
	West	4.41	4.42	4.41	4.40	4.47	3.58	3.64	3.70	3.67	3.76
	West North Central	2.91	2.92	2.92	2.92	3.08	2.34	2.37	2.35	2.35	2.50
Constant cluster	Central	4.38	4.34	4.32	4.34	4.67	4.54	4.49	4.46	4.50	4.60
	East North Central	3.36	3.36	3.34	3.36	3.67	3.46	3.41	3.38	3.43	3.48
	Northeast	3.55	3.51	3.49	3.53	3.90	3.54	3.49	3.46	3.50	3.60
	Northwest	1.48	1.48	1.50	1.49	1.84	1.46	1.46	1.46	1.46	1.53
	South	2.47	2.47	2.46	2.47	2.90	2.49	2.47	2.47	2.48	2.64
	Southeast	3.91	3.88	3.83	3.90	4.54	3.94	3.92	3.85	3.93	4.08
	Southwest	4.29	4.27	4.26	4.27	4.60	4.34	4.34	4.35	4.32	4.47
	West	5.20	5.21	5.21	5.19	5.28	5.10	5.11	5.08	5.08	5.08
	West North Central	3.04	3.04	3.04	3.04	3.18	3.09	3.08	3.08	3.08	3.09
Varying cluster	Central	4.39	4.34	4.32	4.34	4.70	4.29	4.26	4.26	4.27	4.50
	East North Central	3.34	3.35	3.34	3.34	3.74	3.30	3.29	3.28	3.29	3.37
	Northeast	3.53	3.49	3.49	3.51	3.94	3.26	3.23	3.27	3.25	3.49
	Northwest	1.51	1.51	1.52	1.53	1.85	1.29	1.30	1.31	1.31	1.39
	South	2.48	2.49	2.48	2.49	2.93	2.31	2.33	2.34	2.33	2.59
	Southeast	3.92	3.89	3.84	3.92	4.54	3.76	3.74	3.70	3.75	4.03
	Southwest	4.45	4.44	4.42	4.42	4.76	3.97	4.02	4.07	4.02	4.30
	West	5.16	5.17	5.14	5.16	5.18	4.18	4.23	4.26	4.25	4.42
	West North Central	2.97	2.97	2.97	2.97	3.11	2.40	2.43	2.45	2.45	2.57

Table 2.3: Regional RMSE results ($\mu g\,m^{-3}$) for daily and spatio-temporal random forest model for different 10-fold cross-validation settings.

		Daily					Spatio-temporal				
Setting	AOD Status	M1a	M2a	M3a	M4a	M5a	M1b	M2b	M3b	M4b	M5b
Random	Central	61.1	61.7	62.1	61.5	51.1	64.7	64.7	63.8	64.1	59.3
	East North Central	69.4	69.6	69.7	69.3	63.3	72.8	73.2	72.6	72.5	70.7
	Northeast	77.0	77.6	77.8	77.4	70.8	81.9	82.0	81.3	81.6	78.4
	Northwest	44.6	44.2	44.0	44.5	28.2	61.2	60.9	60.0	60.1	53.6
	South	68.0	68.1	68.4	68.0	57.9	74.5	74.0	73.2	73.1	67.5
	Southeast	70.4	70.6	70.8	70.2	59.5	73.2	73.1	72.8	72.3	68.3
	Southwest	46.6	47.3	47.7	47.4	42.4	61.1	60.6	59.3	61.0	51.9
	West	50.7	50.5	50.8	50.9	49.2	68.7	67.7	65.7	66.4	64.8
	West North Central	39.2	39.1	38.9	38.8	33.3	61.5	60.4	60.6	60.7	55.3
Constant cluster	Central	48.0	49.0	49.4	49.2	40.1	45.0	46.0	46.7	45.9	43.5
	East North Central	66.0	66.1	66.6	66.2	60.8	64.4	65.6	66.3	65.1	65.1
	Northeast	73.8	74.4	74.7	74.1	68.1	74.0	74.7	75.1	74.5	73.3
	Northwest	45.4	45.0	44.3	45.0	29.1	45.6	45.6	45.3	45.5	43.7
	South	64.5	64.6	64.9	64.5	53.5	64.3	64.9	65.0	64.6	61.5
	Southeast	64.4	64.9	65.9	64.5	53.5	64.1	64.4	65.6	64.2	62.5
	Southwest	41.9	42.5	42.8	42.6	35.1	40.5	40.4	40.3	40.9	37.5
	West	34.6	34.6	34.5	35.1	32.0	36.6	36.6	37.3	37.2	37.5
	West North Central	34.0	34.1	34.1	34.0	28.8	32.1	32.4	32.6	32.4	31.6
Varying cluster	Central	47.3	48.5	48.9	48.6	39.0	50.8	51.4	51.5	51.5	45.4
	East North Central	66.5	66.5	66.7	66.5	59.1	68.5	69.1	69.3	69.0	67.4
	Northeast	74.3	74.8	74.9	74.6	67.6	79.0	79.3	78.7	79.1	75.3
	Northwest	42.5	42.4	42.0	41.9	26.7	58.1	57.7	57.0	56.8	52.1
	South	64.3	64.2	64.4	64.1	52.7	70.2	69.9	69.6	69.8	63.1
	Southeast	64.3	64.8	65.8	64.2	53.2	68.0	68.3	69.0	68.1	63.5
	Southwest	37.5	37.8	38.5	38.5	30.6	53.0	52.0	50.9	52.8	43.7
	West	33.8	33.6	34.4	34.1	32.8	58.8	57.7	57.3	57.5	52.4
	West North Central	36.6	36.7	36.8	36.6	31.4	59.8	58.8	58.1	58.4	53.4

Table 2.4: Regional R^2 (x100) results for daily and spatio-temporal random forest model for different 10-fold cross-validation settings.

2.6 Discussion

We highlight the main findings of this study in three points. First, we emphasize the importance of constructing testing and cross-validation data that mimic the missing data patterns for both AOD and $PM_{2.5}$ prediction. Previously reported metrics for AOD gap-filling using RF may be over-stated if using out-of-bag (OOB) metrics (Bi et al., 2019; Stafoggia et al., 2019), as using large contiguous areas for testing suggests substantially lower R^2. Different cross-validation settings for $PM_{2.5}$ model evaluation also suggest that performance varies considerably based on the manner of holdout, echoing the findings of previous studies that "spatial" cross-validation performance metrics are typically worse than random cross-validation metrics. In our study, our spatial cross-validation procedures leave out spatially clustered sets of monitors as in several recent studies (Murray et al., 2019; Young et al., 2016; Sarafian et al., 2019). Our results show roughly similar performance metrics for $PM_{2.5}$ estimation compared to previous RF results when using the random cross-validation setting, with ≈ 0.80 (≈ 2.99) vs. 0.81 (2.78) R^2 (RMSE) for summer 2011 in Hu et al. (2017). We fit data for July 2011 only, and without additional variables such as convolutional layers for land use terms.

Second, we demonstrate how super learner approaches combining a large-scale spatial statistical method and machine learning predictions can improve upon the performance of each constituent predictor, and how the super learner method can be further modified for the particular task of AOD gap-filling. Future work should examine extensions to more machine learning and spatial statistical methods. For example, several recent studies have highlighted a number of spatial statistical methods with promising predictive performance and low computational costs (Cressie et al., 2010; Bradley et al., 2016; Heaton et al., 2019), and using these in an ensemble approach may provide further improvements. Spatial data present additional theoretical challenges for super learner methods given that the training data and testing data

will generally not be independent of each other (Davies and Van Der Laan, 2016). A limitation of the current study is the limited time frame and the use of daily rather than spatio-temporal AOD gap-filling models. We focused on July 2011 as there was on average less missingness in AOD in the summer than in other months, and we limited our analysis to a single month due to the high computational cost of fitting daily models with 10-fold cross-validation for the super learner. Future studies may consider expanding the timeframe of spatial prediction beyond a month and including spatio-temporal models that may better utilize the available observed data. Throughout our analysis, we assume that the missing data mechanism for AOD is missing at random (MAR); that is, we assume that AOD's missingness mechanism depends only on the observed values of AOD and other covariates (Little and Rubin, 2019). However, there is some evidence to suggest AOD's missingness is informed by its values, which should be further studied in future work (Grantham et al., 2018).

Finally, we demonstrate that imputed AOD using our proposed ensemble method can have a very small impact on particular RF models for estimating $PM_{2.5}$ concentrations, depending on the cross-validation setting. With a convolution-layer of $PM_{2.5}$ and a rich set of other features, we generally find that AOD (imputed or not), is not strictly needed for good prediction of $PM_{2.5}$ in RF models as judged by R^2 and RMSE. However, population-level metrics like R^2 and RMSE may be misleading in masking improved small-scale predictions, and we find subtle differences in the monthly predicted values between models with and without gap-filled AOD as a predictor. Similarly, Huang et al. (2019) find meaningful differences in PM2.5 predictions in models with and without AOD, particularly in areas with sparse monitors and on high-pollution days. A limitation of the current study is the lack of certain variables for AOD gap-filling and $PM_{2.5}$ estimation; previous work has found that the inclusion of cloud and snow fractions may improve AOD gap-filling and produce meaningful visual improvements in $PM_{2.5}$ estimation (Bi et al., 2019). Moreover, finer resolution

AOD products such as multi-angle implementation of atmospheric correction (MAIAC) derived AOD may provide greater prediction power for $PM_{2.5}$ (Goldberg et al., 2019) at the expense of increasing the computational costs of gap-filling.

Chapter 3

A framework for assessing COVID-19 testing site spatial access

3.1 Introduction

Readily accessible testing is critical to understanding and stopping the spread of COVID-19, and has remained an important issue throughout the pandemic in the United States. The *spatial* accessibility of testing sites reflects the availability and the travel distance to local testing sites (Guagliardo, 2004). Areas with fewer testing locations will likely deter more local residents from getting tested for COVID-19, curtailing disease surveillance and the ability to identify emerging hotspots. Reporting from earlier in the summer of 2020 suggests considerable geographic variability in testing access for large cities in the United States (Kim et al., 2020; Bronner, 2020), where predominantly black and Hispanic areas were likely to be near testing sites with greater demand than predominantly white areas in a number of large U.S. cities.

In this study, we hope to add to the current understanding of testing site access

in Atlanta, Georgia. We build on an approach from the environmental justice literature that previously examined population exposure to toxic release sites (Waller et al., 1997, 1999). In this approach, the distribution of exposure is assessed within each demographic subgroup and the distributions are then compared. Waller et al. (1997) introduced the use of empirical CDFs (ECDFs) as a way to compare these distributions among racial subpopulations for exposure to toxic release sites. Here, we instead consider measures of access to testing sites rather than exposure to toxic release sites. In this framework, the population in a given small area is assigned a common access value. As a straight-forward example, this measure could include the distance from the area centroid to the nearest COVID-19 testing site. For each small area, we also have complete information (or estimates) of various demographic subgroup population counts. For the encompassing geographic area of interest (e.g., the metropolitan Atlanta area), we then construct ECDFs of access within each demographic subgroup in order to assess how the distributions of access compare. In our analysis, we assess both the metropolitan Atlanta area as well as Fulton County, the most populous county and home to the majority of the city of Atlanta. We consider both public testing sites as well private testing sites such as drive-thru locations.

We build on the approach laid out by Waller et al. (1997, 1999) by considering a more elaborate measure of spatial access in the COVID-19 testing context. In addition to distance to nearest testing site, we consider a potential demand measure for nearby testing sites that has recently been used in an analysis by FiveThirtyEight (Bronner, 2020). This potential demand measure accounts for situations more complex than a simpler, naive distance-based measure, and it is related to two-stage catchment area and gravity-based model measures that account for population demand (Guagliardo, 2004; Apparicio et al., 2017). For example, if there is a centrally-located testing site in a population-rich area but no other testing sites nearby, distance to nearest testing site would imply easy access, while the second measure we construct would adjust for

the very large potential patient demand and dearth of other testing sites.

Next, we consider a Monte Carlo approach for assessing whether the observed ECDFs for different demographic subgroup partitions are consistent with different testing site sampling schemes. In practice, observed ECDFs may be a result of numerous factors driving the placement of testing sites. In determining where to place public testing sites, for example, county and city governments may work with underserved community groups to ensure better access. Finally, we propose several ideas for future extensions and research, including using additional measures of spatial access, accounting for the underlying uncertainty in population estimates from the American Community Survey, incorporating spatial equity measures into optimization procedures for finding candidate locations for mobile testing site locations, extending data collection across time and analyzing spatio-temporal placement of testing sites, and incorporating various indicators of COVID-19 incidence. Given the importance of research related to COVID-19, we emphasize that the data collection and analyses presented here are on-going and intended to provide a foundation for future extensions to assess spatial access.

3.2 Data

3.2.1 Testing sites

Our focus is on testing sites in the Atlanta area, and we collect information on testing sites from Fulton County as well as the broader metropolitan Atlanta area. We divide testing sites into four categories: (1) county or Georgia Department of Public Health (DPH) free testing sites, sometimes in collaboration with the non-profit organization Community Organized Relief Effort (CORE); (2) sites that have partnered with com-

munities or governments to provide free testing[1]; (3) health centers as defined by the Health Resources & Services Administration (HRSA) Health Center Program[2] that report COVID-19 testing; (4) other non-profit or charity sites; (5) private testing sites, such as urgent care facilities or other sites like drive-thrus or curbside testing sites.

Although there do exist publicly accessibly repositories with testing site location data (see, e.g., URISA's GISCorps, Coders Against COVID (findcovidtesting.com), and Esri (2020), hereafter referred to as GISCorps data), due to the diverse collection of companies and agencies providing testing and the changing set of locations over time, we find occasional issues with these repositories, such as the inclusion of sites that are no longer in operation. For our work, data on public testing sites were manually obtained from county websites, CORE, public health district websites, and the Georgia DPH website for the week of September 28 through October 3, 2020[3][4]. These testing sites were open for at least 1 day during this target week. Currently, our analysis does not take into account the number of days each testing site was in operation, nor do we collect data on testing site capacity, although future work may include this information. The Appendix contains a list of websites that were used for this process.

Data for health centers were obtained from the HRSA website[5] on September 29, 2020 within 100 miles of midtown Atlanta if the sites confirmed providing COVID-19

[1]These sites include 4 Walmart partnerships with eTrueNorth and a CVS community partnership site

[2]According to the HRSA "Find a Health Center" tool, "health centers provide services regardless of patients' ability to pay and charge for services on a sliding fee scale."

[3]Fulton County's website at `https://www.fultoncountyga.gov/covid-19/covid-testing-sites` provides the most comprehensive listing of testing sites (including CORE sites) for the upcoming week. Many of these sites are not listed in the Georgia DPH site – furthermore some site dates and locations are revised as the week goes on. We initially collected information on testing sites on September 28, 2020 and revised this on October 1, 2020.

[4]Currently (as of October 13, 2020), the Georgia DPH Testing website (`https://dph.georgia.gov/covidtesting`) now embeds Castlight's searchable map, which surfaces public and private testing sites.

[5]`https://findahealthcenter.hrsa.gov/tool`

testing. When a health center operator reported multiple sites at the same address, only one was kept. We note that being a HRSA health center does not guarantee free COVID-19 testing[6].

Collecting information on a plausible set of non-profit and private testing sites relied on several sources of information. Broadly, our goal was to include sites that performed some form of "on-demand" testing. First, we collected information on urgent care clinic providers that reported COVID-19 testing in the metropolitan Atlanta area. Several of these operators reported drive-thru or curbside testing to accommodate patients seeking tests. These sites included urgent care networks such as Peachtree Immediate Care (31 sites), Piedmont Urgent Care by WellStreet (17 sites), Wellstar Urgent Care (16 sites), Northside Family Medical & Urgent Care (6 sites), the Northeast Georgia Physician's Group Urgent Care (6 sites), and others. We additionally included several drive-thru and pharmacy sites that reported COVID-19 testing, including CVS[7] (146 sites in Georgia), Kroger Health's The Little Clinic (18 sites), Walgreens (9 sites), and Walmart in collaboration with eTrueNorth or Quest diagnostics (13 sites).

We also used additional sources of information on COVID-19 testing to add additional testing sites. First, we used Google Maps and Castlight's list of COVID-19 testing sites for various ZIP codes in the Atlanta area to identify other potential testing sites that may have been missed, after verifying their validity. Second, we used a list compiled by Gwinnett, Newton, and Rockdale counties[8] for additional free and private testing sites in the metro Atlanta area. Finally, we used a list of compiled

[6]As an example, MedLink is an HRSA health center, but their website notes that "COVID-19 oropharangyl or nasopharangyl test is $80." This is a lower out-of-pocket cost than most urgent care centers. See `https://web.archive.org/web/20201113222444/http://www.medlinkga.org/local.cfm?id=124`

[7]CVS notes on their website that COVID-19 tests are free for eligible uninsured persons under a Federal program, as a result of the CARES act. Future work may focus on re-categorizing such nominally "private" testing sites in a larger set of free testing sites for analyses.

[8]See `https://www.gnrhealth.com/wp-content/uploads/2020/09/COVID-Testing-09.08.20-3.pdf`

testing sites from GISCorps after removing sites already contained in our database and checking the remaining sites for validity.

Our compiled set of testing sites has its limits which may impact analyses. We chose not to include testing sites that only tested certain populations such as employee screening sites or VA clinics, nor did we include hospitals or medical centers, as they are likely to only perform testing in the course of performing other forms of medical care. Finally, we did not include a comprehensive list of doctor's offices that provide COVID-19 testing, although this may be an avenue many symptomatic or exposed persons take[9]. The HRSA health centers include a variety of different kinds of facility types, and we include all of these[10]. Outside of the HRSA health centers, several primary care clinics and family medicine clinics reported doing COVID-19 testing, but we choose not to include these testing sites in our analysis set[11]. For county/CORE sites, where possible, we used the latitude/longitude provided by the Georgia DPH testing map – otherwise these addresses were geocoded along with the other testing sites[12]. We consider the process for constructing the set of testing sites as preliminary and any conclusions based on these should bear these limitations in mind.

3.2.2 Population and geographic area

Population information is obtained from the American Community Survey (ACS) 5-year estimates for 2014-2018 at the Census block group level (Manson et al., 2020).

[9]The Georgia DPH website says, "You can seek a COVID-19 test at your doctor's office."

[10]Future work may choose to limit to certain kinds of HRSA health centers after a more comprehensive investigation.

[11]For example, Wellstar reports "Wellstar is conducting COVID-19 testing across all hospitals, health parks, offices and urgent care locations. All Wellstar physicians can refer patients for screening and COVID-19 testing, and anyone experiencing symptoms should contact one of the above locations." However, we only include the Wellstar urgent care locations rather than all primary care offices, as these are not likely to be straightforward avenues for persons to get quickly tested.

[12]Counties in the North Georgia Health District, which includes Cherokee, Fannin, Gilmer, Murray, Pickens, and Whitfield, do not have their addresses published on the their website `https://www.nghd.org/pr/34-/1175-update-now-3-locations-for-free-covid-19-testing-in-north-ga.html`. We use the County Health Department addresses, but we acknowledge some potential error here.

Although race and Hispanic ethnicity are not mutually exclusive, we use estimates of the total population and mutually exclusive groups of black non-Hispanic, white non-Hispanic, and Hispanic persons. From the ACS, we also obtain block group estimates of persons above the age of 19 with and without health insurance, and estimates of the number of persons living below and above the poverty level. From IPUMS NHGIS, we additionally obtain shapefiles for block groups, counties, and the Atlanta 2010 Census-defined urbanized area (UA). We consider testing site analyses focusing on just the Atlanta UA area, as well as an analysis focused on the Fulton County area. Census urban area classifications help limit our analysis to more densely populated areas, where distance to a testing site can more plausibly serve as a proxy for spatial access (Bronner, 2020). We use the 2010 Census centers of population for block group locations.

3.3 Methodology for assessing testing site inequity

3.3.1 Group-specific ECDFs of spatial access

Following Waller et al. (1997, 1999), consider a spatial access measure, x_i, for persons in Census-defined area i. That is, we assume that the measure is the same for all persons within an area, and we define the measure based on the center of population. In the context of the current study, this may be some function of distance to testing sites from the block group center. We return to possible definitions of the spatial access measure shortly. In this formulation, we construct group-specific ECDFs of access,

$$G_j(x) = \frac{\sum_{i:x_i \leq x} n_{ij}}{\sum_i n_{ij}} \tag{3.1}$$

for each demographic subgroup j, where n_{ij} denotes the number of persons in group j in area i provided by the Census. In the subsequent analysis, we consider race/ethnicity

groupings, i.e., $j \in \{\mathrm{B}, \mathrm{W}, \mathrm{H}\}$ for black non-Hispanic, white non-Hispanic, and Hispanic, respectively, as well as above or below the poverty level, and insured or uninsured. By constructing these group-specific ECDFs, we are able to consider differences at particular access thresholds (e.g., $G_W(x') - G_B(x')$ for some access measure x'), as well as integrated differences between ECDFs, $\int_0^\infty (G_W(x) - G_H(x))dx$. For example, these group-specific ECDFs allow us to answer what proportion of the underlying group population is within 2 km of a testing site.

3.3.2 Spatial access definitions

A large literature exists in the area of health services and health geography research examining different forms of spatial access to healthcare, including distance to nearest site, two-step floating catchment areas, and gravity-based model measurements (see Guagliardo (2004); Apparicio et al. (2017); Luo and Qi (2009) for a non-exhaustive review of these methods). There are many factors to consider in such spatial access measures, such as the type of distance (Euclidean or travel distance), the spatial unit of reference, and the measurement itself (Apparicio et al., 2017). In the framework of Guagliardo (2004), our focus is on measures of *potential spatial* access. In this framing, we consider the distance and availability of testing sites, but we do not consider utilization measures.

For this analysis, we consider two measures of spatial access. First, we consider a simple distance to nearest testing site measure. For area center i and testing site $k = 1, ..., K$, denote the distance as d_{ik}. Then the spatial access measure is defined as

$$x_i = \min_k d_{ik}. \tag{3.2}$$

Despite being relatively easy to interpret, nearest distance measures have known problems for measuring spatial access in urban areas, as they do not account for

congestion or population demand for testing sites (Guagliardo, 2004).

We also consider a more elaborate spatial access measure that takes into account potential demand at testing sites. As an example, a centrally-located testing site may be relatively close to a large population, but if it is the sole testing site for those persons, then there is likely to be congestion and community need for a greater number of testing sites. This second measure is used in recent reporting by FiveThirtyEight (Kim et al., 2020; Bronner, 2020), and we attempt to replicate the procedure in the following steps, with some modification.

1. For each area i, we define the nearest 10 testing sites by the set R_i. The assumption is that persons in area i will seek a test only at sites $k \in R_i$. We additionally (and deviating slightly from the FiveThirtyEight analysis) assume that persons will only seek testing sites within 40 kilometers (≈ 24.85 miles) of their block group centroid. Thus, R_i may be a set consisting of 10 or fewer testing sites.

2. We assign the population n_i for an area to the sites in R_i in a manner inversely proportional to distance. Thus, we can define weights $w_{ik} \propto \frac{1}{d_{ik}}$ or $w_{ik} \propto \frac{1}{d_{ik}^2}$, such that $\sum_{k \in R_i} w_{ik} = 1$. For the remainder, we consider the latter form, where the weights are inversely proportional to the squared distance[13]. The number of persons allocated to each test site k from area i is then denoted as $m_{ik} = w_{ik} n_i$.

3. After assigning every area's population to testing sites based on distance, we calculate the number of persons allocated to each test site k as

$$m_k = \sum_i w_{ik} n_i = \sum_i m_{ik}.$$

This measure for the testing site is called the **potential patient demand**.

[13]Work is on-going to assess sensitivity to using $\frac{1}{d_{ik}}$ instead. In general, $\frac{1}{d_{ik}^2}$ will place greater weight on closer testing sites than $\frac{1}{d_{ik}}$.

4. The access measure derived from this at the area level is the **potential community need**, or the average potential demand for nearby testing sites (c_i). To calculate this value, a weighted average of the potential patient demand at nearby testing sites is taken. The weights here are the same as before, w_{ik}, which are inversely proportional to distance or squared distance. If we let w_{ik} be equal to 0 for sites $k \notin R_i$, then we have:

$$c_i = \sum_k w_{ik} m_k.$$

Thus, this value represents a weighted average of the potential patient demand at nearby testing sites.

The proposed measure has much in common with other gravity model-based measures in the spatial access literature that attempt to account for population demand (Guagliardo, 2004; Apparicio et al., 2017; Luo and Qi, 2009). In these methods, the distance decay coefficient (determining the spatial weights w_{ik}) is sometimes informed by data and may vary by location, although we consider only squared distance here. A shortcoming of the measure considered here is that we do not account for variation in testing site capacity or days of operation. We consider incorporating operating schedule information and additional spatial access measures in future work.

3.3.3 Application to Atlanta-area data

At present time, we have constructed a set of testing sites for the metropolitan Atlanta area consisting of both public testing sites (i.e., county/CORE), health centers that provide COVID-19 testing, community partnerships, and private testing sites for the week of September 28, 2020. We posit that the wealthier areas of Fulton and Cobb counties may depend less on free public testing sites, and policymakers may in turn determine the placement of free public testing sites in response to where there is a

greater economic need and a lack of health insurance. To this end, we construct the nearest-distance and potential demand measures for three progressively larger sets of testing sites:

1. **Public**: County/CORE testing sites;
2. **Public** + **HRSA**: In addition to the above, we additionally include community partnership sites and HRSA health centers;
3. **Public** + **Private**: In addition to the above, we additionally include the remaining private testing sites, as defined in Section 2.

As mentioned previously, a shortcoming of this study is that we largely lack testing site capacity information, and we do not draw any distinction between sites that are operating 5 or 6 days, for example, and pop-up sites that are operating on a single day.

We consider two geographic areas as previously mentioned. First, we conduct an analysis that is limited to areas with at least 50% of their block group area in the 2010 Census-defined Atlanta UA, which contains the majority of Fulton, DeKalb, Cobb, and Gwinnett counties as well as parts of surrounding counties. Second, we consider an area consisting of all block groups in Fulton County. Importantly, we only make these area restrictions *after* the access values are estimated. For each of the two spatial access measures and each of the two geographic area definitions, we calculate the ECDFs among three distinct partitions of the underlying population:

1. Race/ethnicity by considering block group estimates of black non-Hispanic, white non-Hispanic, and Hispanic persons;
2. Poverty status, by considering estimates of persons above and below the poverty level;

3. Health insurance status, by considering estimates of persons above the age of 19 with and without health insurance.

3.3.4 A Monte Carlo approach to assessing ECDF curves

We extend the ECDF comparison approach in the Atlanta area by adding an additional visual tool for assessing test site access. The observed testing site placements are likely the result of targeting of particular communities as well as financial and physical constraints by county and city governments. Thus, differences in the observed ECDF curves for demographic subgroups may be consistent with some plausible pattern of testing site placement. To assess whether this is the case, we consider different simulation-based approaches for selecting testing sites and re-calculating the two spatial access measures and resulting ECDF curves repeatedly.

To focus this exercise, we limit our attention to an ECDF analysis of Fulton County using the county/CORE public testing sites only. We assume that testing sites outside of Fulton County are fixed, but that the 30 testing sites inside Fulton County can be placed at the centroid of any of the constituent 544 block groups[14]. We consider two separate sampling schemes:

1. **Population-based**: Sample block group centroids without replacement using the population share of Fulton County as the probability of being selected.

2. **Poverty-based**: Sample block group centroids without replacement using the block group's share of persons living under the poverty level as the probability of being selected.

We conduct this Monte Carlo exercise 200 times and we plot the resulting ECDF curves for comparison with the observed ECDF curves.

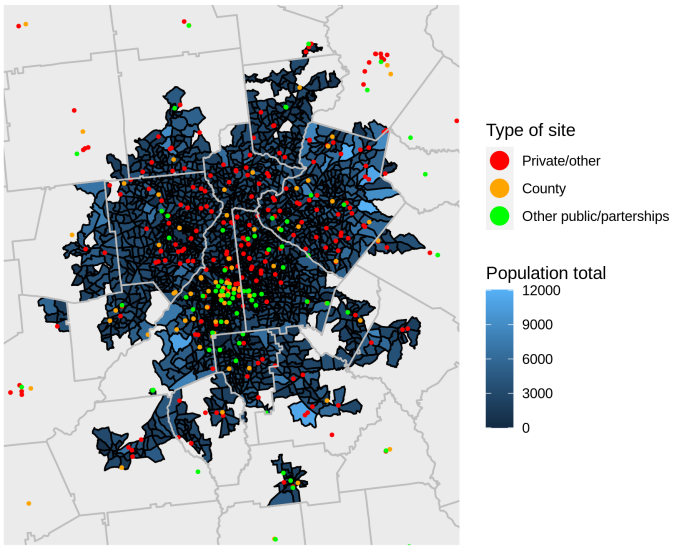
[14]I consider making the testing site placement more flexible in the future

3.4 Results

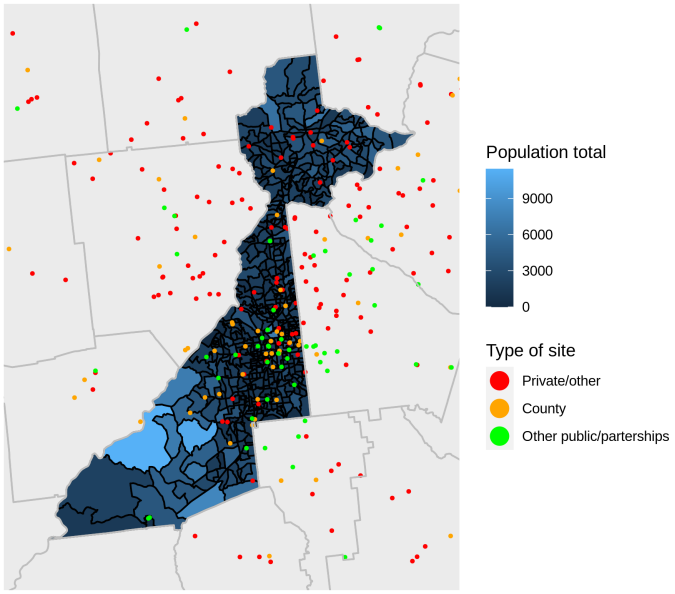
In the Atlanta UA, there were 2132 block groups, while in Fulton County, there were 544 block groups. Figure 3.1 shows the block groups and surrounding testing site locations for these two study areas of interest. Fulton County has a large number of public testing sites relative to other areas in the metropolitan Atlanta area. There is a large cluster of testing sites in the south-central part Fulton County and part of DeKalb county, while other areas have a sparser set of public and HRSA health center sites, partially filled in with other types of testing sites.

The first spatial access measure under consideration, distance to nearest testing site (Figure 3.2), shows that black non-Hispanic and Hispanic persons are actually more likely to be in close proximity to public testing sites than white non-Hispanic persons in the Atlanta UA. Limiting the geographic scope to Fulton County, the relative advantage for black non-Hispanics remained while Hispanics were more similar to non-Hispanic Whites (Figure 3.2b). When expanding to the additional testing sites such as HRSA health centers, these patterns are largely the same (Figures 3.2c and 3.2d). For the full set of public and private testing sites, however, differences between the different racial and ethnic groups appear to be substantially less. Based on distance to nearest testing site, this analysis suggests that public testing sites are specifically targeting under-served populations, while other forms of testing sites are largely serving other communities.

When using the modified FiveThirtyEight measure in Figure 3.3, we find broadly similar but more exaggerated patterns. In particular, for Fulton County, the non-Hispanic Black population is considerably better situated for access to public testing sites (Figure 3.3b). This pattern once again is substantially diminished when we expand to the full set of testing sites, suggesting some strategic placement of the public testing sites towards more under-served areas. In the Appendix, we find broadly similar patterns, with public testing sites favoring uninsured persons and persons under



(a) Block group populations and testing sites in Atlanta UA



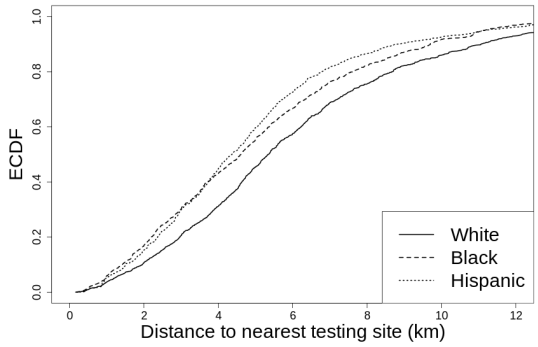
(b) Block group populations and testing sites in Fulton

Figure 3.1: Block groups and testing sites in two areas of interest: (1) Atlanta UA and (2) Fulton County; population totals from 2014-2018 ACS.

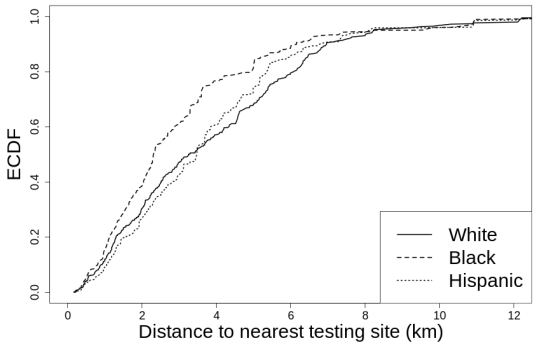
the poverty level (Figures C.1-C.4). When the FiveThirtyEight measure is applied to a subset of the testing sites, as in Figures 3.3a-3.3d, there are some limitations in interpretation of the measure. This potential demand measure is constructed by assuming all persons are allocated to only the subset of testing sites considered; in practice some persons will go to public testing sites, while others will seek tests at drive-thrus or other pop-up testing sites. Even in Figures 3.3e-3.3f, the full set of testing sites included in our analysis will exclude doctor's offices, employee screening sites, and hospitals, as mentioned in Section 2. We think it is plausible that if one were to include all of these additional testing sites, there would likely be more parity between the subgroup ECDFs in Figure 3.3f. Nonetheless, we highlight the *change* in ECDF differences across different sets of testing sites, which suggest that public and private testing sites are filling in gaps in patient demand with respect to each other.

The Monte Carlo analysis for the public set of testing sites in Fulton County demonstrates that the pattern of testing sites is much better explained by the poverty-based placement of testing sites. For the nearest distance access measure, Figure 3.4b shows that when placing testing sites proportional to block group counts of persons under the poverty level, the observed ECDF lines for white non-Hispanics, black non-Hispanics, and Hispanics are roughly in line with the ECDF lines based on the Monte Carlo simulations. Similar patterns hold for the poverty level and health insurance status (Figures 3.4c-3.4f). Similarly, the Monte Carlo analysis for the FiveThirtyEight measure shows substantially more visual agreement when using the poverty-based placement of test sites rather than the population-based approach. These results give additional evidence to Fulton County strategically placing testing sites in under-served areas.

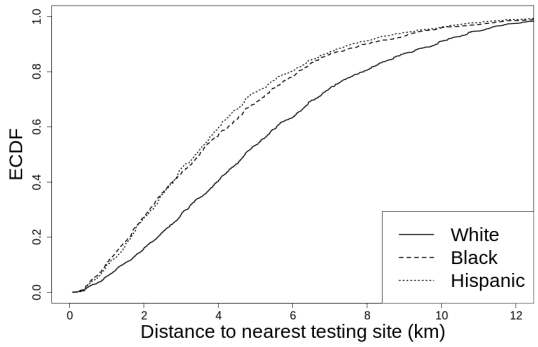
Figure 3.6 shows majority black, white, and Hispanic block groups, respectively, in the Atlanta UA and Fulton County areas to help illustrate patterns in the ECDF analyses for the potential demand measure. Predominantly white non-Hispanic block



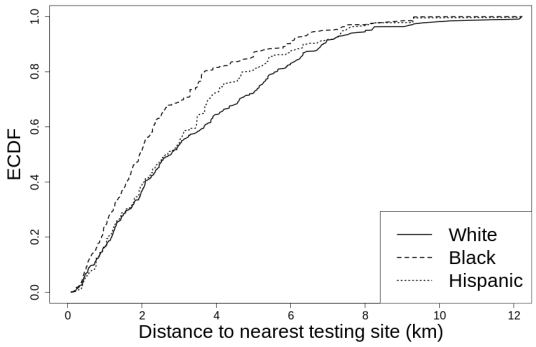
(a) Atlanta UA, county sites only



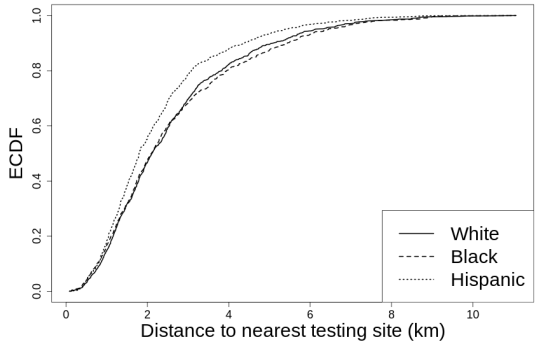
(b) Fulton County, county sites only



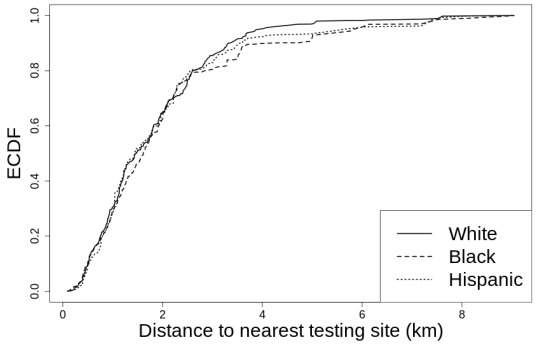
(c) Atlanta UA, county and other publicly funded sites



(d) Fulton County, county and other publicly funded sites

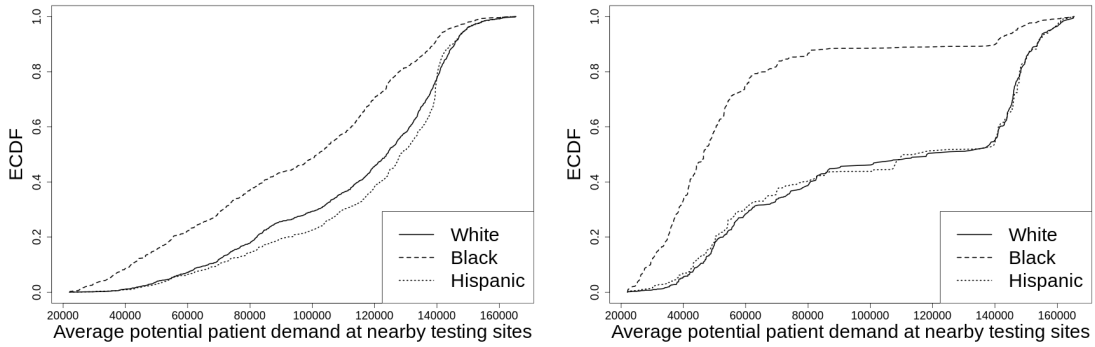


(e) Atlanta UA, all testing sites



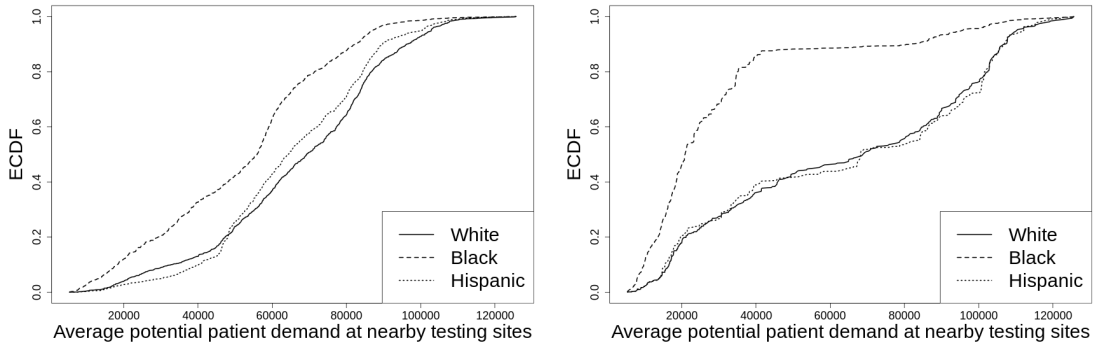
(f) Fulton County, all testing sites

Figure 3.2: ECDF comparisons for distance to nearest testing site among white non-Hispanic, black non-Hispanic, and Hispanic persons in the Atlanta UA (left column) and Fulton County (right column), for public sites (top row), public sites together with other community and HRSA health center sites (middle row), and all public and private testing sites together (bottom row).



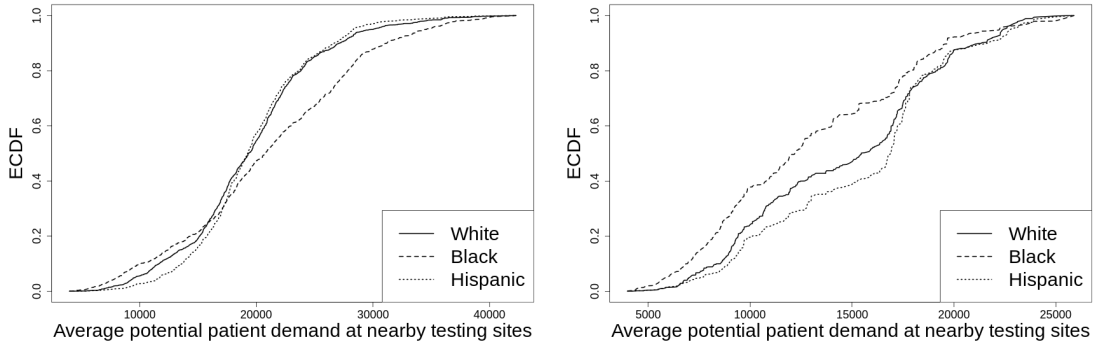
(a) Atlanta UA, county sites only

(b) Fulton County, county sites only



(c) Atlanta UA, county and other publicly funded sites

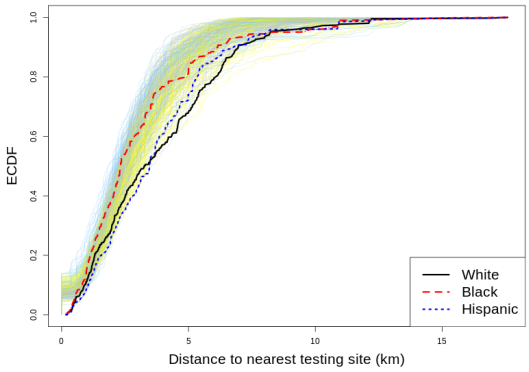
(d) Fulton County, county and other publicly funded sites



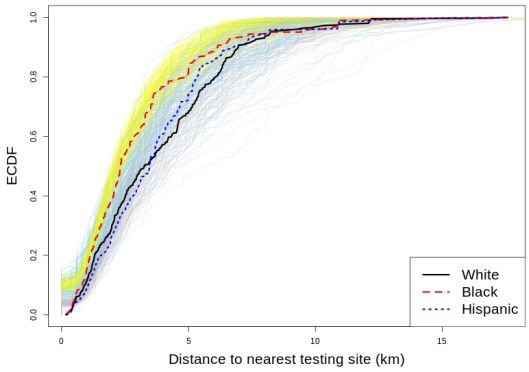
(e) Atlanta UA, all testing sites

(f) Fulton County, all testing sites

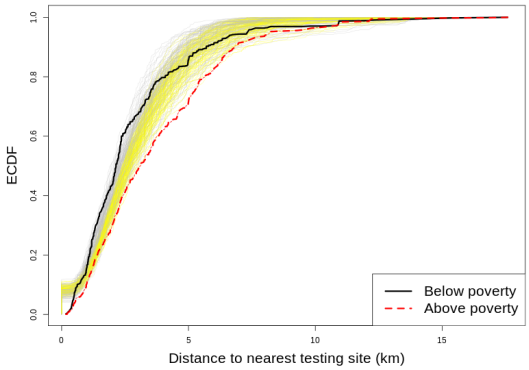
Figure 3.3: ECDF comparisons for potential demand at nearby testing sites among white non-Hispanic, black non-Hispanic, and Hispanic persons in the Atlanta UA (left column) and Fulton County (right column), for public sites (top row), public sites together with other community and HRSA health center sites (middle row), and all public and private testing sites together (bottom row).



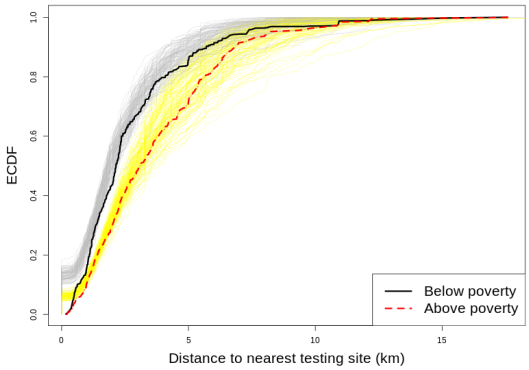
(a) Population-based placement of testing sites



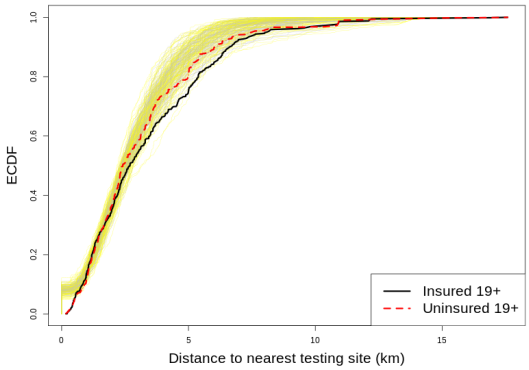
(b) Poverty-based placement of testing sites



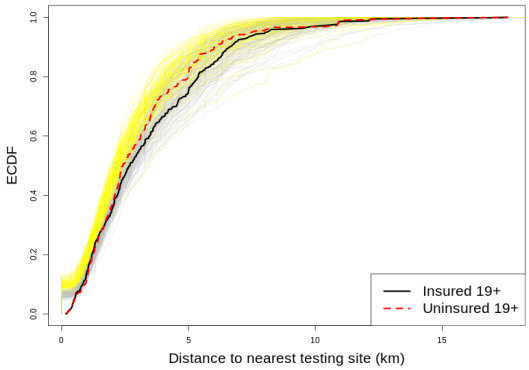
(c) Population-based placement of testing sites



(d) Poverty-based placement of testing sites

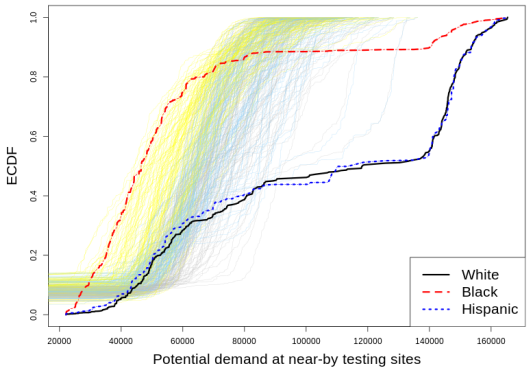


(e) Population-based placement of testing sites

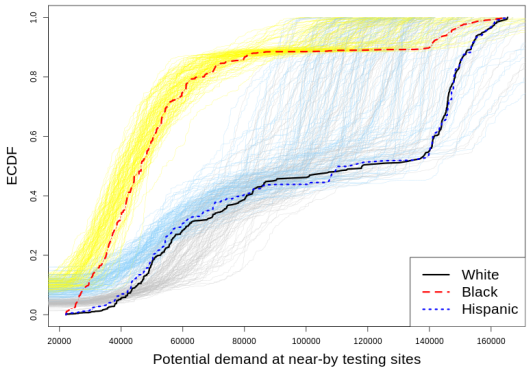


(f) Poverty-based placement of testing sites

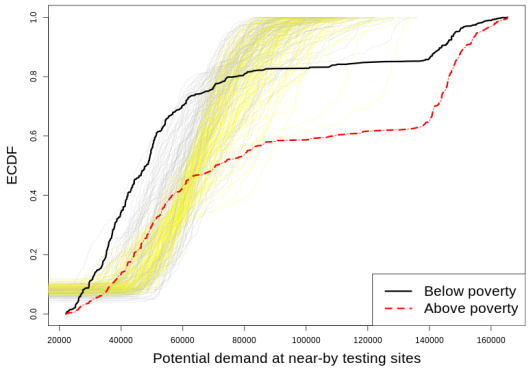
Figure 3.4: Results of Monte Carlo sampling schemes on ECDF for distance to nearest testing site, for sampling scheme using basing test site placement on underlying population (left column) and underlying population of those under the poverty level (right column). The gray, yellow, and sky blue lines denote the Monte Carlo ECDF lines for black, red, and blue observed ECDF lines, respectively.



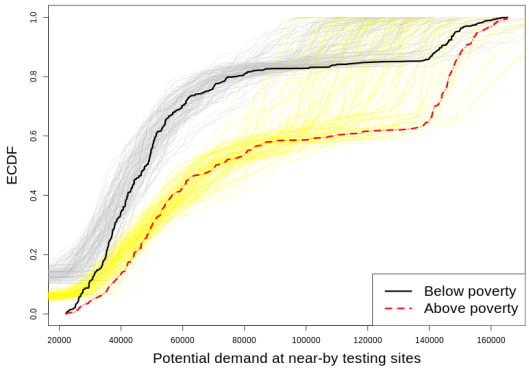
(a) Population-based placement of testing sites



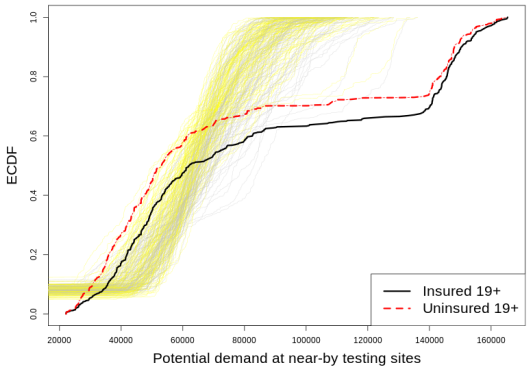
(b) Poverty-based placement of testing sites



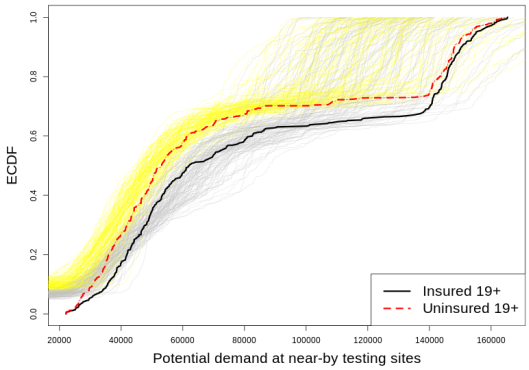
(c) Population-based placement of testing sites



(d) Poverty-based placement of testing sites



(e) Population-based placement of testing sites



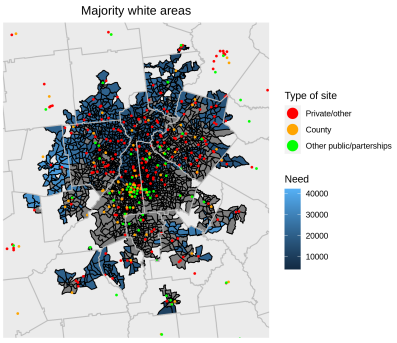
(f) Poverty-based placement of testing sites

Figure 3.5: Results of Monte Carlo sampling schemes on ECDF for potential demand at nearby testing sites, for sampling scheme using basing test site placement on underlying population (left column) and underlying population of those under the poverty level (right column). The gray, yellow, and sky blue lines denote the Monte Carlo ECDF lines for black, red, and blue observed ECDF lines, respectively.

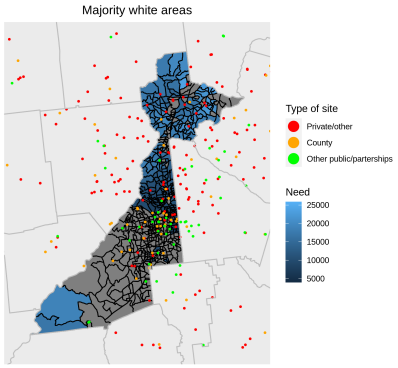
groups on the western part of Atlanta UA in Paulding County appare to have few testing sites, while areas in Cobb and North Fulton are apparently well-served by a collection of predominantly private testing sites. Black areas outside of the urban core of Atlanta in southeast DeKalb and Rockdale Counties have very few testing sites.

3.5 Discussion

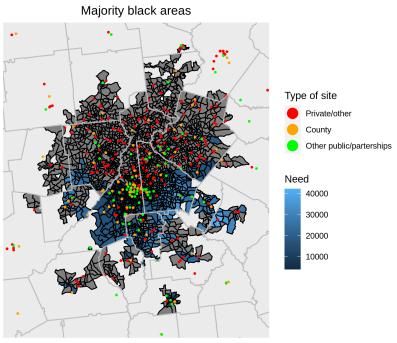
While providing insight into patterns of placement of different types of testing centers, several limitations of the analysis remain and offer room for future improvement. Several of these limitations are a result of the set of testing site data collected and the measures used. First, although we made every effort to include a reasonable set of public and private testing sites, we may have missed certain testing sites. Second, we do not account for testing site capacity or days of operation of the testing sites in question. This may, as a result, overstate the relative access of black non-Hispanics in Fulton County, as many testing sites operate on one or two days in a week. Third, the FiveThirtyEight measure may be inappropriate for application to subsets of testing sites; implicit in the calculation is that everyone in the nearby block groups must obtain COVID-19 tests at public testing sites, for example, which may skew these measures. Future work may examine analyses that allocate portions of the block group populations to different classes of testing sites in constructing measures, as well as considering different distance-based weights for different areas depending on population density, transit access and vehicle ownership. Fourth, we did not include a number of testing sites, such as employee screening sites, VA clinics, and primary care providers. Finally, we do not account for differences in disease burden in different demographic subgroups in examining test site spatial access. Understanding disproportionate disease burden in particular demographic subgroups may be important for



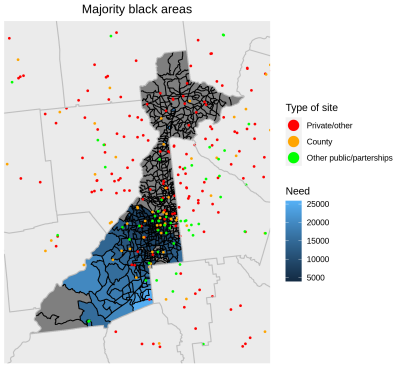
(a) Majority white



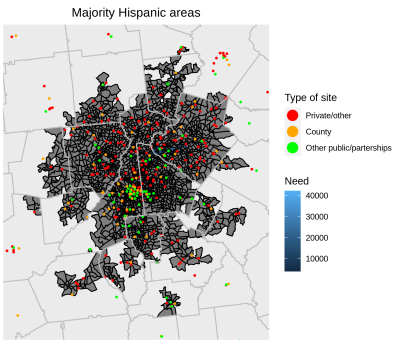
(b) Majority white



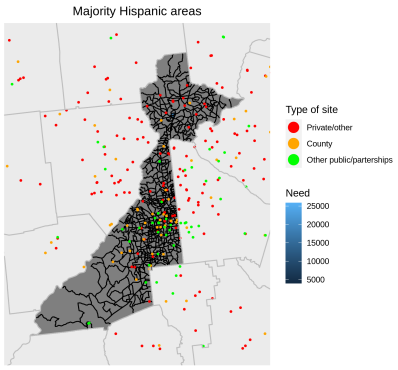
(c) Majority black



(d) Majority black



(e) Majority Hispanic



(f) Majority Hispanic

Figure 3.6: Potential patient demand for majority black, white, and Hispanic block groups for Atlanta UA (left column) and Fulton (right column)

understanding any inequity in observed testing site access. For example, differences in disease burden in different race and ethnicity groups may partially explain differences in public testing site access if communities and governments are attempting to place testing sites where there are the most new cases of COVID-19.

Additional approaches also exist for analyzing potential spatial access. In particular, Stamm et al. (2017) analyze spatial access to vaccines in response to the H1N1a influenza virus in 2009, using detailed data on the availability of flu vaccine supplies and service providers. They measure spatial access through an optimization framework; this approach has the benefit of allowing the choice of one person (e.g., obtaining a vaccine at a particular location) to impact the rest of the system.

3.6 Future extensions

We highlight several areas for possible extensions to the above analysis.

3.6.1 Optimization methods for determining testing site placement

Having assessed ECDF differences, we may also begin to inquire about the *optimal* placement of new testing sites. Building upon Waller et al. (1997, 1999), we may consider moving or adding an additional testing site in order to address both spatial access and overall coverage. Define the optimization problem as:

$$\max \Big\{ \sum_{j=1}^{m} e_j - \lambda |D(G_W(x_i), G_B(x_i))| \Big\}. \tag{3.3}$$

Let e_j denote whether block group j is covered by a testing site or not, and λ denotes a term weighting the importance of spatial equity. $D(G_W(x), G_B(x))$ denotes the difference between the two race-specific CDFs and x is the access measure (nearest-

distance to testing site, or potential demand at nearby testing sites). The difference can be approximated as a discrete sum of rectangular areas:

$$D(G_W(x), G_B(x)) = \sum_{l=1}^{L} \{G_W(x_l) - G_B(x_l)\}(x_l - x_{l-1}). \tag{3.4}$$

3.6.2 Incorporating uncertainty about block group population estimates

Constructing the ECDFs requires knowledge about Census block group population characteristics. Previously, we assumed n were known (omitting subscripts for the time being); now we suppose that we instead have estimates $\hat{n}$. We assume a model such that true values are drawn from a (possibly truncated) normal distribution:

$$n \sim N(\hat{n}, \hat{\sigma}^2).$$

We could imagine that this model could also incorporate spatial correlation. In words, if we think a census block group under-counted blacks relative to the truth, then neighboring block groups probably also under-counted, so that the errors are correlated. We consider three basic error models, where we observe $\hat{n}_{ACS,ij}$ for area i from the ACS and subgroup j, along with the ACS margin of error, MOE_{ij}. We assume that the true population count could have one of the following structures:

- **Marginal** model:

$$n_{ij} \sim N\Big(\hat{n}_{ACS,ij}, \Big(\frac{MOE_{ij}}{1.645}\Big)^2\Big) \tag{3.5}$$

- **Spatially correlated counts** model:

$$\boldsymbol{n}_j \sim N\Big(\hat{\boldsymbol{n}}_{ACS,j}, \boldsymbol{\Sigma_{0,j}}\Big), \tag{3.6}$$

where diagonals of $\mathbf{\Sigma_{0,j}}$ are $\left(\frac{MOE_{ij}}{1.645}\right)^2$ and off-diagonals reflect correlation between counts in two areas. These correlations could be estimated using neighbors or some distance function, possibly calibrated using the 2010 Census and 2008-2012 ACS, for example.

- **CAR** model:

$$n_{ij} = \hat{n}_{ACS,ij} + \epsilon_{ij} + v_{ij}, \tag{3.7}$$

where $\epsilon_{ij} \sim N\left(0, \left(\frac{MOE_{ij}}{1.645}\right)^2\right)$ denotes an independent error term and v_{ij} denotes the random effect with the CAR prior.

The analytical approach here is to use Monte Carlo methods for simulating estimates of population size for Census block groups, and re-calculating the ECDF curves, as well as the integrated differences, for each sample draw. Thus, we can propagate uncertainty about the population counts into the ECDF analyses.

3.6.3 Spatio-temporal extensions

With weekly data on public testing sites and tracking changes in private testing sites, we would be able to access changes in testing site spatial access over time, as well as contribute to crowdsourced projects that will improve access to data for other researchers (in particular, see the GISCorps data). However, collecting the public testing site data is labor intensive. Future work may include contacting the Georgia DPH and county health departments to obtain testing site locations and schedules directly.

3.6.4 Incorporating disease case data into analyses and optimization

An optimization framework for placing new testing sites would ideally incorporate disease data as well as spatial access to testing sites. Sub-county data on COVID-19 are somewhat limited, however. Fulton and DeKalb Counties provide the total and new number of positive cases in each ZIP code area. Figure 3.7 shows the total number of cases as of August 11 to 12, 2020 in ZCTAs in Fulton and DeKalb Counties. There are several issues with these data. In particular, the epidemiological reports from the counties do not appear to release the total number of tests, so we only have access to the number of positive tests, with no small area data on the positivity rate. Nonetheless, we may use increases in the counts of cases at the ZIP code level, together with spatial access measures and the existing set of testing sites, to consider where to place additional mobile testing sites. These approaches could lead to better surveillance while ensuring fairness in testing site access.

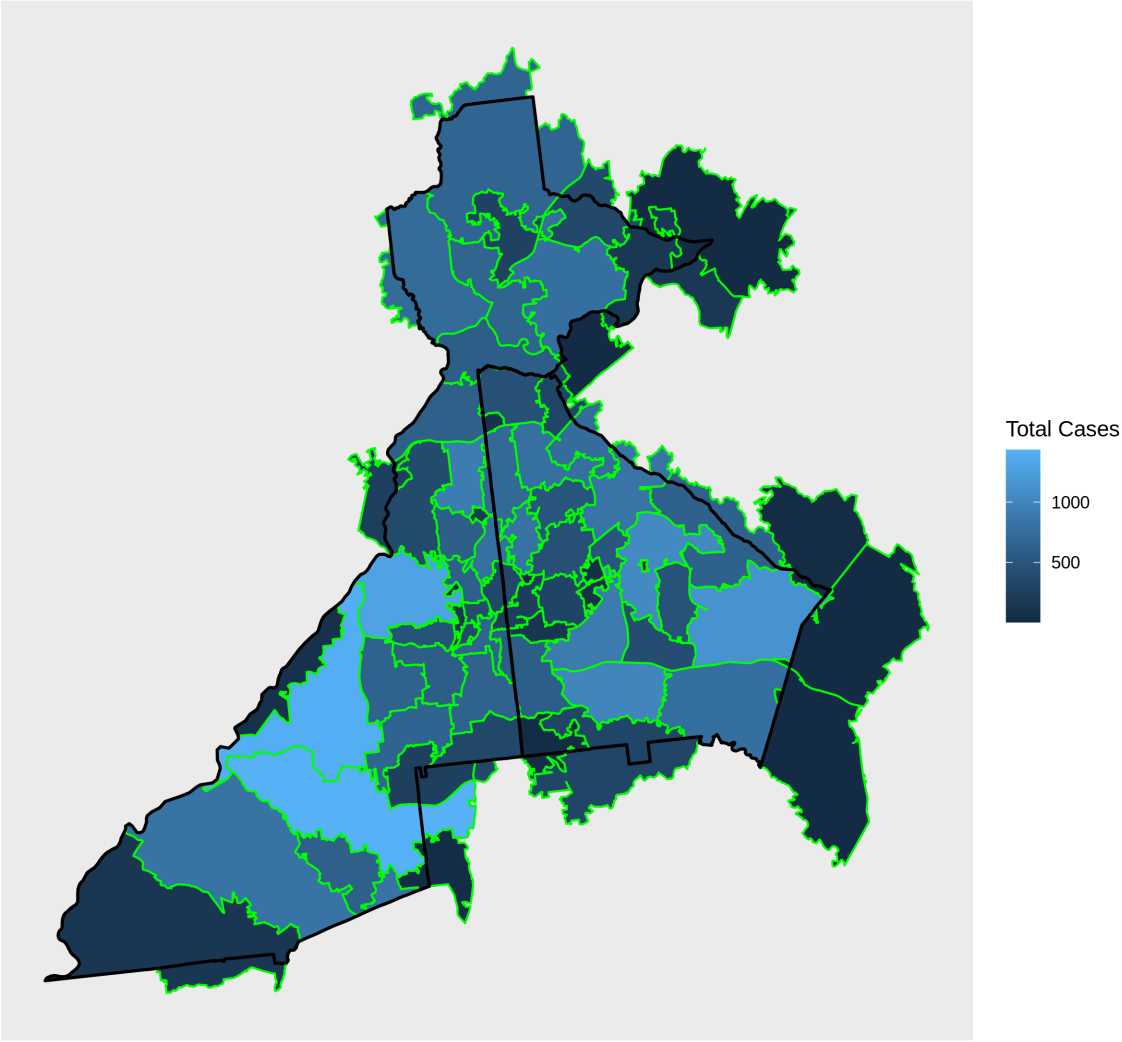


Figure 3.7: Total COVID-19 cases as of August 11-12, 2020. Based on data from DeKalb and Fulton county epidemiological reports; data are preliminary and subject to change.

Appendix A

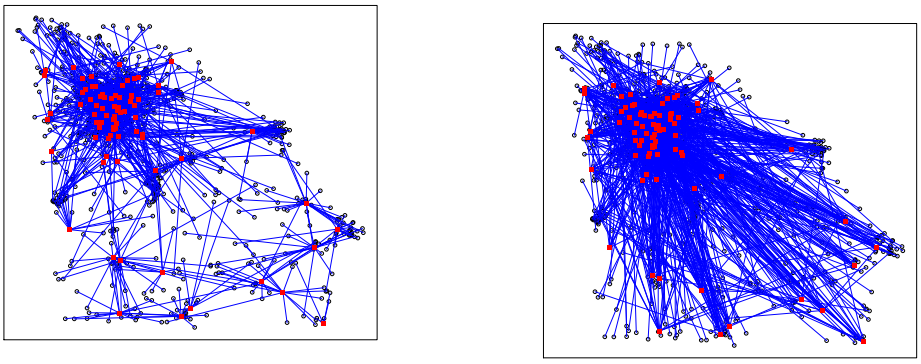
Supplemental Materials to "Propensity score matching for multi-level and spatial data"

A.1 Simulation Study

A.1.1 Data generation

Two scenarios were considered for patient assignment to facilities. Figure A.1 shows the visual pairing of patients to facilities.

Figure A.2 demonstrates the different Matern patterns utilized for the generation of the the unobserved spatial covariate U from an example dataset.



(a) Distance related to facility assignment

(b) Random facility assignment

Figure A.1: Two background settings for facility assignment
Blue line segments (–) indicate assignment of patients (∘) to facilities (▪).

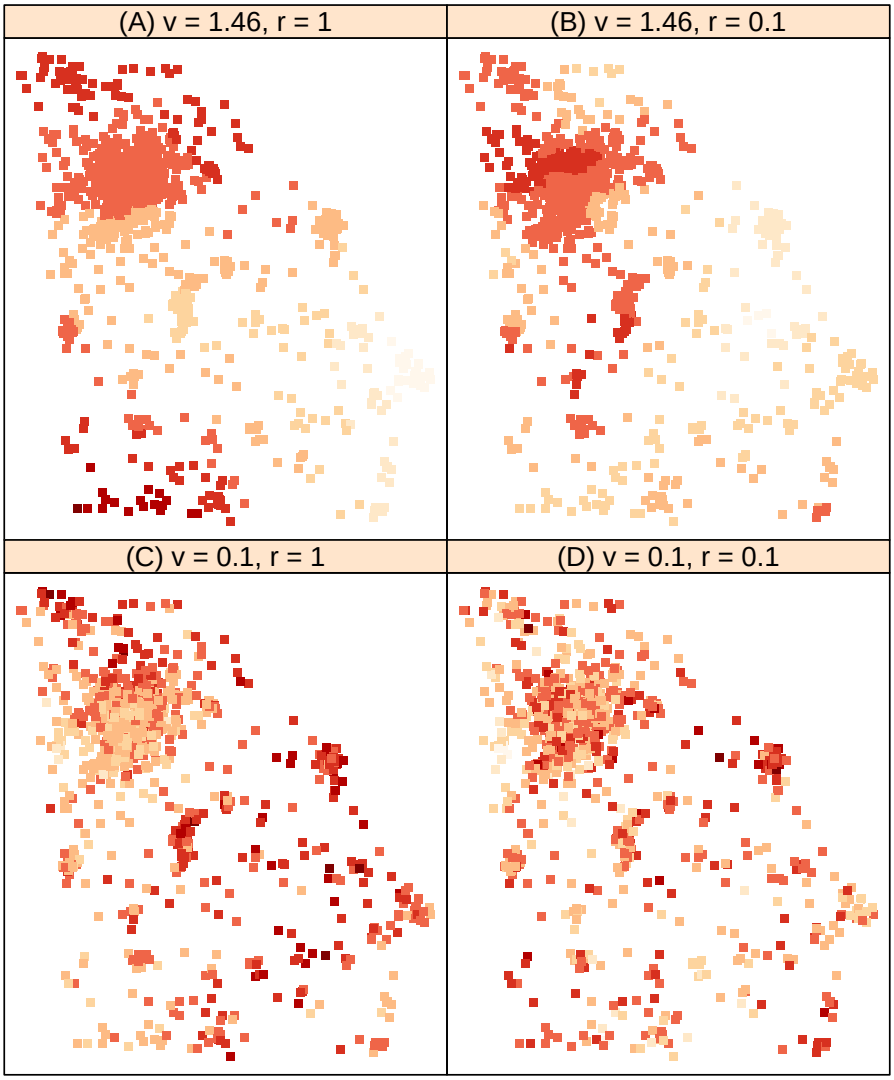


Figure A.2: Example datasets demonstrating different Matern parameters for smoothness (s) and range (r), with darker colors indicating higher values and lighter values indicating lower values.

A.1.2 DAPSm method parameters and full results

Utilizing the DAPSm method requires several choices on the part of the user. First, the weight must be chosen, with a value of 0 meaning that matching is based entirely on the distance between treated and controls, and a value of 1 meaning that matching is based entirely on the propensity score difference. Second, the researcher must choose what caliper type and what caliper value to use based on the weighted DAPS. The DAPSm package in R allows for a caliper to be based on either the DAPS itself or on the propensity score.

In the simulations, we consider the following methods:

- Choosing the optimal weight (described below) for each simulation together with a caliper of 0.2 standard deviations on the propensity score. This is presented in the main results and is the preferred combination.
- Choosing the optimal weight for each simulation together with a caliper of 0.3 standard deviations on the DAPS.
- An approach that always uses a constant weight of 0.3 and a caliper of 0.3 on the DAPS score.
- An approach that always uses a constant weight of 0.3 and a caliper of 0.2 on the PS.

This applies both to the overall DAPSm method as well as in combination with the within-cluster (WC) method, and with each of the specified propensity scores (single-level, random effects, fixed effects). The optimal method proceeds by first trying to find the smallest weight where the matched dataset results in covariates being under a cutoff of 0.10 in absolute standardized difference between treated and control units. If this cannot be met, or if the only weight that satisfies this criteria is 1 (giving no weight to distance and fully weighting the propensity score difference),

the procedure is then to pick the smallest weight that meets a cutoff of 0.15, and then finally 0.2, in absolute standardized difference for all covariates. This second step trying 0.15 and 0.20 as cutoffs allows for a weight of 1 to be chosen. If this procedure fails to find a weight for all 3 cutoff points, a weight of 1 is used. In finding the smallest weight, weights are attempted between 0 to 1 in increments of 0.025. When used in conjunction with the within-cluster methods, only the portion of the dataset that is not matched in that first step is used for the purposes of calculating balance on the covariates and making a decision about the weight.

In general, the 4 methods attempted were broadly similar across the 500 simulations. For conciseness, in the main paper text, only the preferred optimal PS caliper-based results are shown. Tables A.2 and A.3 show bias, relative mean-squared error, and the proportion of treated subjects matched the 3 additional methods using the DAPSm technique.

Table A.1: Average weights chosen for "optimal" DAPS matching approaches across simulation settings.

	Continuous Outcome Setting								Binary Outcome Setting							
Method	1-A	1-B	1-C	1-D	2-A	2-B	2-C	2-D	3-A	3-B	3-C	3-D	4-A	4-B	4-C	4-D
PS Caliper = 0.2																
DAPS-S	0.25	0.23	0.24	0.27	0.24	0.24	0.27	0.27	0.23	0.26	0.26	0.26	0.27	0.28	0.25	0.25
DAPS-RE	0.55	0.54	0.59	0.56	0.63	0.62	0.63	0.62	0.55	0.53	0.57	0.55	0.63	0.62	0.62	0.65
DAPS-FE	0.24	0.23	0.26	0.22	0.25	0.25	0.23	0.24	0.24	0.26	0.26	0.25	0.25	0.25	0.23	0.27
WC+DAPS-S	0.64	0.64	0.65	0.65	0.60	0.63	0.63	0.63	0.69	0.64	0.64	0.64	0.59	0.62	0.62	0.62
WC+DAPS-RE	0.85	0.87	0.89	0.86	0.88	0.88	0.89	0.90	0.88	0.86	0.87	0.87	0.88	0.89	0.89	0.88
WC+DAPS-FE	0.74	0.72	0.71	0.69	0.70	0.71	0.67	0.70	0.70	0.71	0.72	0.69	0.65	0.70	0.69	0.69
DAPS Caliper = 0.3																
DAPS-S	0.42	0.42	0.44	0.43	0.44	0.43	0.45	0.42	0.42	0.44	0.44	0.43	0.43	0.43	0.44	0.43
DAPS-RE	0.15	0.15	0.15	0.15	0.15	0.14	0.15	0.14	0.15	0.15	0.16	0.15	0.14	0.14	0.14	0.14
DAPS-FE	0.37	0.35	0.34	0.38	0.36	0.37	0.34	0.37	0.37	0.39	0.35	0.37	0.38	0.37	0.36	0.33
WC+DAPS-S	0.69	0.66	0.66	0.68	0.67	0.66	0.67	0.66	0.67	0.66	0.67	0.67	0.67	0.67	0.66	0.67
WC+DAPS-RE	0.55	0.52	0.53	0.48	0.48	0.49	0.52	0.52	0.53	0.54	0.55	0.51	0.52	0.50	0.49	0.53
WC+DAPS-FE	0.68	0.69	0.68	0.68	0.67	0.67	0.64	0.67	0.69	0.68	0.66	0.65	0.66	0.64	0.66	0.67

This table summarizes the average weight chosen for the DAPS-related methods across 500 datasets generated for each simulation setting. The method using the propensity score caliper of 0.2 (top) is the method presented in the main text.

Table A.2: Summary of Simulations for Continuous Outcome – Additional DAPSm methods

	Mean ATT Estimate (True = 1)								Relative MSE								Proportion treated subjects matched							
Method	1-A	1-B	1-C	1-D	2-A	2-B	2-C	2-D	1-A	1-B	1-C	1-D	2-A	2-B	2-C	2-D	1-A	1-B	1-C	1-D	2-A	2-B	2-C	2-D
CW-PS-S	2.82	2.78	2.86	2.87	2.88	2.86	2.97	2.98	75.66	74.32	66.86	76.07	79.48	76.01	80.91	79.62	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
CW-PS-RE	1.65	1.70	1.79	1.78	1.78	1.78	1.86	1.83	12.15	14.18	14.24	15.56	16.34	15.61	17.35	15.88	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.94
CW-PS-FE	1.04	1.05	1.12	1.13	1.05	1.08	1.11	1.14	1.56	1.27	1.57	1.71	1.36	1.43	1.56	1.73	0.94	0.94	0.94	0.94	0.95	0.95	0.94	0.94
WC+CW-PS-S	1.93	1.95	2.02	2.03	1.99	2.00	2.09	2.10	21.19	22.12	20.99	24.19	23.22	23.21	25.97	25.91	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC+CW-PS-RE	1.29	1.34	1.42	1.42	1.36	1.41	1.45	1.43	3.49	4.42	4.79	5.23	4.64	5.19	5.33	5.21	0.95	0.95	0.95	0.95	0.96	0.96	0.95	0.95
WC+CW-PS-FE	1.00	1.04	1.10	1.10	1.01	1.07	1.09	1.10	1.13	1.15	1.22	1.30	0.98	1.26	1.20	1.45	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95
CW-DAPS-S	2.77	2.76	2.81	2.85	2.89	2.85	2.95	2.96	72.15	72.41	63.59	74.36	80.02	75.87	79.35	77.75	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.96
CW-DAPS-RE	1.77	1.81	1.88	1.89	1.89	1.88	1.98	1.94	15.56	17.76	17.05	19.30	19.92	19.32	21.92	19.55	0.87	0.88	0.88	0.88	0.87	0.88	0.87	0.87
CW-DAPS-FE	1.10	1.09	1.16	1.18	1.08	1.12	1.19	1.20	2.35	1.98	2.24	2.41	1.92	2.42	2.57	2.68	0.81	0.81	0.81	0.81	0.80	0.80	0.80	0.80
WC+CW-DAPS-S	1.86	1.89	1.96	1.98	1.95	1.98	2.06	2.07	18.45	19.69	18.82	21.99	21.59	22.50	24.47	24.31	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
WC+CW-DAPS-RE	1.32	1.37	1.44	1.45	1.39	1.42	1.48	1.47	3.79	4.67	5.15	5.71	4.79	5.37	5.98	5.76	0.91	0.92	0.91	0.92	0.91	0.91	0.91	0.91
WC+CW-DAPS-FE	1.03	1.07	1.13	1.11	1.03	1.08	1.12	1.13	1.30	1.33	1.47	1.48	1.17	1.45	1.56	1.52	0.87	0.88	0.88	0.87	0.86	0.87	0.87	0.87
OPT-PS-S	2.82	2.77	2.87	2.88	2.88	2.86	2.98	2.98	76.07	73.47	67.06	76.73	79.53	76.13	81.35	79.05	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
OPT-PS-RE	1.61	1.68	1.75	1.78	1.76	1.76	1.82	1.80	11.23	13.83	13.17	15.26	15.65	15.12	16.27	15.15	0.94	0.95	0.95	0.95	0.95	0.95	0.94	0.94
OPT-PS-FE	1.05	1.06	1.13	1.14	1.05	1.08	1.11	1.14	1.43	1.15	1.56	1.57	1.24	1.24	1.44	1.64	0.94	0.94	0.94	0.94	0.95	0.95	0.94	0.94
WC+OPT-PS-S	1.94	1.95	2.03	2.03	2.01	2.01	2.09	2.11	21.42	22.50	21.44	24.32	23.98	23.56	26.02	25.91	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC+OPT-PS-RE	1.28	1.34	1.41	1.41	1.36	1.40	1.44	1.43	3.47	4.41	4.71	5.10	4.70	5.14	5.24	5.28	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
WC+OPT-PS-FE	1.00	1.05	1.11	1.10	1.02	1.08	1.10	1.10	1.12	1.19	1.27	1.26	1.11	1.31	1.27	1.31	0.94	0.95	0.94	0.95	0.95	0.95	0.95	0.95
OPT-DAPS-S	2.78	2.75	2.84	2.85	2.88	2.84	2.97	2.95	73.17	71.70	65.29	74.41	79.50	74.88	80.52	77.04	0.93	0.93	0.94	0.94	0.93	0.93	0.94	0.94
OPT-DAPS-RE	2.07	2.10	2.19	2.19	2.24	2.25	2.34	2.32	28.68	29.59	28.82	32.15	36.18	35.07	37.78	36.78	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.97
OPT-DAPS-FE	1.20	1.22	1.30	1.27	1.18	1.24	1.31	1.32	3.43	3.48	4.01	3.49	3.22	3.72	4.52	4.67	0.86	0.86	0.87	0.86	0.85	0.86	0.86	0.86
WC+OPT-DAPS-S	1.88	1.88	1.98	1.98	1.95	1.97	2.05	2.06	18.98	19.66	19.53	22.45	21.47	22.30	24.22	23.57	0.95	0.96	0.96	0.96	0.95	0.95	0.96	0.96
WC+OPT-DAPS-RE	1.35	1.41	1.49	1.50	1.48	1.49	1.55	1.55	4.93	5.68	6.22	7.05	6.93	7.20	7.89	7.89	0.94	0.95	0.95	0.95	0.96	0.95	0.95	0.95
WC+OPT-DAPS-FE	1.04	1.08	1.15	1.13	1.06	1.10	1.15	1.15	1.30	1.40	1.58	1.63	1.48	1.50	1.73	1.73	0.92	0.92	0.92	0.92	0.92	0.92	0.91	0.92

"CW-PS" represents the DAPS matching technique using a constant weight (CW) of 0.3 for all methods with a propensity score (PS) caliper of 0.2. "CW-DAPS" represents a constant weight of 0.3 with a DAPS caliper of 0.3. "OPT-PS" is presented in the main text and represents the method of choosing an optimal weight based on balance of observed covariates, with a PS caliper of 0.2.
"OPT-DAPS" is an optimal method of choosing weights with a DAPS caliper of 0.3.

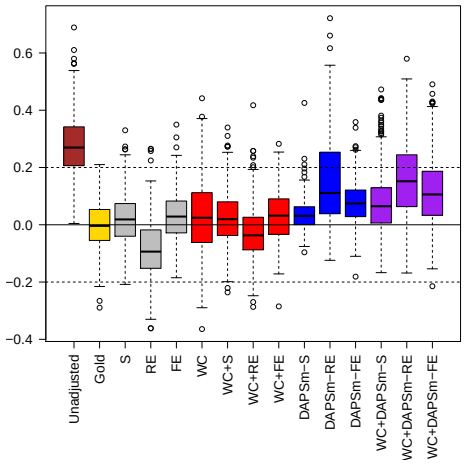
Table A.3: Summary of simulations for binary outcome – additional DAPSm methods

	Mean ATT Estimate (x100) (True ≈ 5)								Relative MSE								Proportion treated subjects matched							
Method	3-A	3-B	3-C	3-D	4-A	4-B	4-C	4-D	3-A	3-B	3-C	3-D	4-A	4-B	4-C	4-D	3-A	3-B	3-C	3-D	4-A	4-B	4-C	4-D
CW-PS-S	11.8	11.7	12.0	11.9	12.4	12.3	12.2	12.4	2.53	2.59	2.59	2.86	2.49	2.80	2.72	3.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
CW-PS-RE	5.5	5.9	6.2	6.4	6.4	6.2	6.6	6.7	1.06	1.01	1.05	1.13	1.03	1.07	1.03	1.19	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95
CW-PS-FE	5.3	5.5	5.9	6.3	6.1	5.7	5.7	6.0	1.06	1.02	0.96	1.08	0.96	0.99	1.00	0.96	0.94	0.94	0.94	0.94	0.94	0.95	0.94	0.94
WC+CW-PS-S	9.0	9.3	9.4	9.3	9.6	9.6	9.5	9.7	1.44	1.74	1.55	1.75	1.55	1.77	1.62	1.89	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC+CW-PS-RE	5.4	5.9	6.2	6.2	6.3	6.3	6.0	6.6	1.00	1.11	1.05	1.15	1.00	1.07	1.01	1.17	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
WC+CW-PS-FE	5.4	5.7	5.9	6.1	6.1	6.1	5.8	6.3	0.98	1.07	0.94	1.07	0.94	1.05	1.05	1.06	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
CW-DAPS-S	11.7	11.8	12.2	12.3	12.5	12.4	12.6	12.8	2.57	2.59	2.64	3.02	2.57	2.82	2.99	3.33	0.96	0.95	0.95	0.96	0.96	0.96	0.96	0.96
CW-DAPS-RE	5.9	6.3	6.9	7.3	6.8	6.9	7.0	7.5	1.04	1.10	1.15	1.28	1.03	1.22	1.28	1.40	0.88	0.87	0.88	0.88	0.87	0.88	0.87	0.88
CW-DAPS-FE	5.2	5.1	5.8	6.3	5.6	5.5	5.8	5.9	1.18	1.22	1.03	1.32	1.11	1.23	1.22	1.19	0.81	0.81	0.81	0.81	0.80	0.80	0.80	0.80
WC+CW-DAPS-S	8.8	9.1	9.4	9.4	9.5	9.7	9.7	9.9	1.48	1.66	1.57	1.79	1.50	1.75	1.70	1.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
WC+CW-DAPS-RE	5.6	6.0	6.6	6.6	6.3	6.6	6.4	6.9	1.04	1.13	1.11	1.16	1.04	1.14	1.08	1.22	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
WC+CW-DAPS-FE	5.2	5.4	5.9	6.0	5.6	5.9	5.7	6.3	1.09	1.11	1.02	1.16	1.02	1.19	1.13	1.20	0.88	0.87	0.88	0.88	0.86	0.87	0.87	0.87
OPT-PS-S	11.9	11.8	12.1	12.0	12.4	12.2	12.3	12.3	2.56	2.60	2.57	2.95	2.49	2.77	2.74	3.03	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
OPT-PS-RE	5.6	6.2	6.4	6.5	6.5	6.5	6.6	6.6	1.08	1.00	1.10	1.21	1.03	1.08	1.08	1.10	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95
OPT-PS-FE	5.3	5.5	5.8	6.3	6.2	5.6	5.8	5.9	1.01	1.04	0.94	1.07	0.93	1.02	1.04	0.95	0.94	0.94	0.94	0.94	0.94	0.95	0.94	0.94
WC+OPT-PS-S	9.1	9.4	9.5	9.5	9.8	9.7	9.6	9.7	1.54	1.74	1.60	1.77	1.61	1.78	1.67	1.87	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC+OPT-PS-RE	5.5	6.1	6.3	6.2	6.7	6.5	6.3	6.4	1.00	1.03	1.09	1.09	1.01	1.13	1.01	1.17	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
WC+OPT-PS-FE	5.3	5.7	5.9	6.0	6.3	6.1	5.8	6.2	1.01	1.05	0.96	1.12	0.91	1.08	1.05	1.18	0.94	0.95	0.95	0.94	0.95	0.95	0.95	0.95
OPT-DAPS-S	11.6	11.7	12.0	12.1	12.3	12.3	12.5	12.7	2.50	2.54	2.53	3.04	2.49	2.86	3.02	3.33	0.94	0.93	0.93	0.93	0.93	0.94	0.93	0.94
OPT-DAPS-RE	9.0	9.1	9.7	9.9	10.0	9.9	10.3	10.7	1.50	1.53	1.66	1.97	1.67	1.82	1.97	2.32	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.97
OPT-DAPS-FE	6.0	5.7	6.5	6.8	6.3	6.1	6.6	6.8	1.18	1.21	1.23	1.17	1.08	1.18	1.29	1.33	0.87	0.86	0.87	0.87	0.86	0.86	0.86	0.87
WC+OPT-DAPS-S	8.7	8.9	9.2	9.2	9.4	9.3	9.4	9.7	1.53	1.62	1.49	1.73	1.46	1.65	1.62	1.93	0.96	0.95	0.95	0.95	0.95	0.96	0.95	0.96
WC+OPT-DAPS-RE	6.2	6.7	7.1	7.2	7.2	7.4	7.3	7.5	1.13	1.11	1.19	1.27	1.11	1.19	1.19	1.37	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
WC+OPT-DAPS-FE	5.3	5.8	6.1	6.2	6.4	6.0	5.9	6.4	0.99	1.12	1.05	1.16	1.00	1.12	1.07	1.26	0.92	0.92	0.92	0.92	0.92	0.92	0.91	0.92

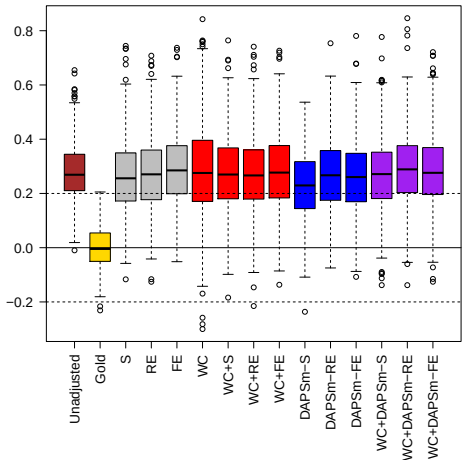
"CW-PS" represents the DAPS matching technique using a constant weight (CW) of 0.3 for all methods with a propensity score (PS) caliper of 0.2. "CW-DAPS" represents a constant weight of 0.3 with a DAPS caliper of 0.3. "OPT-PS" is presented in the main text and represents the method of choosing an optimal weight based on balance of observed covariates, with a PS caliper of 0.2.
"OPT-DAPS" is an optimal method of choosing weights with a DAPS caliper of 0.3.

Table A.4: Additional simulation results for $n = 1000$, $J = 40$ facilities

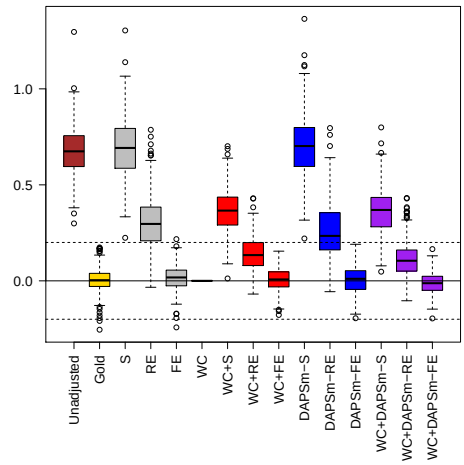
Continuous	Mean ATT Estimate (True = 1)								Relative MSE								Proportion treated subjects matched							
Method	1-A	1-B	1-C	1-D	2-A	2-B	2-C	2-D	1-A	1-B	1-C	1-D	2-A	2-B	2-C	2-D	1-A	1-B	1-C	1-D	2-A	2-B	2-C	2-D
Gold	1.02	0.99	1.01	0.99	1.02	0.98	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.98	0.98	0.97	0.98	0.98
S	2.64	2.73	2.73	2.77	2.83	2.89	2.97	2.95	64.22	67.09	62.52	72.82	80.52	78.94	82.40	81.20	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
RE	1.49	1.56	1.62	1.66	1.49	1.58	1.63	1.64	7.32	9.02	9.47	11.96	7.22	9.04	9.71	10.40	0.97	0.97	0.97	0.97	0.96	0.96	0.97	0.97
FE	1.03	1.11	1.13	1.14	1.03	1.09	1.15	1.18	1.20	1.53	1.31	1.75	1.33	1.52	1.80	2.05	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC	1.02	1.09	1.11	1.15	1.01	1.07	1.13	1.14	0.89	1.04	1.15	1.52	0.98	0.99	1.24	1.28	0.79	0.79	0.79	0.79	0.76	0.76	0.76	0.76
WC+S	1.53	1.60	1.60	1.69	1.64	1.72	1.76	1.77	8.47	9.40	8.72	12.37	11.73	13.16	13.61	14.26	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC+RE	1.11	1.17	1.21	1.26	1.11	1.18	1.22	1.23	1.41	1.77	1.87	2.63	1.47	1.97	1.95	2.11	0.97	0.97	0.97	0.97	0.97	0.96	0.97	0.97
WC+FE	0.99	1.06	1.09	1.12	0.98	1.05	1.10	1.11	1.01	1.04	0.99	1.33	1.16	1.05	1.07	1.20	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
DAPSm-S	2.64	2.67	2.70	2.74	2.85	2.86	2.95	2.93	64.24	63.17	59.61	69.82	82.29	76.58	79.88	80.16	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
DAPSm-RE	1.41	1.40	1.49	1.52	1.48	1.50	1.53	1.55	5.26	5.37	6.43	8.12	7.52	7.02	7.31	7.89	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
DAPSm-FE	1.05	1.06	1.11	1.14	1.07	1.07	1.11	1.15	1.27	1.05	1.32	1.62	1.38	1.21	1.27	1.56	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
WC+DAPSm-S	1.59	1.65	1.65	1.70	1.68	1.71	1.76	1.76	9.79	10.65	9.80	12.88	12.78	12.80	13.39	13.60	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC+DAPSm-RE	1.10	1.15	1.19	1.22	1.14	1.17	1.20	1.21	1.25	1.54	1.59	2.21	1.64	1.75	1.78	1.95	0.97	0.97	0.97	0.97	0.97	0.96	0.96	0.96
WC+DAPSm-FE	1.00	1.05	1.08	1.11	1.02	1.06	1.09	1.09	0.97	0.99	1.01	1.28	1.04	1.01	1.03	1.05	0.96	0.96	0.97	0.96	0.97	0.97	0.97	0.97
Binary	Mean ATT Estimate (x100) (True ≈ 5)								Relative MSE								Proportion treated subjects matched							
Method	3-A	3-B	3-C	3-D	4-A	4-B	4-C	4-D	3-A	3-B	3-C	3-D	4-A	4-B	4-C	4-D	3-A	3-B	3-C	3-D	4-A	4-B	4-C	4-D
Gold	4.4	5.0	5.2	5.1	4.8	4.7	4.9	4.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.98	0.97	0.98	0.98	0.98
S	10.5	11.3	12.2	12.1	11.3	12.2	12.4	12.3	2.17	2.43	2.66	2.60	2.36	3.24	3.08	2.72	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
RE	5.5	6.5	6.9	7.0	5.5	6.3	6.6	6.9	1.13	1.19	1.11	1.22	1.07	1.12	1.21	1.12	0.97	0.97	0.97	0.97	0.96	0.96	0.97	0.97
FE	5.0	6.0	6.3	6.6	5.5	5.8	6.3	6.4	0.93	0.97	1.06	1.15	0.97	1.20	1.09	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC	4.8	5.7	6.1	6.2	5.0	5.3	5.9	5.9	1.30	1.26	1.09	1.17	1.16	1.32	1.27	1.16	0.79	0.78	0.79	0.78	0.76	0.76	0.76	0.76
WC+S	7.0	8.0	8.5	8.7	7.9	8.3	8.6	8.8	1.23	1.33	1.38	1.38	1.22	1.58	1.50	1.38	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC+RE	5.1	6.1	6.5	6.7	5.3	5.8	6.1	6.2	1.12	1.08	1.02	1.05	0.95	1.12	1.04	1.00	0.97	0.97	0.97	0.97	0.96	0.96	0.97	0.97
WC+FE	5.2	6.2	6.5	6.7	5.7	6.0	6.4	6.6	1.08	1.08	1.07	1.07	0.91	1.12	1.06	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
DAPSm-S	10.9	11.4	12.0	12.0	11.7	11.8	12.4	12.2	2.32	2.41	2.65	2.65	2.49	2.97	2.99	2.73	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
DAPSm-RE	5.1	5.8	6.4	6.6	6.0	5.9	6.4	6.5	1.08	1.04	1.06	1.07	1.04	1.10	1.23	1.06	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
DAPSm-FE	5.0	5.6	6.0	6.4	5.4	5.7	6.2	6.1	1.09	1.02	1.01	1.08	1.07	1.08	1.16	0.97	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
WC+DAPSm-S	7.4	8.3	8.7	8.8	8.2	8.3	8.6	8.6	1.32	1.43	1.45	1.40	1.27	1.66	1.48	1.39	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
WC+DAPSm-RE	5.0	6.0	6.5	6.6	5.9	5.9	6.4	6.2	1.10	1.15	1.08	0.96	0.97	1.11	1.11	1.00	0.96	0.97	0.97	0.97	0.96	0.96	0.97	0.96
WC+DAPSm-FE	5.0	6.1	6.3	6.5	6.0	5.8	6.2	6.3	1.11	1.07	1.04	1.01	0.94	1.08	1.05	0.99	0.96	0.97	0.97	0.97	0.97	0.96	0.97	0.97



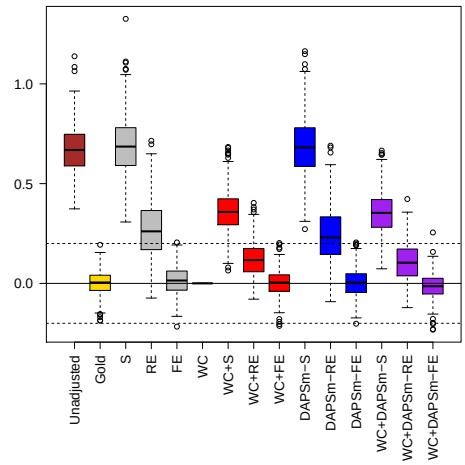
(a) U: S2-A (Smoothness 1.46, Range 1)



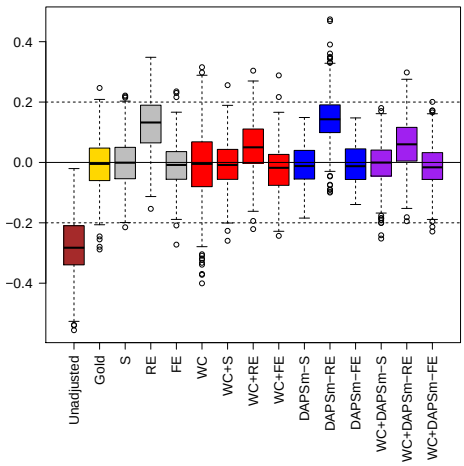
(b) U: S2-D (Smoothness 0.1, Range 0.1)



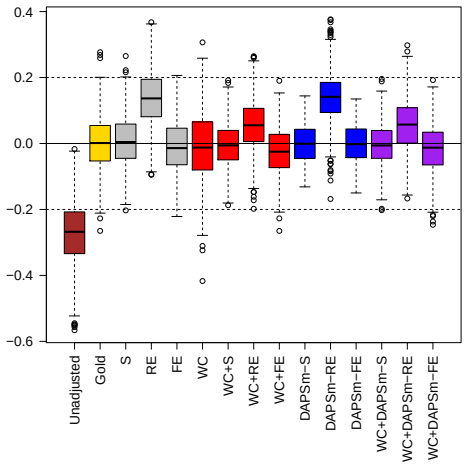
(c) V: S2-A



(d) V: S2-D

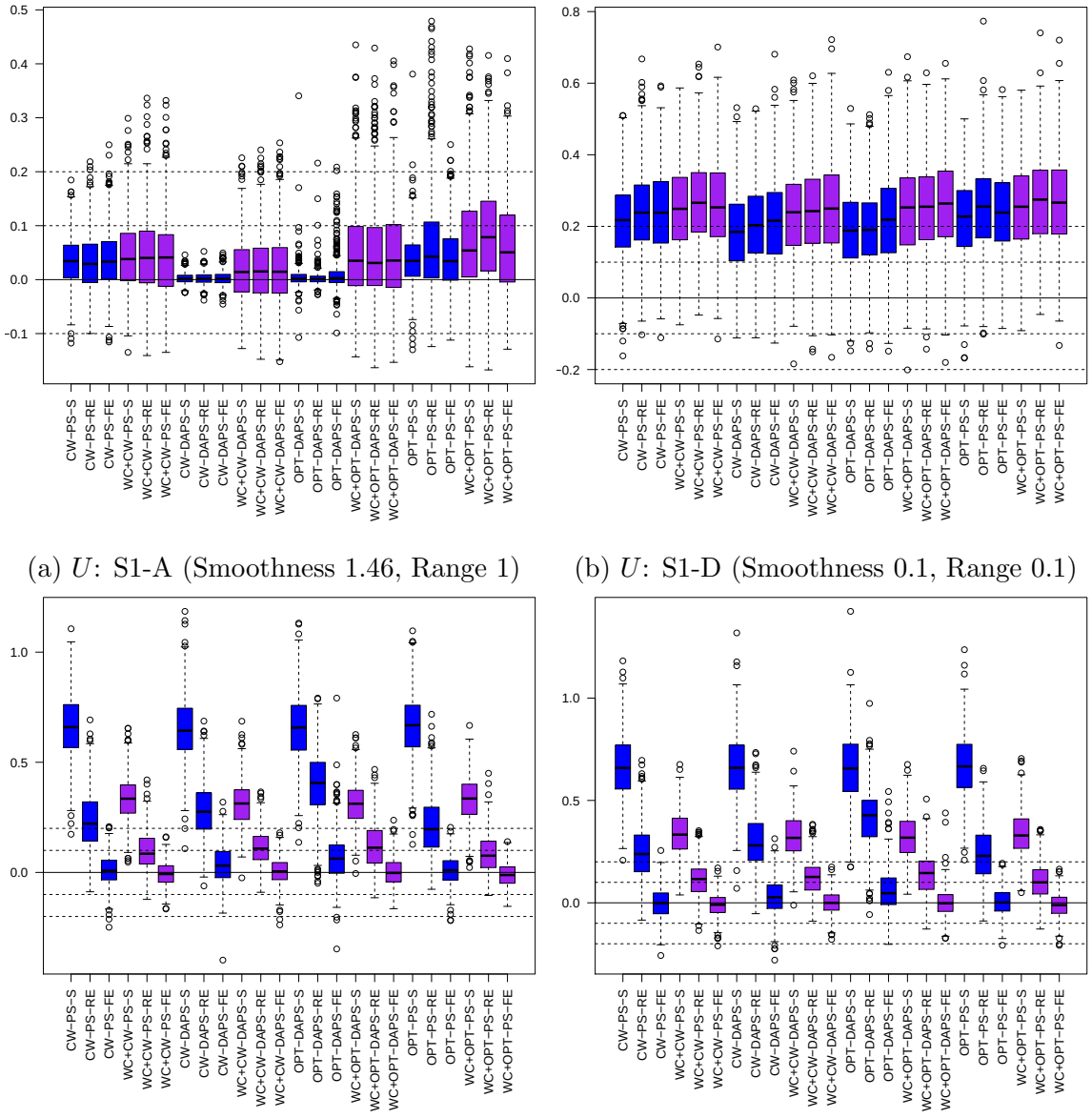


(e) X_4: S2-A



(f) X_4: S2-D

Figure A.3: Standardized difference for U, V, and X_4 in random facility assignment setting (S2-A and S2-D)



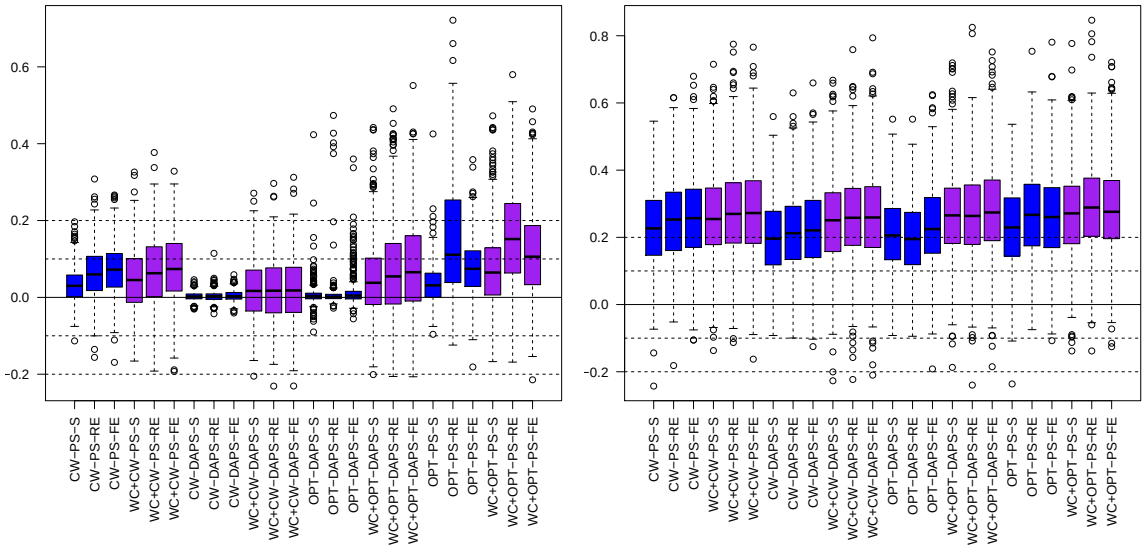
(a) U: S1-A (Smoothness 1.46, Range 1)

(b) U: S1-D (Smoothness 0.1, Range 0.1)

(c) V: S1-A

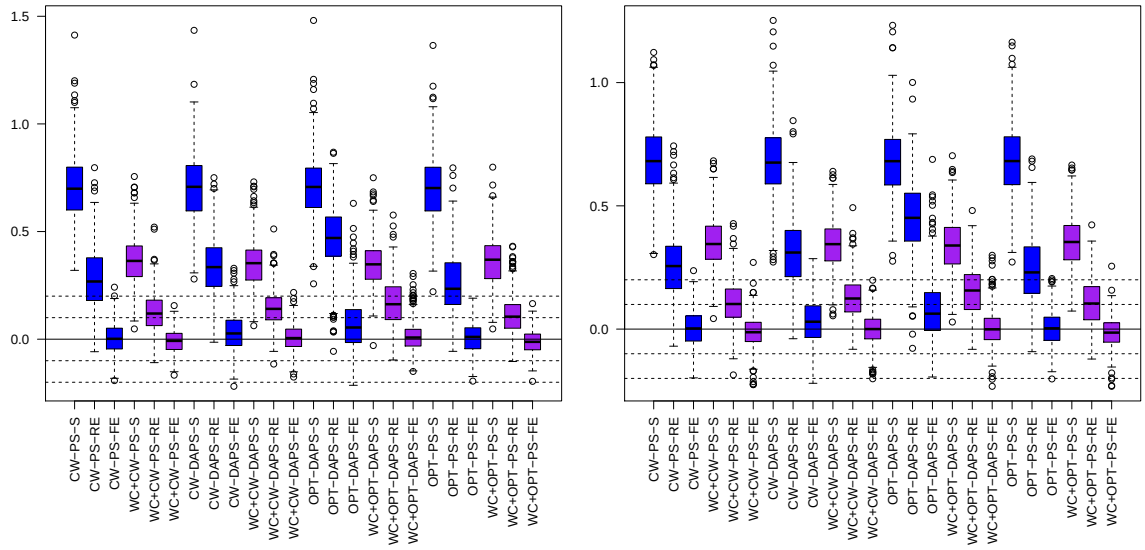
(d) V: S1-D

Figure A.4: Standardized difference for U and V in distance-based setting (S1-A and S1-D) for various DAPS-based methods. "OPT" refers to optimal method of choosing weights described in text. "CW" refers to the constant-weight method (0.3). "DAPS" refers to the DAPS-based caliper of 0.3. "PS" refers to the PS-based caliper of 0.2.



(a) U: S2-A (Smoothness 1.46, Range 1)

(b) U: S2-D (Smoothness 0.1, Range 0.1)



(c) V: S2-A

(d) V: S2-D

Figure A.5: Standardized difference for U and V in random facility assignment setting (S2-A and S2-D) for various DAPS-based methods. "OPT" refers to optimal method of choosing weights described in text. "CW" refers to the constant-weight method (0.3). "DAPS" refers to the DAPS-based caliper of 0.3. "PS" refers to the PS-based caliper of 0.2.

A.2 Data Application

A.2.1 Sample construction

Sample exclusions from the 2017 USRDS (3,081,768 patients):

- 47,463 not included in the ADR
- 2,429,508 patients identified by USRDS as not starting dialysis between January 1, 2012 and December 31, 2016.
- 16,846 with a missing facility ID (from *SAF.RXHIST*).
- 536,751 patients that did not begin dialysis at a facility in GA, NC, and SC.
- 4,167 patients with initial treatment modalities including uncertain dialysis or other (from *SAF.RXHIST60*).
- 3,553 patients between 18 and 80 years old (inclusive).
- 739 patients removed with no matching medical evidence form.
- 131 patients whose facility did not correspond to any facility in SAF.FACILITY in the incidence year/facility year
- 60 patients with no supervising physician signature date.
- 4,526 patients with a physician signature date more than 62 days after first ESRD date.
- 33 patients with a signature date prior to first ESRD service date.
- 1,555 patients excluded because not in categories of "INFORMED" or "NOT ASSESSED"
- 21,603 patients excluded because facility is not in GA (i.e. is in NC or SC)

- 587 pre-emptive waitlists excluded
- 1,337 pre-emptive referrals excluded.
- 350 who do not report living in GA at time of first ESRD service date based on reported ZIP code – matched to SAS's ZIPCODE database to determine state of residence.
- 7,579 excluded from 2014-2016 incidence; focus analysis only on 2012 and 2013 incidence years.
- 74 removed who were in transplant facilities, VA facilities, or had missing values for BMI/GFR-EPI/Co-morbidities.

A.2.2 Covariates

The variables included in propensity score estimation models are: incidence year, incidence age, race, Hispanic ethnicity, log-BMI, log GFR-EPI score, height, log weight, dialysis type, access type, primary cause of ESRD (broad grouping), pre-ESRD nephrology/EPO/dietary care, insurance variables, current employment status, hemoglobin, albumin. These are taken from the 2728 form.

In addition, we include many binary co-morbidities from the 2728 form – diabetes (insulin), diabetes (oral), diabetes (retinopathy), history of hypertension, Atherosclerotic heart disease (ASHD), congestive heart failure (CHF), other cardiac disease, peripheral vascular disease, amputation, cerebrovascular disease (CVA, TIA), inability to ambulate, inability to transfer, need assistance with daily activities, institutionalized, alcohol dependence, drug dependence, tobacco use, COPD, malignant neoplasm (cancer), and toxic nephropathy.

Facility variables include number of patients at start and end of incidence year, profit/non-profit status, hospital or free-standing, and various quantities related to

the number of FTE social workers and registered nurses (see notes below). Other covariates include county-level household poverty rate, county-level percent non-Hispanic white, and the total population of the county. County-level variables are ACS 5-year 2014 estimates. County-level variables are attached based on geocoding of ZIP codes rather than the variable recorded in the USRDS – these differ for 169 patients out of 4,906. In various places, I re-code categorical variables to reduce sparsity in certain categories and to simplify the propensity score estimation. For some continuous variables, I use a log-transformation or otherwise convert to a categorical if there are large numbers of missing values.

- Race is re-coded as White, Black, or Other.
- Hispanic (based on 2728) is a binary variable.
- Dialysis type is re-coded as HEMO vs. CAPD/CCPD/Other
- Access Type is recoded such that N/A and Other are grouped together; the other categories are AVF, Graft, and Cath
- Pre-ESRD nephrology, pre-ESRD EPO, and pre-ESRD dietary are coded as 3-category variables: Yes, No, and Unknown/Missing.
- Employment status (current) is recoded as (Unemployed, Med LOA, Other), (Employed FT/PT, Student, or Homemaker) and (Ret-age/Ret-dis).
- Hemoglobin is put into 3 categories: < 10, >= 10, or NA (missing).
- Serum Albumin is coded into 3 categories: Low < $3.5g/dL$, >= $3.5g/dL$, or NA/Missing.
- Log transformations are made for: BMI, GFR-EPI, weight and county total population.

- Facility variables include the number of patients being treated at the beginning and end of the incidence year (survey period).
- Facility variable for non-profit status (3 categories): For-profit, non-profit, or unknown (based on SAF.FACILITY).
- Facility variable for free-standing or hospital-based – with the exception of the fixed effects models which do not include this variable due to convergence issues.
- Facility variables are also included for number of FTE social workers and FTE registered nurses; I also use a ratio of FTE social workers to number of patients at the end of the year, and a ratio of FTE registered nurses to number of patients at the end of the year. A handful of patients are in facilities with 0 patients at the end of a year – in these cases, I use the average of patients at the start of the year and end of the year.

A.2.3 Balance

The standard deviation used to calculate the absolute standardized differences are based on the standard deviation of the treated group in the full unmatched sample. This is done based on recommendations from Stuart (2010). Figure 1.2 demonstrate balance for various methods as compared to the unadjusted method using love plots. Ideal balance is achieved when absolute standardized differences are below 0.1. We additionally provide select tables comparing balance on some important covariates before and after matching for pre-ESRD nephrology care, hemodialysis, and access type.

	N		Hemodialysis		
Method	Control	Treated	Controls	Treated	Std. difference
Unadjusted	4428	478	90.1	97.3	0.44
S	472	472	97.5	97.2	-0.01
RE	264	264	93.6	95.5	0.12
FE	295	295	94.9	95.6	0.04
WC	206	206	92.2	95.1	0.18
WC+S	471	471	95.5	97.2	0.10
WC+RE	283	283	94.0	95.4	0.09
WC+FE	310	310	93.5	95.8	0.14
DAPSm-S	460	460	95.7	97.2	0.09
DAPSm-RE	257	257	93.4	95.7	0.14
DAPSm-FE	256	256	95.3	96.1	0.05
WC+DAPSm-S	463	463	95.5	97.2	0.11
WC+DAPSm-RE	278	278	92.8	95.3	0.15
WC+DAPSm-FE	284	284	93.7	95.8	0.13

Table A.5: Hemodialysis (vs. CAPD/CCPD/Other) in unmatched and matched samples for control and treated subjects, as well as the standardized difference.

	Pre-ESRD Nephrology Care								
	Yes			No			Unknown		
Method	C	T	SD	C	T	SD	C	T	SD
Unadjusted	59.8	43.9	-0.32	30.9	32.4	0.03	9.2	23.6	0.34
S	42.8	44.5	0.03	30.9	32.4	0.03	26.3	23.1	-0.07
RE	57.2	56.1	-0.02	29.9	31.1	0.02	12.9	12.9	0.00
FE	54.6	53.2	-0.03	30.5	32.2	0.04	14.9	14.6	-0.01
WC	57.3	56.3	-0.02	28.6	31.1	0.05	14.1	12.6	-0.03
WC+S	45.9	44.6	-0.03	30.4	32.1	0.04	23.8	23.4	-0.01
WC+RE	58.7	55.5	-0.06	26.9	31.1	0.09	14.5	13.4	-0.02
WC+FE	57.1	51.9	-0.10	29.4	34.5	0.11	13.5	13.5	0.00
DAPSm-S	48.5	45.2	-0.07	28.9	32.4	0.07	22.6	22.4	-0.01
DAPSm-RE	52.9	54.5	0.03	32.7	32.7	0.00	14.4	12.8	-0.04
DAPSm-FE	54.7	53.9	-0.02	31.2	33.2	0.04	14.1	12.9	-0.03
WC+DAPSm-S	46.4	45.1	-0.03	29.4	32.6	0.07	24.2	22.2	-0.05
WC+DAPSm-RE	56.8	54.7	-0.04	29.9	32.4	0.05	13.3	12.9	-0.01
WC+DAPSm-FE	58.1	53.9	-0.09	28.9	32.4	0.08	13.0	13.7	0.02

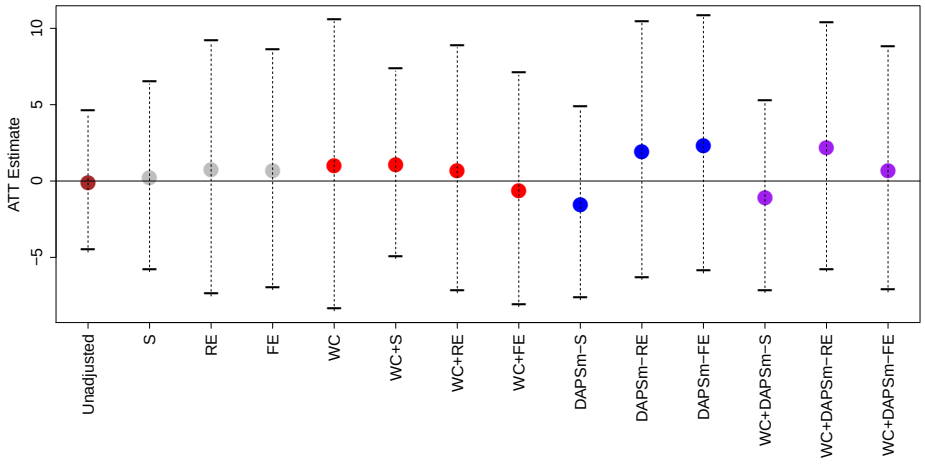
Table A.6: Pre-ESRD nephrology care in unmatched and matched samples for control and treated subjects, as well as the standardized difference.

	AVF			Graft			Catheter			Other/NA		
Method	C	T	SD	C	T	SD	C	T	SD	C	T	SD
Unadjusted	12.2	11.3	-0.03	2.4	4.0	0.08	75.5	82.2	0.18	9.9	2.5	-0.47
S	10.6	11.4	0.03	3.2	4.0	0.04	83.7	82.0	-0.04	2.5	2.5	0.00
RE	15.2	12.5	-0.08	2.3	2.7	0.02	76.1	80.3	0.11	6.4	4.5	-0.12
FE	14.9	12.5	-0.07	2.7	3.1	0.02	77.3	80.3	0.08	5.1	4.1	-0.07
WC	12.1	12.1	0.00	1.0	2.4	0.07	79.6	81.1	0.04	7.3	4.4	-0.19
WC+S	11.0	11.5	0.01	2.5	4.0	0.08	82.2	82.0	-0.01	4.2	2.5	-0.11
WC+RE	13.4	12.4	-0.03	2.1	3.2	0.05	78.8	80.2	0.04	5.7	4.2	-0.09
WC+FE	14.5	12.3	-0.07	1.9	3.2	0.07	77.4	80.6	0.08	6.1	3.9	-0.14
DAPSm-S	11.7	11.5	-0.01	3.5	3.9	0.02	80.4	82.0	0.04	4.3	2.6	-0.11
DAPSm-RE	11.7	11.7	0.00	1.9	3.1	0.06	80.5	80.9	0.01	5.8	4.3	-0.10
DAPSm-FE	11.7	12.1	0.01	1.6	3.5	0.10	82.8	80.5	-0.06	3.9	3.9	0.00
WC+DAPSm-S	12.1	11.4	-0.02	2.8	3.9	0.06	80.8	82.1	0.03	4.3	2.6	-0.11
WC+DAPSm-RE	11.9	11.9	0.00	1.8	3.2	0.07	79.5	80.6	0.03	6.8	4.3	-0.16
WC+DAPSm-FE	13.0	12.7	-0.01	1.8	3.5	0.09	79.2	79.9	0.02	6.0	3.9	-0.14

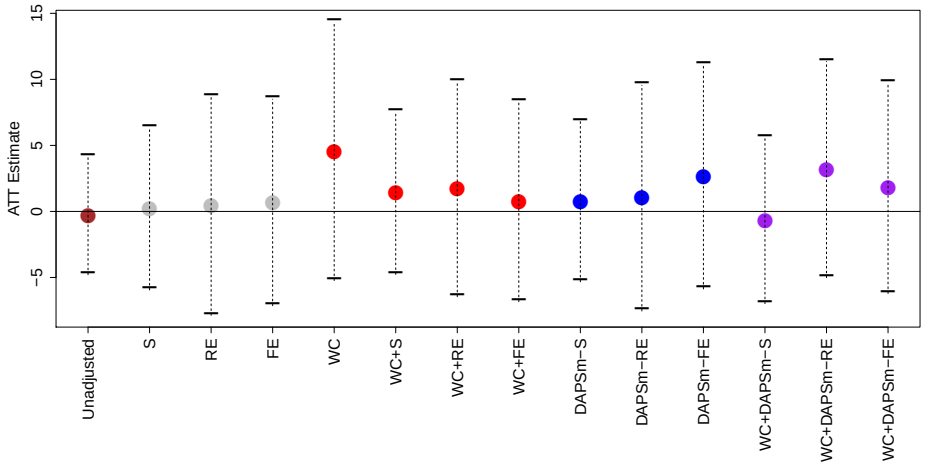
Table A.7: Proportions for various access types on first outpatient dialysis among controls (C), treated (T), and standardized differences (SD) between treated and controls.

A.2.4 Adjusting after matching

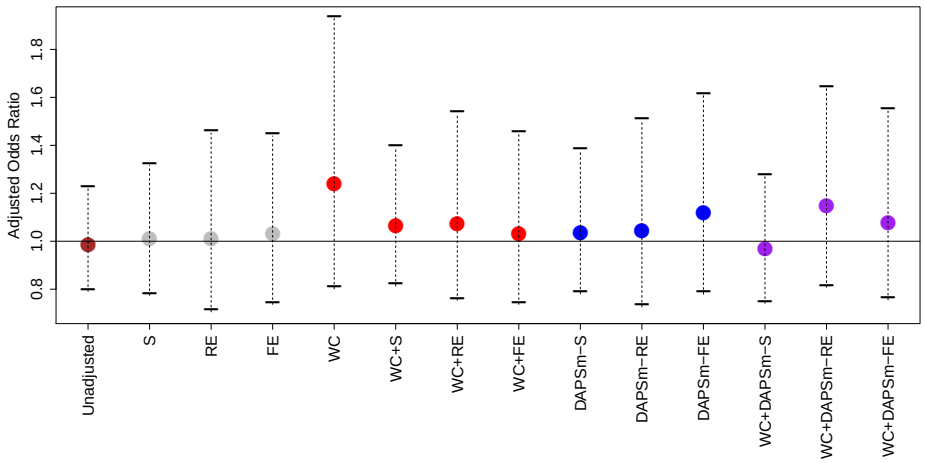
The recommendation by Ho et al. (2007) is to use matching as a pre-processing procedure. After matching, one may utilize parametric methods that will be less sensitive to the particular functional form. We use linear and logistic regression models after matching to adjust for any lingering imbalance in the dataset. Figure A.6 compares 1-year referral estimates after matching with no adjustment, with adjustment in a linear probability model, and with adjustment in a logistic regression model. Adjustment is made for any covariates with an absolute standardized difference greater than 0.1 after matching.



(a) Outcome: 1-year referral



(b) Estimate from linear probability model adjustment



(c) Adjusted odds ratios from logistic regression model adjustment

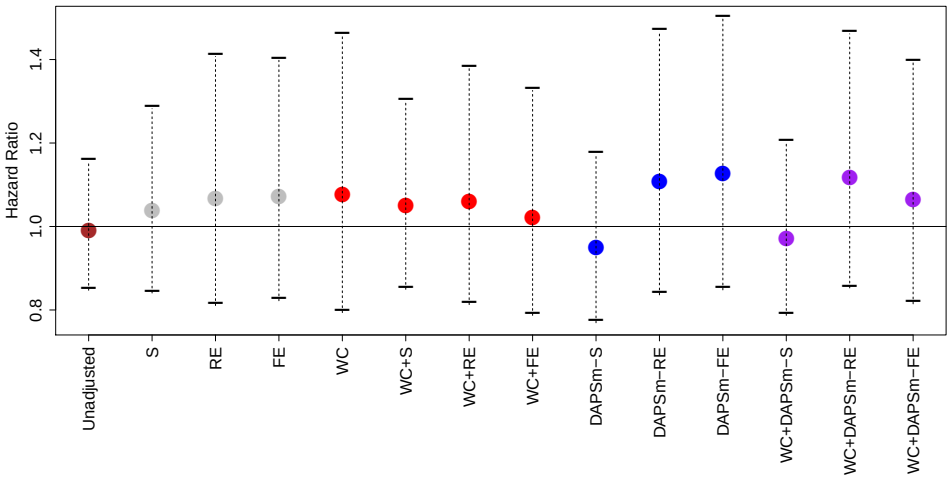
Figure A.6: Estimates of ATT for 1-year referral after matching with no adjustment, with linear model adjustment, and with logistic regression adjustment

A.2.5 Hazard ratio estimate

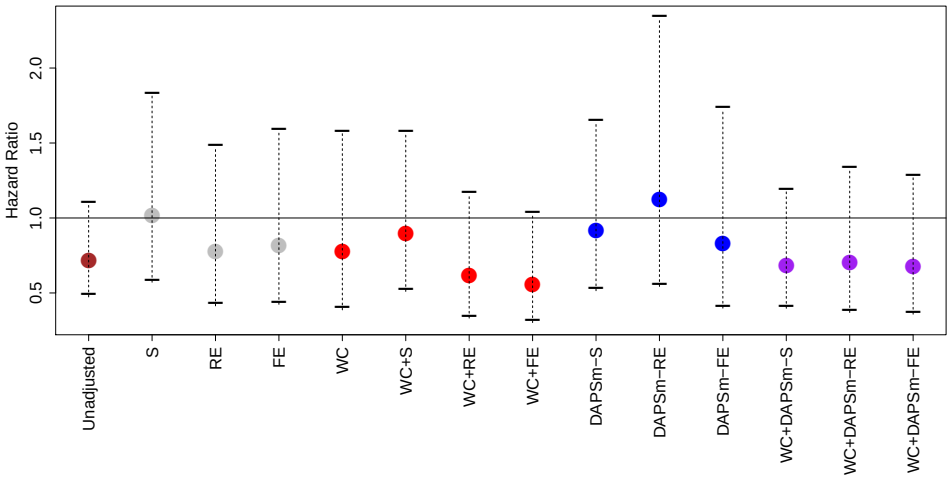
In addition to analyzing 1-year binary outcomes, we also estimate the hazard ratio using the Cox proportional hazards model to account for competing risks. We estimate cause-specific hazard ratios for referral and waitlisting, as well as hazard ratios for death within the first year. For referral and waitlisting, all observations are right-censored at 1-year if no event occurs, and persons are treated as censored if death (competing risk) occurs at the time that death occurs, following Austin et al. (2016) for constructing cause-specific hazard estimates.

	Cause-Specific Hazard Ratio Estimate (95% CI)		
Method	Referral	Waitlist	Death
Unadjusted	0.99 (0.85, 1.16)	0.72 (0.47, 1.09)	0.88 (0.67, 1.15)
S	1.04 (0.84, 1.28)	1.02 (0.57, 1.82)	1.18 (0.80, 1.73)
RE	1.07 (0.81, 1.41)	0.78 (0.41, 1.47)	0.98 (0.58, 1.64)
FE	1.07 (0.82, 1.40)	0.82 (0.42, 1.58)	1.18 (0.73, 1.89)
WC	1.08 (0.79, 1.46)	0.78 (0.39, 1.56)	1.11 (0.60, 2.07)
WC+S	1.05 (0.85, 1.30)	0.89 (0.51, 1.57)	1.11 (0.76, 1.62)
WC+RE	1.06 (0.81, 1.38)	0.62 (0.33, 1.15)	1.24 (0.75, 2.06)
WC+FE	1.02 (0.79, 1.33)	0.56 (0.30, 1.02)	1.42 (0.88, 2.31)
DAPSm-S	0.95 (0.77, 1.17)	0.92 (0.51, 1.64)	0.92 (0.64, 1.33)
DAPSm-RE	1.11 (0.84, 1.47)	1.13 (0.54, 2.33)	1.60 (0.91, 2.81)
DAPSm-FE	1.13 (0.85, 1.50)	0.83 (0.40, 1.72)	1.18 (0.70, 1.99)
WC+DAPSm-S	0.97 (0.79, 1.20)	0.68 (0.40, 1.18)	0.99 (0.68, 1.43)
WC+DAPSm-RE	1.12 (0.85, 1.46)	0.70 (0.37, 1.33)	1.28 (0.77, 2.12)
WC+DAPSm-FE	1.07 (0.81, 1.39)	0.67 (0.36, 1.27)	1.33 (0.80, 2.22)

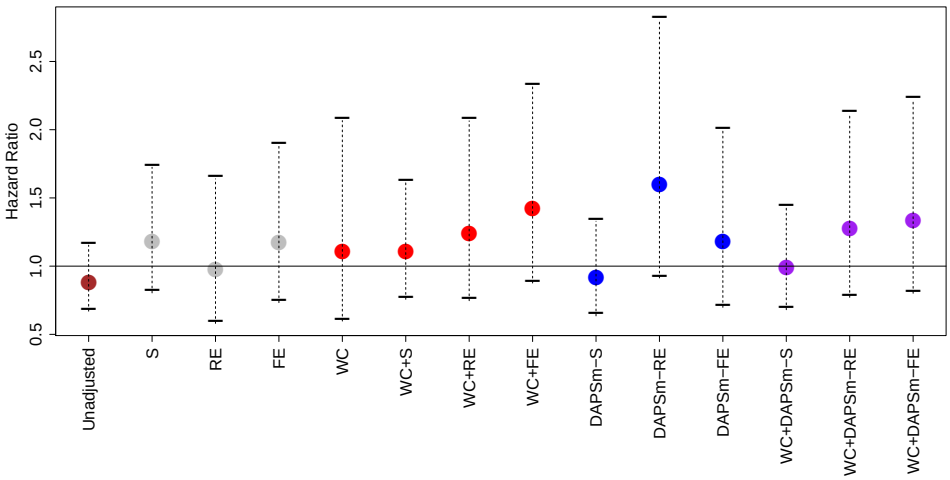
Table A.8: Estimates of cause-specific hazard ratios for treatment (not assessed) from Cox proportional hazards models for 1-year of follow-up. For time-to-referral and time-to-waitlist, death (competing risk) is coded as a censoring event. See Austin et al. (2016).



(a) Referral



(b) Waitlisting



(c) Death

Figure A.7: Estimates of hazard ratio (1-year follow-up) for being not assessed vs. informed

A.2.6 DAPSm tuning parameters and sensitivity results

For the DAPSm methods presented in the main text, we utilized a propensity score caliper of 0.2. By matching on the DAPS score within a propensity caliper, we are ensured that the estimated propensity score difference between matched treated and control patients will not be too large, regardless of the DAPS weight chosen. The alternative approach, which we consider in the Appendix, is to use a caliper on the DAPS score itself. The other choice that may impact results is the weight chosen to estimate the DAPS score.

In the main text, we considered weights between 0 and 1 with increments of 0.05, and we used the following criteria to determine a weight. For the DAPSm-S, DAPSm-RE, and DAPSm-FE methods, we pick the smallest weight that balances all observed covariates under a cutoff of 0.15 absolute standardized difference. This resulted in a weight of 0.05 for all three methods. For the methods that use DAPSm in a second stage following the within-cluster matching stage, we pick a weight of 0.5 as a natural balance between distance and the propensity score differences in constructing the DAPS score.

Given the potential importance of parameter choices, we present full results on sensitivity of outcome estimates to the weight chosen in the DAPSm methods, as well as also using the DAPS caliper in lieu of the PS caliper. In particular these results show a much greater variability in effect estimates when using a DAPS caliper instead of a PS caliper. Smaller weights (more heavily weighting distance between treated and control patients in the DAPS calculation) lead to reduced estimates of the impact of not being assessed, and in some cases negative (more in line with intuition). Nonetheless, substantial uncertainty remains in the estimates regardless of the tuning parameters, along with a substantial portion of the treated units that are unmatched when taking into account the facility in the form of a fixed or random effect.

Figure A.8 shows the sensitivity of results to weight choice in the single-stage

DAPSm-S, DAPSm-RE, and DAPSm-FE methods with a propensity score caliper of 0.2. Figure A.9 shows the same results but using a caliper of 0.3 on the distance-adjusted propensity score for different weights. Figure A.10 and A.11 similarly show the same results when the DAPSm method is taken for the second stage after a first stage of within-cluster matching.

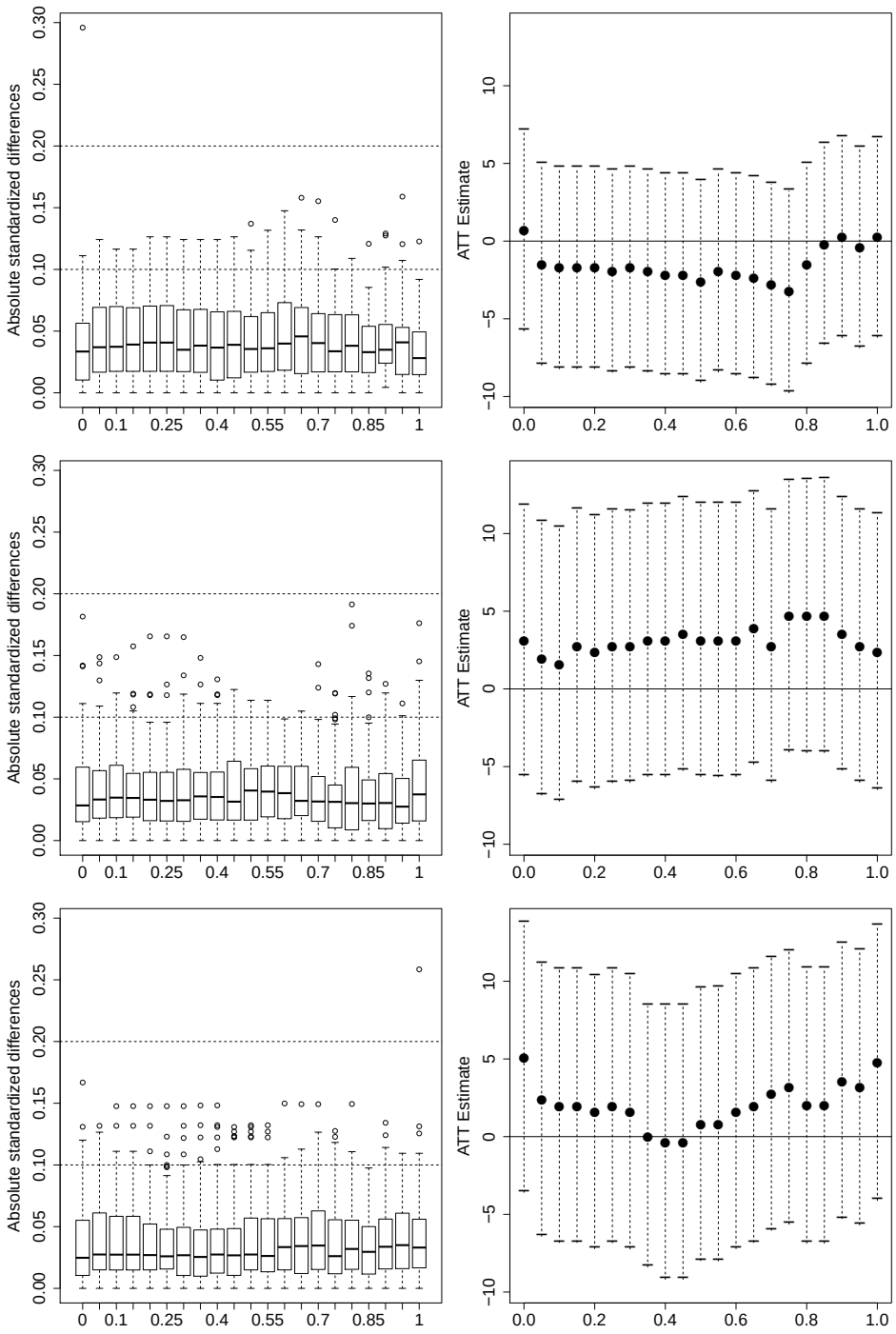


Figure A.8: The x-axis in each figure refers to the weight used in the DAPS score. The first column shows the balance across covariates for each weight. The second column shows outcome estimates for each weight. The rows correspond to the DAPSm-S, DAPSm-RE, and DAPSm-FE methods, respectively, using a caliper of 0.2 on the propensity score.

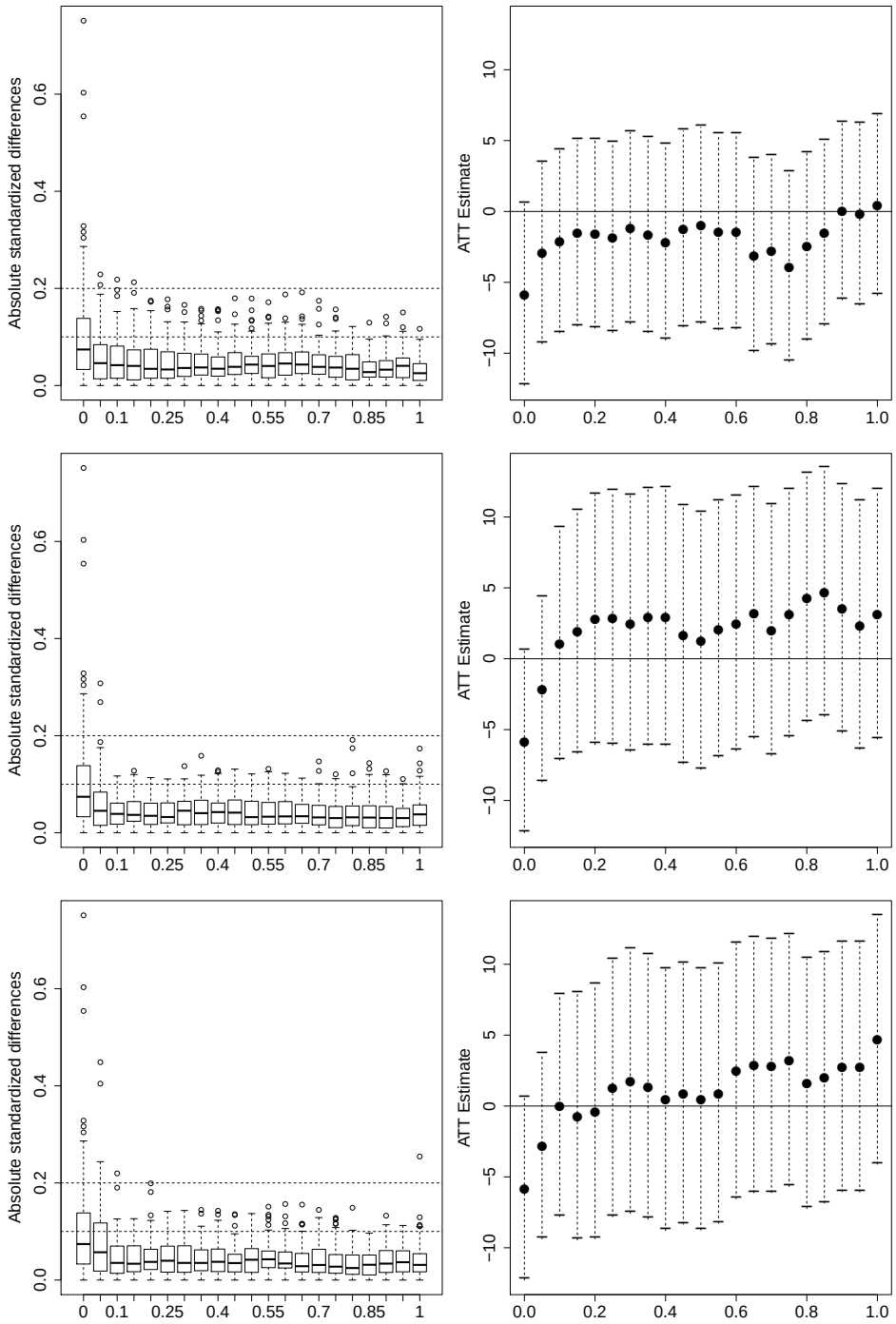


Figure A.9: The x-axis in each figure refers to the weight used in the DAPS score. The first column shows the balance across covariates for each weight. The second column shows outcome estimates for each weight. The rows correspond to the DAPSm-S, DAPSm-RE, and DAPSm-FE methods, respectively, using a caliper of 0.3 on the distance-adjusted propensity score rather than the propensity score as in figure A.8.

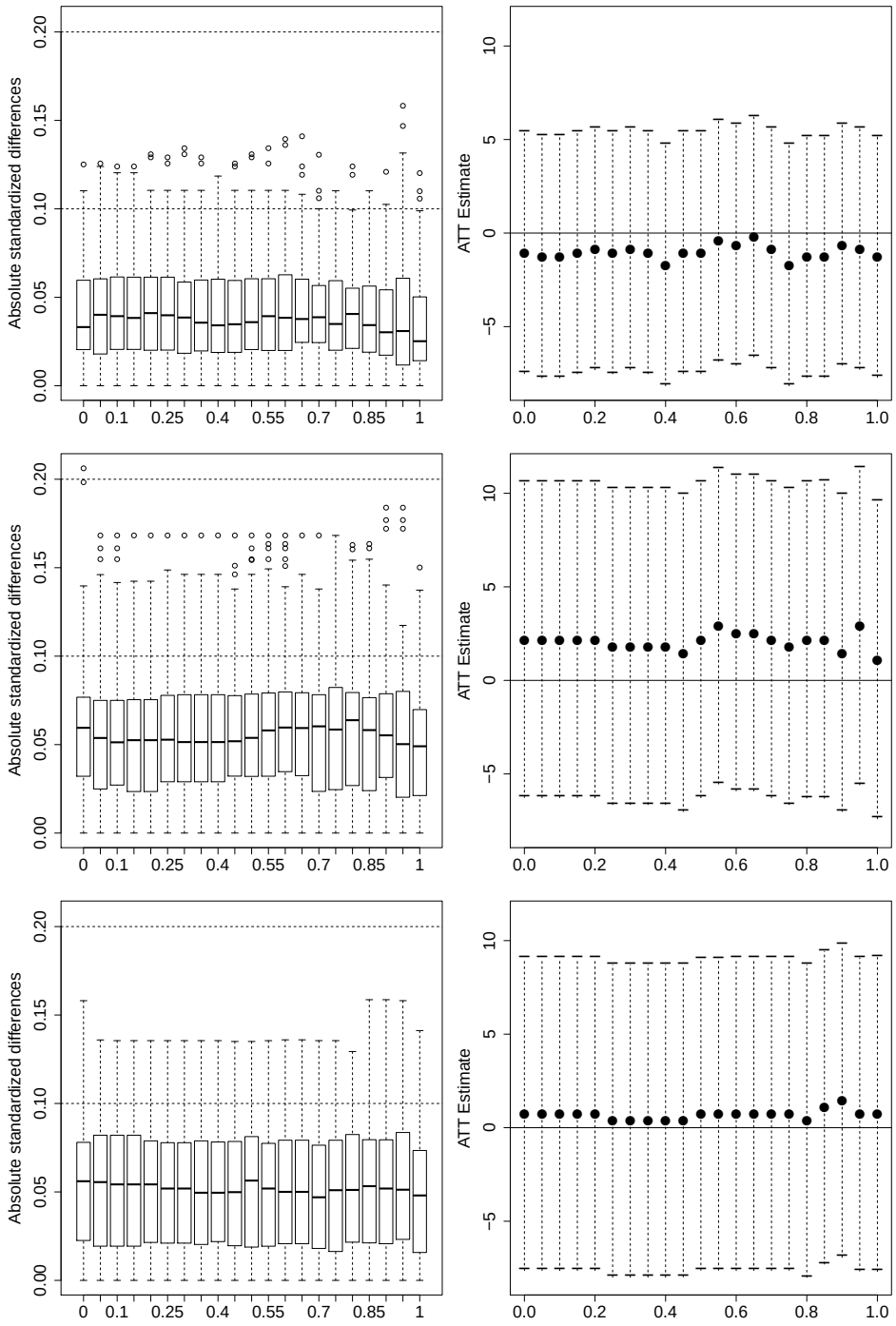


Figure A.10: The x-axis in each figure refers to the weight used in the DAPS score. The first column shows the balance across covariates for each weight. The second column shows outcome estimates for each weight. The rows correspond to the WC+DAPSm-S, WC+DAPSm-RE, and WC+DAPSm-FE methods, respectively, using a caliper of 0.2 on the propensity score.

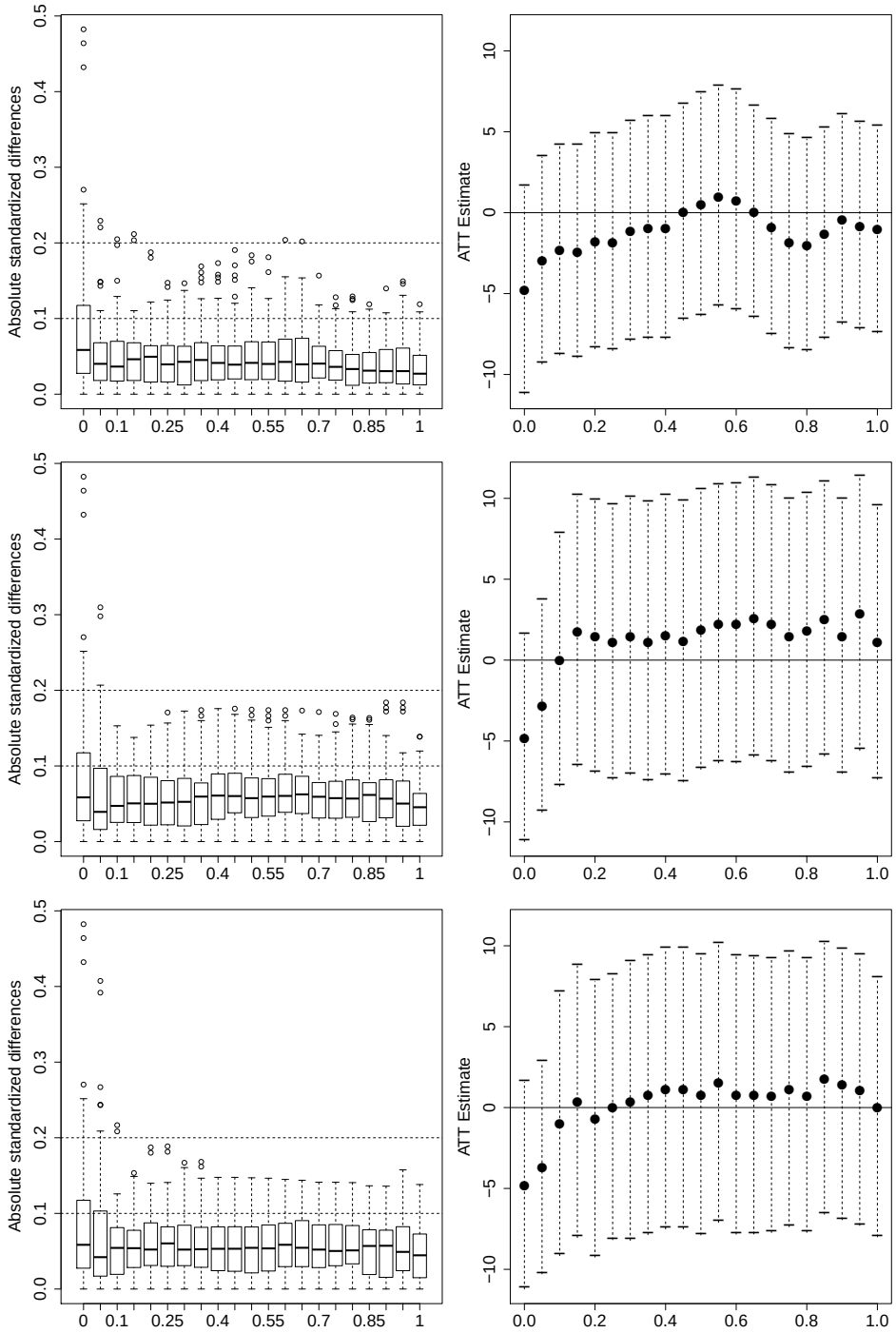


Figure A.11: The x-axis in each figure refers to the weight used in the DAPS score. The first column shows the balance across covariates for each weight. The second column shows outcome estimates for each weight. The rows correspond to the WC+DAPSm-S, WC+DAPSm-RE, and WC+DAPSm-FE methods, respectively, using a distance-adjusted propensity score caliper of 0.3 rather than a caliper on the propensity score as in Figure A.10.

Appendix B

Supplemental Materials to "Imputing satellite-derived aerosol optical depth using a multi-resolution spatial model and random forest for $PM_{2.5}$ prediction"

Animations are best viewed in Acrobat Reader.

B.1 Additional AOD Figures and Tables

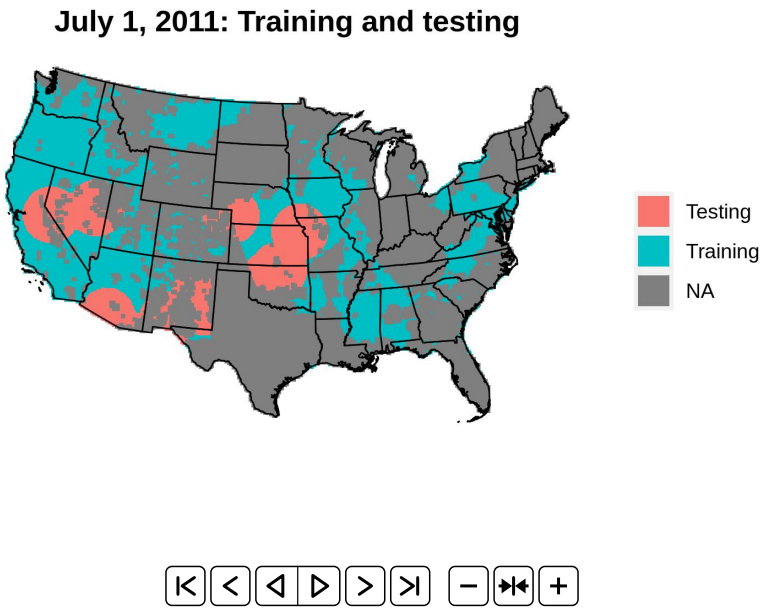


Figure B.1: Daily split between training and testing data.

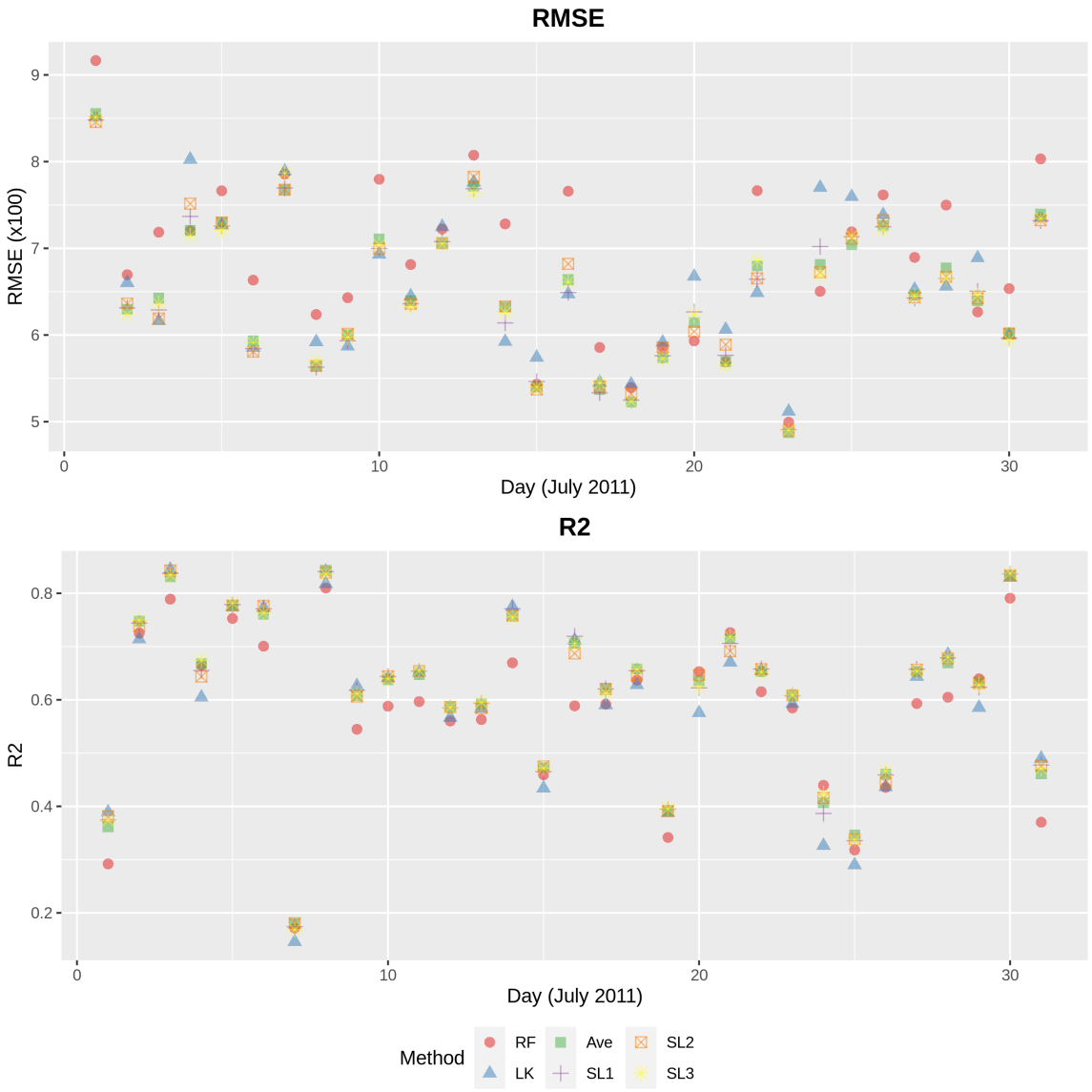


Figure B.2: Root mean-squared error (RMSE) and R^2 across days for LK, RF, average of RF and LK (Ave), SL: Overall (SL1), SL: Daily (SL2), and SL: Distance-based (SL3) methods.

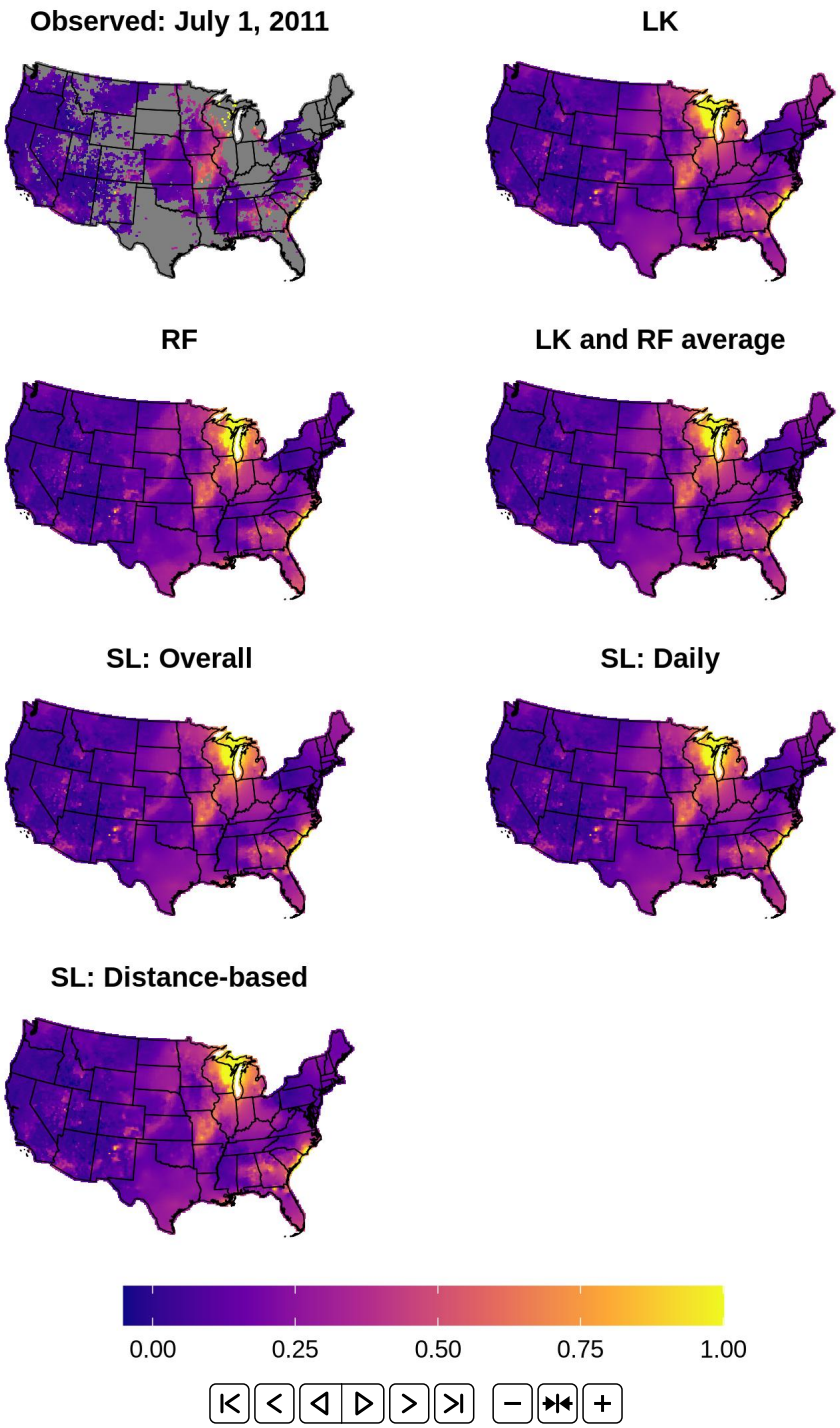


Figure B.3: Daily observed and predicted AOD values. Values outside of range are truncated for display.

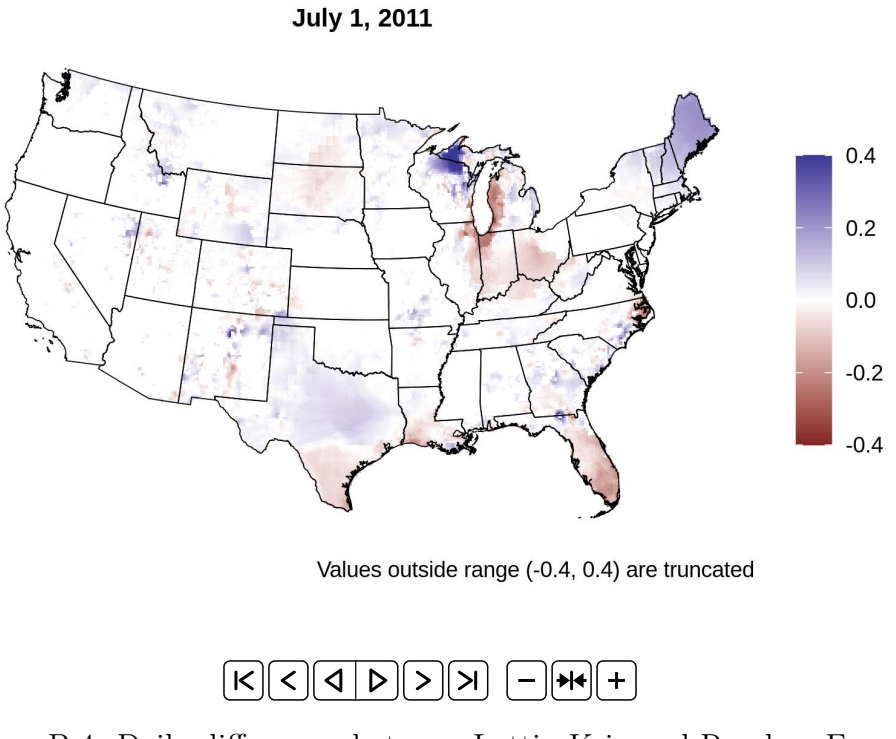


Figure B.4: Daily differences between LatticeKrig and Random Forest

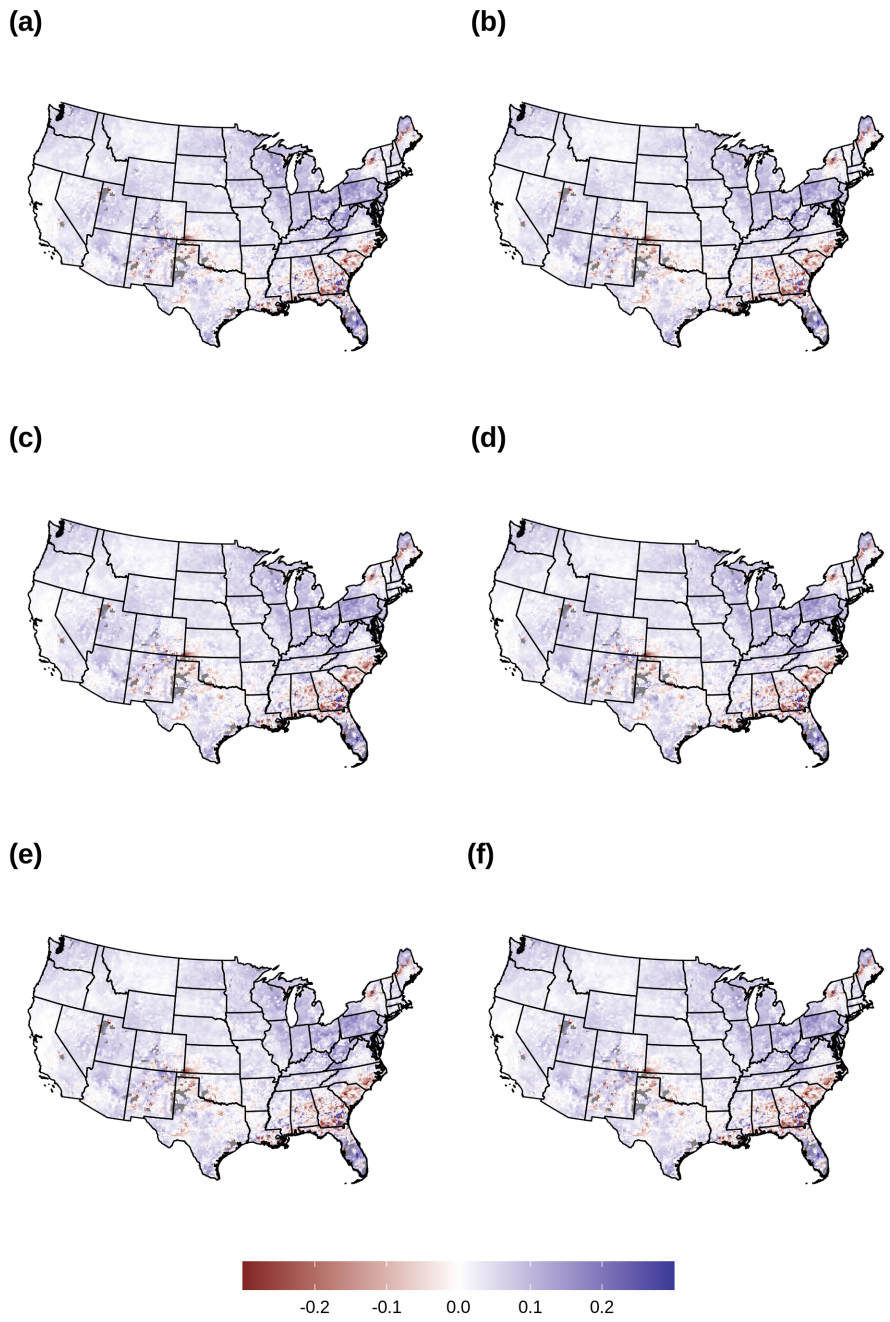


Figure B.5: Difference in average predictions and observed daily values for July 2011: **(a)** LK; **(b)** RF; **(c)** Average of LK and RF; **(d)** SL: Overall; **(e)** SL: Daily; **(f)** SL: Distance-based. Differences outside of range of $(-0.3, 0.3)$ are trimmed for figure appearance.

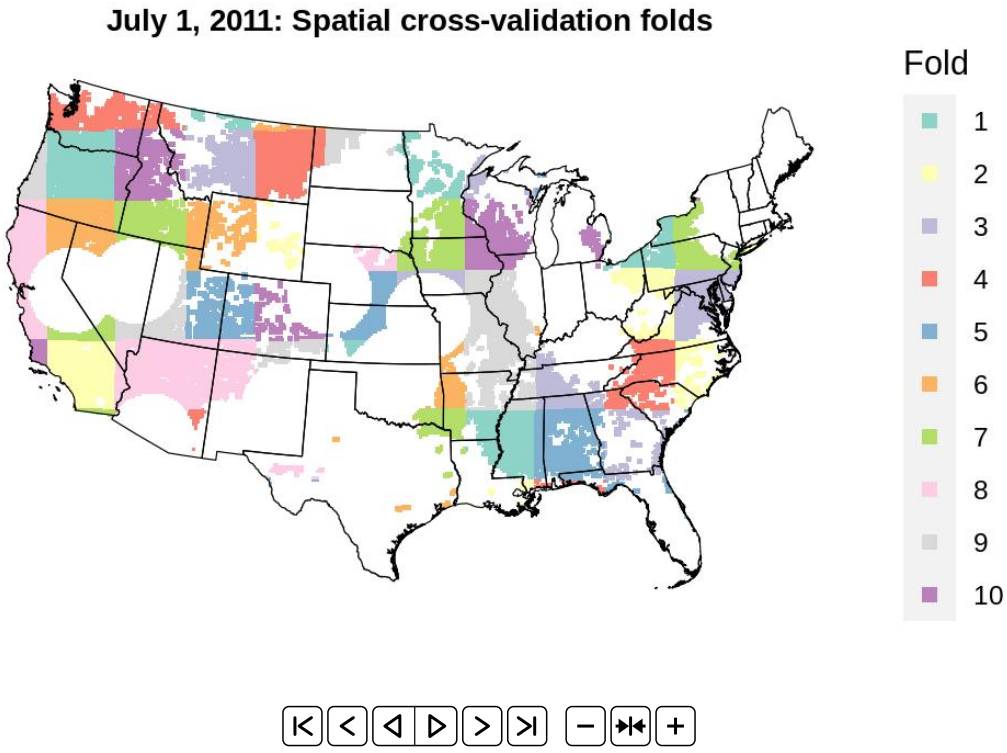


Figure B.6: Daily 10-fold spatially clustered CV. Each color represents a distinct fold, generated by the R package `blockCV`.

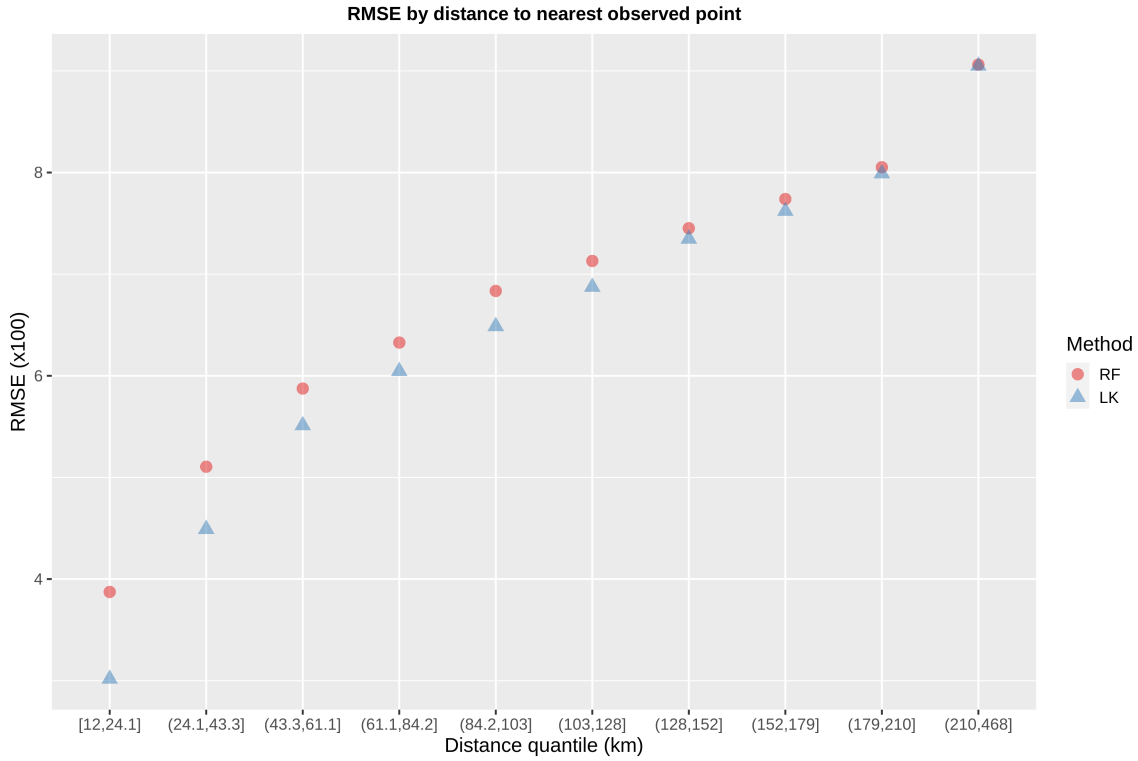


Figure B.7: Comparison of LatticeKrig and Random Forest at different distances between test data and training data across all days.

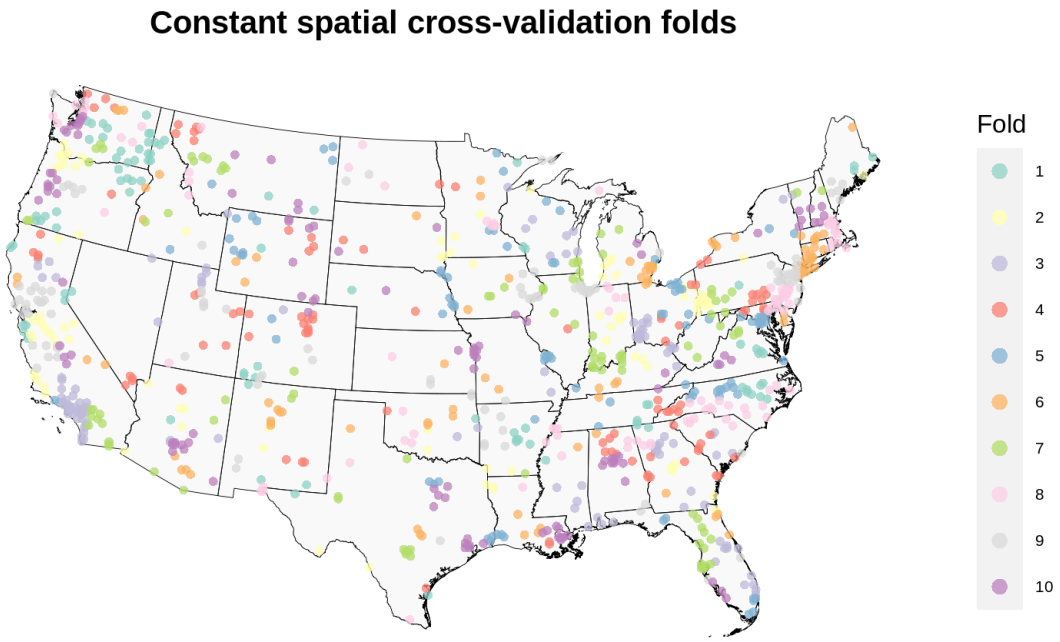


Figure B.8: Constant spatial clustering cross-validation map for $PM_{2.5}$ analyses.

B.2 Additional $PM_{2.5}$ Figures and Results

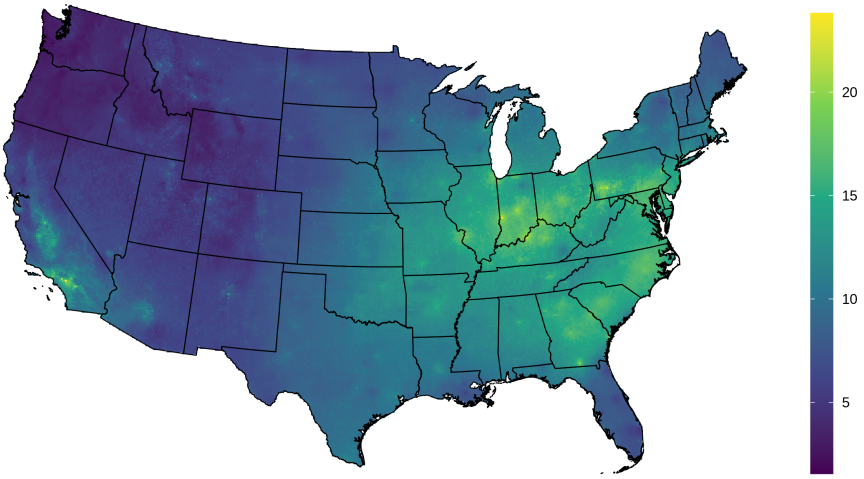


Figure B.9: Average July 2011 $PM_{2.5}$ predicted map using imputed AOD random forest model for `mtry` = 8.

Features	M1	M2	M3	M4	M5
Convolution layer $PM_{2.5}$	35.46	33.19	30.65	33.54	26.32
CMAQ-X Coordinate	18.77	13.52	13.12	15.88	12.18
GEOS-Chem		6.69	6.31		5.79
CMAQ-Y Coordinate	7.83	6.58	6.05	6.92	4.91
Convective available potential energy	8.71	6.64	5.66	7.20	4.60
Pressure at surface	6.24	5.95	5.30	6.40	4.76
Surface DW longwave radiation flux	7.46	6.35	5.14	6.62	5.61
Temperature	6.24	5.72	4.94	5.93	4.49
Imputed AOD			4.43		
Elevation	5.95	4.81	4.40	5.02	4.69
AOD/GEOS-Chem combination				3.47	
Observed AOD					2.99
Potential evaporation	3.42	3.11	3.07	3.18	2.86
Population density	3.47	3.05	2.81	3.03	2.79
Relative humidity	3.30	2.83	2.67	2.91	1.74
Day	2.52	2.49	2.24	2.31	1.46
Impervious surface (%)	2.48	2.01	1.89	2.07	1.88
Surface DW shortwave radiation flux	1.92	1.83	1.83	1.81	1.58
Percent forest cover	1.64	1.38	1.23	1.37	1.53
u-direction wind-speed	1.27	1.20	1.09	1.10	0.67
v-direction wind speed	1.28	1.13	0.98	1.10	1.18
Precipitation	0.85	0.59	0.68	0.72	0.15
Total length of local road	0.92	0.76	0.65	0.78	0.83
Faction of total precipitation that is convective	0.64	0.51	0.58	0.53	0.03
Day of the Week	0.50	0.50	0.43	0.46	0.37
AOD Missing Indicator	0.17	0.17	0.18	0.32	
Total length of limited-access road	0.10	0.10	0.09	0.09	0.11
Total length of highway	0.13	0.11	0.08	0.10	0.11
EPA 2011 emission inventory	0.07	0.06	0.06	0.07	0.07

Table B.1: Feature importance (permutation-based, mean decrease in accuracy) from spatio-temporal random forest model based on `mtry` = 4.

Description_Features	M1	M2	M3	M4	M5
Convolution layer $PM_{2.5}$	50.44	47.58	45.01	48.13	39.51
CMAQ-X Coordinate	19.59	13.80	12.02	17.62	10.36
GEOS-Chem		5.90	5.69		5.38
CMAQ-Y Coordinate	6.13	5.32	4.82	5.48	4.06
Pressure at surface	5.47	5.21	4.59	4.93	4.11
Surface DW longwave radiation flux	5.92	5.13	4.56	5.65	4.14
Convective available potential energy	6.13	5.15	3.99	5.59	2.82
Temperature	4.63	4.61	3.96	4.67	2.99
Imputed AOD			3.50		
Elevation	4.17	3.71	3.33	3.77	3.68
AOD/GEOS-Chem combination				2.70	
Observed AOD					2.30
Population density	3.15	2.84	2.67	2.84	2.52
Relative humidity	2.49	2.16	2.18	2.40	1.15
Potential evaporation	2.21	2.04	1.99	2.06	1.93
Impervious surface (%)	2.03	1.77	1.65	1.74	1.55
Day	1.27	1.27	1.33	1.17	0.78
Surface DW shortwave radiation flux	1.25	1.16	1.18	1.14	0.93
Percent forest cover	1.30	1.12	1.16	1.16	1.22
u-direction wind-speed	0.85	0.87	0.77	0.75	0.56
v-direction wind speed	0.84	0.78	0.67	0.79	0.96
Precipitation	0.63	0.53	0.51	0.61	0.17
Total length of local road	0.63	0.55	0.48	0.56	0.70
Faction of total precipitation that is convective	0.47	0.41	0.42	0.40	0.03
Day of the Week	0.22	0.25	0.22	0.23	0.17
AOD Missing Indicator	0.08	0.08	0.08	0.12	
Total length of limited-access road	0.08	0.06	0.07	0.07	0.09
Total length of highway	0.09	0.08	0.07	0.08	0.08
EPA 2011 emission inventory	0.06	0.05	0.05	0.05	0.06

Table B.2: Feature importance (mean decrease in accuracy) from pooled random forest model based on `mtry` = 8.

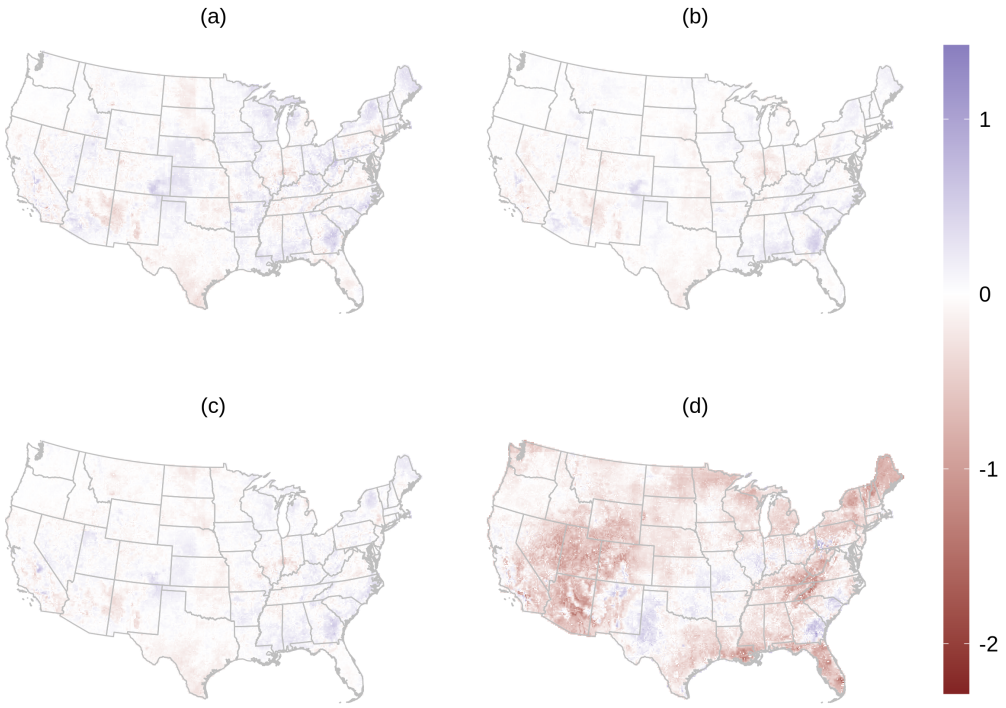


Figure B.10: Difference between the imputed AOD RF model (M3) and other RF models in average July 2011 $PM_{2.5}$ predictions for `mtry` = 8: **(a)** M1: model with no AOD or GEOS-Chem; **(b)** M2: GEOS-Chem; **(c)** M4: Replacing missing values of AOD with GEOS-Chem; **(d)** M5: Train on observed AOD, and predict by replacing missing AOD values with imputed values.

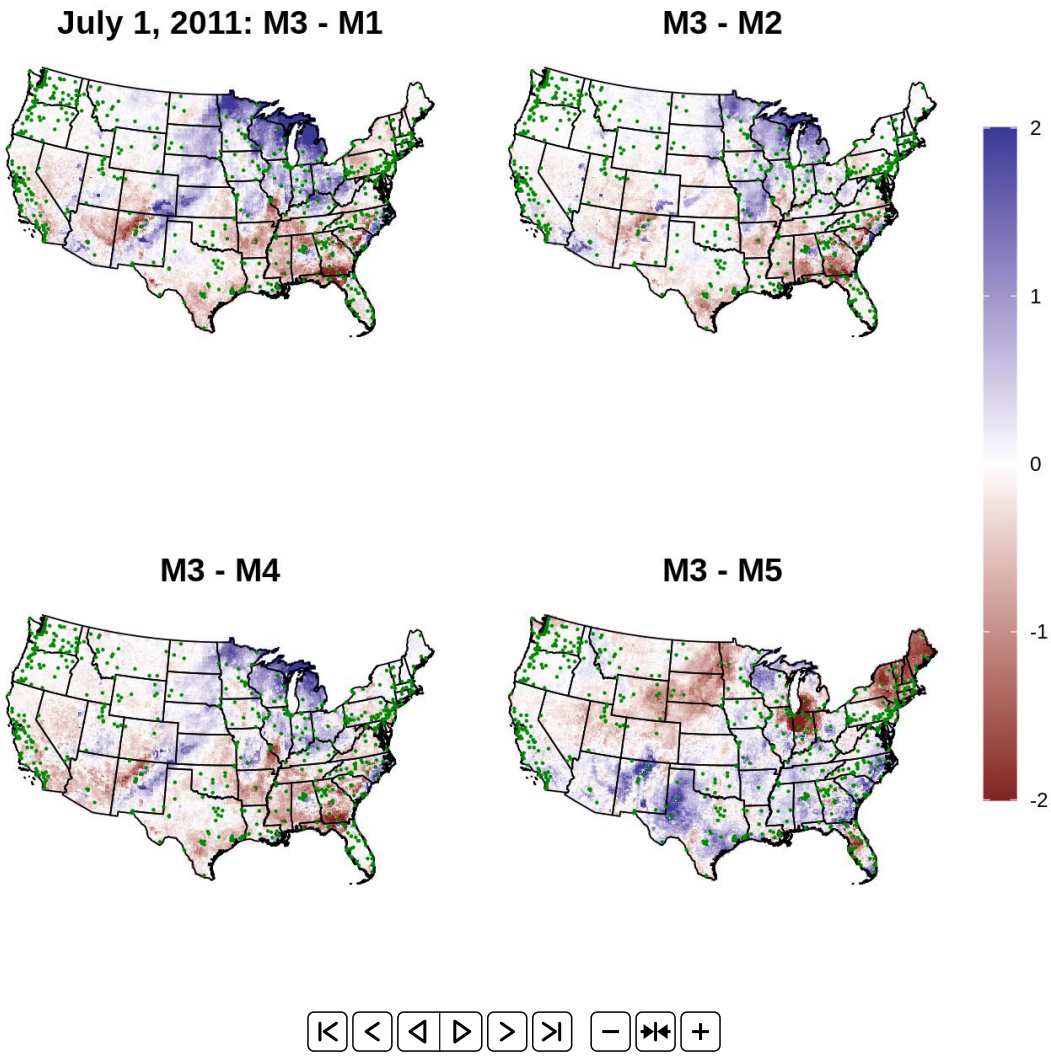
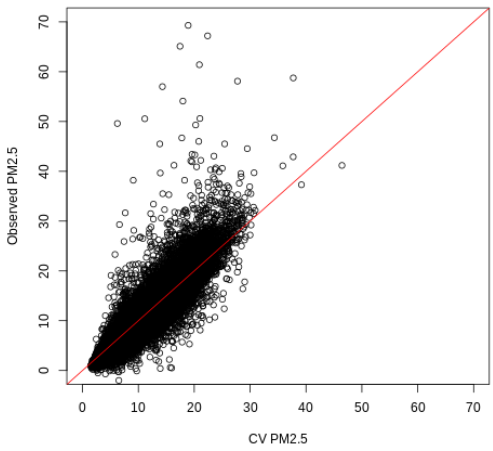
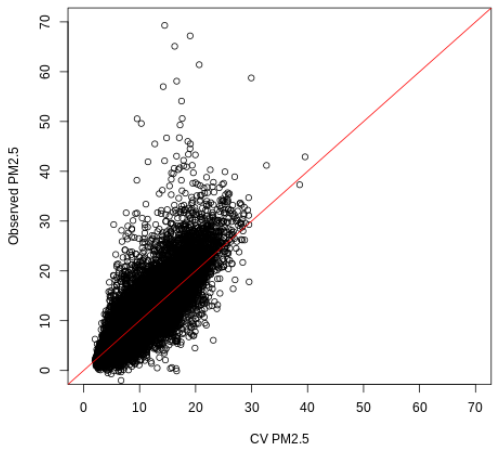


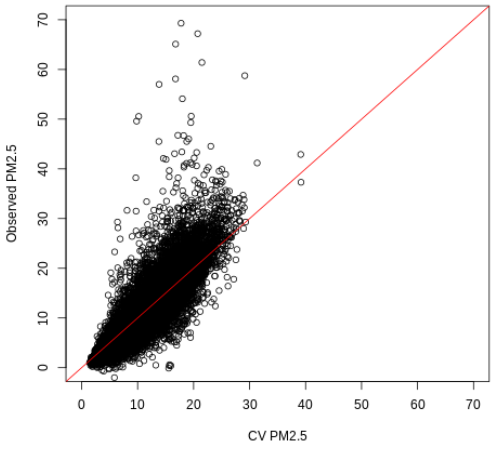
Figure B.11: Difference between the imputed AOD RF model (M3) and other RF models in daily $PM_{2.5}$ predictions for `mtry` = 4. M1: model with no AOD or GEOS-Chem; M2: GEOS-Chem; M4: Replacing missing values of AOD with GEOS-Chem; M5: Train on observed AOD, and predict by replacing missing AOD values with imputed values. Green points denote cells with observed $PM_{2.5}$ monitors. Values outside of range truncated for display.



(a) Random

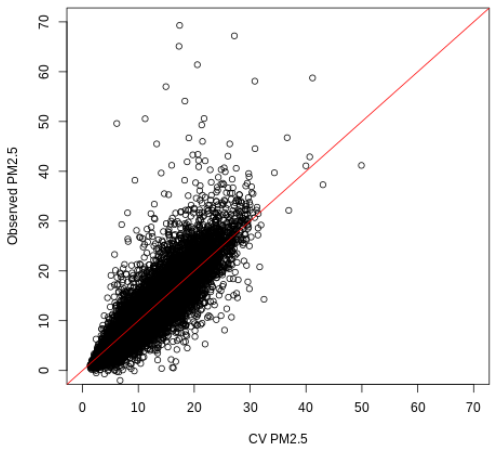


(b) Constant spatial cluster

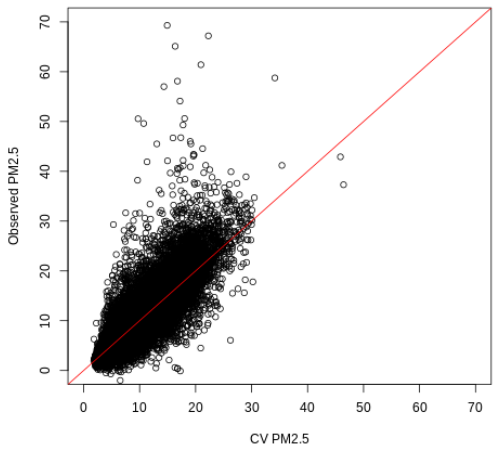


(c) Varying spatial cluster

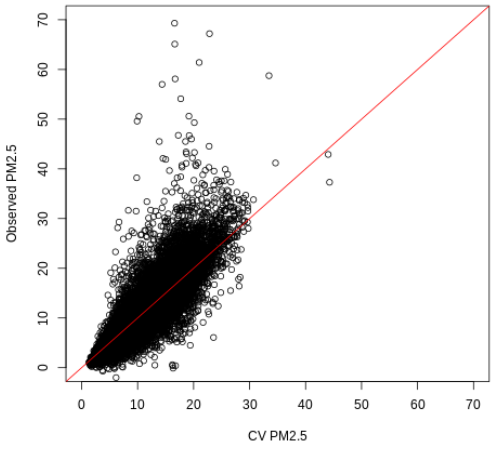
Figure B.12: Scatter plots comparing observed $PM_{2.5}$ values with cross-validation predictions from spatio-temporal random forest models including imputed AOD (M3) with `mtry` $= 4$ for **(a)** random cross-validation, **(b)** constant spatially clustered cross-validation, and **(c)** varying spatially clustered cross-validation. Axes limited to $(0, 70)$ for clarity of visual presentation. Red line is the $y = x$ line.



(a) Random

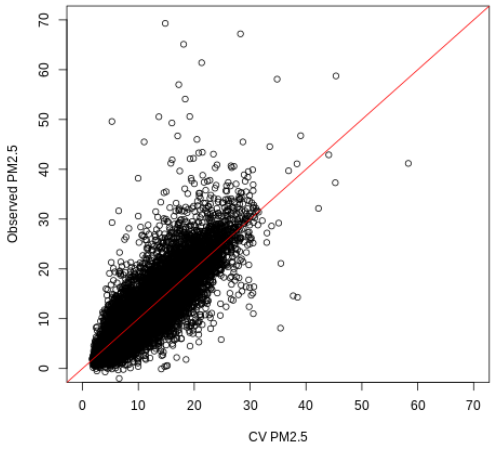


(b) Constant spatial cluster

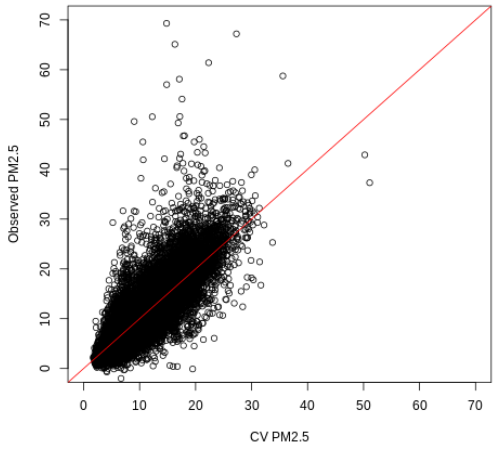


(c) Varying spatial cluster

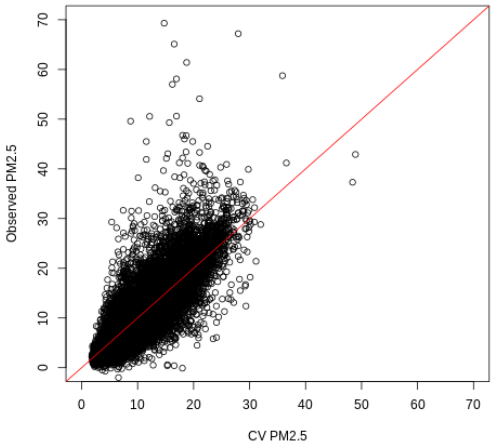
Figure B.13: Scatter plots comparing observed $PM_{2.5}$ values with cross-validation predictions from spatio-temporal random forest models including imputed AOD (M3) with `mtry` $= 8$ for **(a)** random cross-validation, **(b)** constant spatially clustered cross-validation, and **(c)** varying spatially clustered cross-validation. Axes limited to $(0, 70)$ for clarity of visual presentation. Red line is the $y = x$ line.



(a) Random



(b) Constant spatial cluster



(c) Varying spatial cluster

Figure B.14: Scatter plots comparing observed $PM_{2.5}$ values with cross-validation predictions from daily random forest models including imputed AOD (M3) with `mtry` $= 8$ for **(a)** random cross-validation, **(b)** constant spatially clustered cross-validation, and **(c)** varying spatially clustered cross-validation. Axes limited to $(0, 70)$ for clarity of visual presentation. Red line is the $y = x$ line.

		Daily					Spatio-temporal				
Setting	AOD Status	M1a	M2a	M3a	M4a	M5a	M1b	M2b	M3b	M4b	M5b
Intercept											
Random	All	-0.19	-0.21	-0.23	-0.23	-0.37	-0.71	-0.73	-0.41	-0.39	-0.75
Random	Missing	-0.26	-0.29	-0.31	-0.27	-0.78	-0.76	-0.79	-0.44	-0.40	-1.01
Random	Observed	-0.09	-0.09	-0.14	-0.16	0.06	-0.64	-0.66	-0.37	-0.35	-0.47
Constant cluster	All	-0.32	-0.33	-0.35	-0.37	-0.51	-0.42	-0.44	-0.47	-0.47	-0.92
Constant cluster	Missing	-0.60	-0.64	-0.65	-0.61	-1.12	-0.73	-0.74	-0.78	-0.73	-1.48
Constant cluster	Observed	0.03	0.05	0.02	-0.04	0.05	-0.02	-0.05	-0.10	-0.12	-0.33
Varying cluster	All	-0.40	-0.41	-0.44	-0.44	-0.53	-1.13	-1.15	-1.15	-1.17	-1.26
Varying cluster	Missing	-0.64	-0.67	-0.69	-0.64	-1.05	-1.33	-1.35	-1.33	-1.32	-1.66
Varying cluster	Observed	-0.07	-0.06	-0.12	-0.14	-0.02	-0.88	-0.90	-0.94	-0.95	-0.84
Slope											
Random	All	1.01	1.01	1.02	1.02	1.00	1.06	1.06	1.03	1.03	1.05
Random	Missing	1.02	1.02	1.02	1.02	1.02	1.07	1.07	1.03	1.04	1.06
Random	Observed	1.00	1.01	1.01	1.01	0.99	1.06	1.06	1.03	1.02	1.04
Constant cluster	All	1.04	1.04	1.04	1.05	1.03	1.06	1.06	1.06	1.06	1.08
Constant cluster	Missing	1.06	1.06	1.06	1.07	1.06	1.08	1.07	1.08	1.08	1.11
Constant cluster	Observed	1.02	1.02	1.03	1.02	1.02	1.03	1.03	1.04	1.03	1.06
Varying cluster	All	1.05	1.05	1.05	1.05	1.03	1.11	1.12	1.11	1.12	1.11
Varying cluster	Missing	1.06	1.06	1.06	1.07	1.05	1.13	1.13	1.12	1.13	1.13
Varying cluster	Observed	1.02	1.02	1.03	1.02	1.01	1.10	1.10	1.11	1.10	1.10

Table B.3: Intercept and slope estimates from daily and spatio-temporal random forest model for different 10-fold cross-validation settings. Internet and slope estimated from linear regression model Observed = β_0 + β_1Predicted.

		Daily					Spatio-temporal				
Setting	AOD Status	M1a	M2a	M3a	M4a	M5a	M1b	M2b	M3b	M4b	M5b
Random	Central	0.07	-0.04	-0.12	0.01	1.19	-1.73	-1.77	-0.83	-0.76	-0.92
	East North Central	-0.66	-0.67	-0.68	-0.65	-0.62	-1.82	-1.91	-1.17	-1.13	-1.49
	Northeast	-0.17	-0.24	-0.28	-0.23	0.24	-1.03	-0.98	-0.49	-0.47	-0.99
	Northwest	0.32	0.35	0.34	0.33	0.70	-0.70	-0.71	-0.35	-0.35	-0.43
	South	0.28	0.23	0.10	0.22	-0.32	-0.84	-0.88	-0.31	-0.17	-1.14
	Southeast	-0.38	-0.39	-0.35	-0.38	-1.61	-0.99	-1.04	-0.44	-0.43	-1.51
	Southwest	-0.00	-0.12	-0.18	-0.20	0.49	-1.47	-1.59	-0.99	-0.89	-1.05
	West	0.10	0.14	0.03	-0.01	0.47	-1.49	-1.61	-1.02	-1.01	-1.21
	West North Central	0.71	0.75	0.72	0.76	1.11	-0.90	-0.87	-0.68	-0.65	-0.74
Constant cluster	Central	-0.11	-0.31	-0.35	-0.43	1.17	-0.88	-0.97	-1.17	-1.09	-1.33
	East North Central	-0.94	-0.93	-0.95	-0.97	-1.04	-1.59	-1.72	-1.79	-1.72	-2.26
	Northeast	-0.35	-0.41	-0.45	-0.41	-0.08	-0.48	-0.57	-0.64	-0.55	-1.32
	Northwest	0.06	0.08	0.09	0.09	0.42	-0.34	-0.31	-0.30	-0.33	-0.57
	South	-0.25	-0.36	-0.54	-0.37	-0.83	-0.83	-0.98	-1.08	-0.93	-2.36
	Southeast	-1.09	-1.15	-1.13	-1.08	-2.45	-1.35	-1.34	-1.35	-1.33	-2.60
	Southwest	-0.22	-0.33	-0.44	-0.45	0.15	0.10	0.02	-0.18	-0.19	-0.60
	West	0.87	1.01	1.12	0.80	1.27	0.21	0.22	0.16	0.11	-0.13
	West North Central	0.84	0.85	0.86	0.86	1.29	0.86	0.81	0.76	0.78	0.59
Varying cluster	Central	0.41	0.20	0.12	0.08	1.75	-1.93	-2.03	-2.28	-2.12	-1.27
	East North Central	-1.11	-1.07	-1.07	-1.09	-0.98	-2.43	-2.55	-2.55	-2.54	-2.41
	Northeast	-0.62	-0.68	-0.69	-0.68	-0.25	-1.71	-1.64	-1.63	-1.69	-1.74
	Northwest	0.22	0.22	0.24	0.24	0.59	-0.89	-0.87	-0.88	-0.89	-0.76
	South	-0.28	-0.38	-0.51	-0.36	-0.67	-1.83	-1.95	-2.01	-1.91	-2.43
	Southeast	-1.14	-1.18	-1.17	-1.11	-2.06	-2.00	-2.08	-1.96	-2.02	-2.70
	Southwest	-0.18	-0.21	-0.40	-0.41	0.33	-2.17	-2.26	-2.43	-2.66	-2.09
	West	0.81	0.90	0.78	0.75	1.25	-2.05	-2.16	-2.27	-2.19	-1.56
	West North Central	0.64	0.65	0.65	0.66	1.05	-1.29	-1.26	-1.32	-1.34	-1.25

Table B.4: Regional intercept estimates for daily and spatio-temporal random forest model for different 10-fold cross-validation settings.

		Daily					Spatio-temporal				
Setting	AOD Status	M1a	M2a	M3a	M4a	M5a	M1b	M2b	M3b	M4b	M5b
Random	Central	1.01	1.02	1.02	1.01	0.93	1.12	1.12	1.06	1.05	1.06
	East North Central	1.02	1.02	1.02	1.02	0.99	1.13	1.13	1.07	1.07	1.08
	Northeast	1.01	1.02	1.02	1.02	0.98	1.08	1.07	1.03	1.03	1.06
	Northwest	0.87	0.86	0.86	0.86	0.71	1.15	1.16	1.07	1.07	1.05
	South	0.96	0.96	0.98	0.96	0.97	1.07	1.07	1.02	1.00	1.07
	Southeast	1.02	1.03	1.02	1.02	1.07	1.07	1.08	1.03	1.03	1.08
	Southwest	0.95	0.96	0.97	0.97	0.82	1.19	1.20	1.12	1.11	1.07
	West	1.01	1.01	1.02	1.02	0.96	1.15	1.17	1.10	1.10	1.11
	West North Central	0.91	0.90	0.91	0.90	0.82	1.14	1.13	1.11	1.11	1.10
Constant cluster	Central	1.07	1.08	1.08	1.09	0.98	1.13	1.13	1.14	1.14	1.15
	East North Central	1.04	1.03	1.03	1.03	1.01	1.09	1.10	1.11	1.10	1.13
	Northeast	1.04	1.04	1.05	1.04	1.01	1.04	1.05	1.05	1.05	1.09
	Northwest	0.91	0.91	0.90	0.90	0.75	1.02	1.02	1.01	1.02	1.03
	South	1.01	1.02	1.04	1.02	1.01	1.06	1.07	1.08	1.07	1.18
	Southeast	1.08	1.08	1.08	1.08	1.13	1.10	1.10	1.10	1.10	1.17
	Southwest	1.00	1.01	1.03	1.03	0.89	0.98	0.99	1.02	1.02	1.02
	West	1.03	1.02	1.00	1.04	0.97	1.09	1.09	1.09	1.10	1.12
	West North Central	0.90	0.90	0.90	0.90	0.80	0.92	0.92	0.93	0.93	0.94
Varying cluster	Central	1.03	1.04	1.05	1.05	0.94	1.18	1.19	1.20	1.19	1.14
	East North Central	1.05	1.04	1.04	1.05	1.01	1.17	1.17	1.17	1.17	1.15
	Northeast	1.07	1.07	1.07	1.07	1.03	1.14	1.13	1.13	1.14	1.13
	Northwest	0.88	0.88	0.88	0.87	0.72	1.18	1.17	1.17	1.17	1.10
	South	1.01	1.01	1.03	1.01	1.00	1.15	1.16	1.17	1.16	1.19
	Southeast	1.08	1.08	1.08	1.08	1.10	1.15	1.16	1.14	1.15	1.18
	Southwest	1.00	1.00	1.03	1.03	0.87	1.30	1.31	1.33	1.37	1.22
	West	1.00	0.99	1.00	1.00	0.93	1.26	1.27	1.29	1.28	1.20
	West North Central	0.92	0.92	0.92	0.92	0.83	1.19	1.19	1.19	1.20	1.16

Table B.5: Regional slope estimates for daily and spatio-temporal random forest model for different 10-fold cross-validation settings.

B.3 Additional LatticeKrig modeling details

We follow the model description of lattice kriging (LatticeKrig or LK) laid out by Nychka et al. (2015). At a high-level, LK models the spatial process using several levels of two-dimensional basis functions, which are laid out on a grid and approximately double with each successive layer. These basis functions are compact, which means that for a particular point only a small number of basis function are used to make the prediction. The coefficients associated with the basis functions are assumed to be correlated, and this structure can flexibly model observed spatial covariance structures. Estimation proceeds through a likelihood-based approach after specifying various tuning parameters.

Following the notation of Nychka et al. (2015), we observe $\{y_i\}$ at locations $\{\boldsymbol{x}_i\}$ for $i = 1, .., n$. We assume $\{y_i\}$ follow an additive model consisting of a mean function based on covariates, a spatial process, and a measurement error term:

$$y_i = \boldsymbol{Z}_i^T \boldsymbol{d} + g(\boldsymbol{x}_i) + \epsilon_i, \tag{B.1}$$

where $\boldsymbol{d}$ is a $p \times 1$ vector of fixed coefficients associated with the covariates $\boldsymbol{Z}_i$, and $g(\boldsymbol{x_i})$ denotes the spatial process. The mean-zero error terms ϵ_i are presumed to be independent and identically distributed, i.e., $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, where $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T$.

The overall spatial process $g(\boldsymbol{x}_i)$ can be written as a sum of L independent spatial processes $g_l(\boldsymbol{x}_i)$:

$$g(\boldsymbol{x}_i) = \sum_{l=1}^{L} g_l(\boldsymbol{x}_i) = \sum_{l=1}^{L} \sum_{j=1}^{m(l)} c_j^l \phi_{j,l}(\boldsymbol{x}_i), \tag{B.2}$$

where $\phi_{j,l}$ denotes the the lth level of resolution's jth basis function, and c_j^l denotes the coefficient associated with this basis function. Although the basis functions and number of levels are fixed (i.e., chosen), the coefficients for each level l, $\boldsymbol{c}^l = (c_1^l, ..., c_{m(l)}^l)^T$

are assumed to follow a multivariate normal with mean zero and covariance $\rho \boldsymbol{Q}_l^{-1}$:

$$\boldsymbol{c}^l \sim N(\boldsymbol{0}, \rho \boldsymbol{Q}_l^{-1}). \tag{B.3}$$

Each level's spatial process is independent with marginal variance $\rho\alpha_l$ subject to the constraint $\sum_{l=1}^{L} \alpha_l = 1$, so that the marginal variance of the overall spatial process $g(\boldsymbol{x}_i)$ is ρ.

Let m denote the total number of basis functions, and for simplicity consider a single level $L = 1$, so that $g(\boldsymbol{x}) = \sum_{j=1}^{m} c_j \phi_j(\boldsymbol{x})$. Then, for any two locations $\boldsymbol{x}$ and $\boldsymbol{x'}$, the covariance is given as:

$$Cov(g(\boldsymbol{x}), g(\boldsymbol{x'})) = \rho \sum_{j=1}^{m} \sum_{k=1}^{m} \boldsymbol{Q}_{j,k}^{-1} \phi_j(\boldsymbol{x}) \phi_k(\boldsymbol{x'}). \tag{B.4}$$

Denote $\boldsymbol{\Phi}$ as the $n \times m$ matrix of basis functions evaluated at the observed locations. The full marginal distribution $\boldsymbol{y}$ is then given as

$$\boldsymbol{y} \sim N(\boldsymbol{Zd}, \rho \boldsymbol{\Phi Q}^{-1} \boldsymbol{\Phi}^T + \sigma^2 \boldsymbol{I}). \tag{B.5}$$

By setting $\lambda = \sigma^2/\rho$ (a noise to signal ratio), and $M_\lambda = \boldsymbol{\Phi Q}^{-1} \boldsymbol{\Phi}^T + \lambda \boldsymbol{I}$, this may be further re-written as

$$\boldsymbol{y} \sim N(\boldsymbol{Zd}, \rho \boldsymbol{M}_\lambda). \tag{B.6}$$

Nychka et al. (2015) provide further details on estimation of the key parameters using the profile log-likelihood such that the likelihood only depends on λ and parameters determining $\boldsymbol{Q}$.

Nychka et al. (2015) propose using two-dimensional radial basis functions (RBF) using the Wendland functions that have a compact support (Wendland, 1995). These

basis functions take the following form for scaled distance $0 \leq d \leq 1$:

$$\phi(d) = (1-d)^6(35d^2 + 18d + 3)/3. \tag{B.7}$$

By default, the distance is scaled to be 2.5 times the grid spacing for each level of resolution. For example, if basis functions are defined to be 100km apart, a particular basis function will be defined as 0 for points outside of a 250km radius of where it is placed. The basis functions are thus defined as:

$$\phi_j^*(\boldsymbol{x}) = \phi(||\boldsymbol{x} - \mu_j||/\theta), \tag{B.8}$$

where μ_j is the location of the basis function, and θ is set to determine the amount of overlap. Nychka et al. (2015) additionally recommend and implement basis normalization by default as part of their estimation. Normalization re-scales the basis functions in the case of a single level as

$$\phi_t(\boldsymbol{x}) = \frac{\phi_t^*(\boldsymbol{x})}{\sqrt{\sum_{j=1}^m \sum_{k=1}^m \boldsymbol{Q}_{j,k}^{-1} \phi_j^*(\boldsymbol{x}) \phi_k^*(\boldsymbol{x})}}, \tag{B.9}$$

such that

$$Cov(g(\boldsymbol{x}), g(\boldsymbol{x})) = \rho \frac{\sum_{j=1}^m \sum_{k=1}^m \boldsymbol{Q}_{j,k}^{-1} \phi_j^*(\boldsymbol{x}) \phi_k^*(\boldsymbol{x})}{\sum_{j=1}^m \sum_{k=1}^m \boldsymbol{Q}_{j,k}^{-1} \phi_j^*(\boldsymbol{x}) \phi_k^*(\boldsymbol{x})} = \rho. \tag{B.10}$$

Thus the normalization process ensures a constant marginal variance. Nychka et al. (2015) recommend this to reduce edge effects and for better approximating stationary covariance functions. For multiple levels of resolution this process is carried out separately for each level of resolution.

B.3.1 Tuning

By default, package implementation of LK uses the profile log-likelihood to estimate λ. We use a mean-squared error (MSE) approach with a validation portion of the data to tune the other parameters of our LK model (e.g., see Supplemental Section 1.2 of Heaton et al. (2019)). The validation data was constructed from the training data in a similar manner to the testing data.

For the fixed covariate portion of the LK model, we include a simple form using just a few covariates. By default, the `LatticeKrig` package includes the spatial coordinates as fixed predictors in $\boldsymbol{Z}$. In addition, we include the interaction between the coordinates, GEOS-Chem, the interaction between each coordinate and GEOS-Chem, and elevation as the fixed predictors in the model. No variable selection was performed – we instead focused on tuning the spatial aspect of the model. As a result, some important variables may have been excluded from the mean model.

The tuning parameters for LK were chosen from values of `a.wght` = (4.1, 4.5, 6, 8, 10, 12) and values of `nu` (determining α_l) = $(0.1, 0.25, 0.5, 0.75, 1, 1.25)$. These 36 combinations of parameters were tried with two possible combinations of `nlevel` and `NC` for a total of 72 combinations; `nlevel` = 4 and NC = 30, or `nlevel` = 5 and NC = 15. Each combination of `nlevel`/`NC` results in close to 50,000 basis functions, with the finest level of resolution having basis functions roughly 20km apart. However, these methods differ with respect to the distance between basis functions at the coarsest level.

For LK, we find the best prediction MSE on the validation data is usually associated with high `a.wght` (10, 12) and small `nu` (0.1, 0.25), together with either combination of `nlevel`/`NC`. We opt for `nlevel = 5`, `NC = 15`, `a.wght = 12`, and `nu = 0.1` as the final set of parameters for this reason.

B.4 Random forest tuning for AOD prediction

We consider 22 possible variables: the projected centroid coordinates, elevation, 2011 emission inventory, forest cover, impervious surface, total lengths of highway, limited-access road, and local road, population density, potential evaporation, surface DW longwave radiation flux, surface DW shortwave radiation flux, convective available potential energy, fraction of total precipitation that is convective, precipitation, relative humidity, temperature, u- and v-direction wind speed, pressure at surface, and GEOS-Chem AOD. For random forest, we tried every value of m (`mtry`) between 1 and 22, where $p = 22$. Additionally, we varied the n_{size} (`min.node.size`) value between 2, 5, and 8. The total combination of parameters tried is 66. For every combination, we set $B = 500$ (`num.trees`). We used a validation data as with tuning LK, constructed in a manner similar to the testing data.

In general the node size n_{size} does not substantially change the prediction MSE for a given m, so we restrict the value to be 5 (the default), and instead focus on selecting m. In validation experiments, we found the best m was around 7, close to the default that would be suggested by the `randomForest` package. We set the number of trees in the daily AOD prediction models to 2000.

We initially considered an additional predictor in the form of nearest-neighbor AOD (nnAOD) for AOD prediction models. In this setting, nnAOD is determined based on the distance between observations in the validation dataset to the training dataset. However, in some preliminary explorations, we found that including nnAOD did not help in the prediction performance on the basis of MSE, and in some settings, the inclusion of this predictor may actually slightly decrease performance. We posit that in the training dataset, there will almost always be near-by observed AOD values. In contrast, the validation or test dataset in a spatial setting will consist of a large number of points that are more distant from the training dataset locations. Thus, even though this predictor was likely highly important for the training fits, random forest

then extrapolates when predicting to the test dataset and performance suffers. Some results demonstrating this decrease in performance are included in following section. These results echo concerns in Hengl et al. (2018) regarding the use of random forest in spatial applications when extrapolating to unobserved areas.

B.5 Consideration of nearest-neighbor AOD and OOB metrics in random forest

We considered two additional features for random forest in AOD prediction: (1) A weighted average of nearest-neighbor AOD observations, and (2) a weighted average of the distances of these nearest-neighbor points. A priori, we expected that these predictors would produce very strong in-sample fit, but that they may not improve (or may even diminish) performance when making predictions out of sample in the setting we observe. This is because we are making spatial predictions in large areas where there is no observed data, and the nearest neighbors may be quite distant and unlike the training data. Furthermore, we posit that the out-of-bag (OOB) metrics from random forest will be misleading in this spatial setting where the training and testing data may be quite unlike each other.

To assess whether the inclusion of these two features can help in prediction, and whether the OOB metric is misleading, we carry out a set of experiments. Based on the training data considered in the main analysis, for each day, we consider 2 separate validation datasets (1) The validation set consists of 10 points and any other point within a 100km radius, and (2) the validation set consists of 10 points and any other point within a 250km radius. Moreover, we consider two particular choices for the m (`mtry`) parameter for random forest in combination with the above validation datasets: (A) $m = 4$ and (B) $m = 12$. For each day, we assess both the OOB mean-squared error (MSE), as well as the validation MSE, for both the case where the two

additional nearest-neighbor features were included and in the case where they were not included.

Our results showed that for all cases, the OOB MSE prefers the inclusion of the additional 2 features. However, the validation MSE tended to actually prefer fewer predictors rather than more predictors on more days when $m = 12$. With $m = 4$, the additional features are strongly predictive but are not selected for as many splits, therefore less of an issue is posed, and the results are similar regardless of the inclusion of the additional nearest-neighbor features. Generally, the OOB MSE is substantially lower than the validation MSE as well, suggesting that OOB metrics are not well suited for this particular kind of spatial prediction problem.

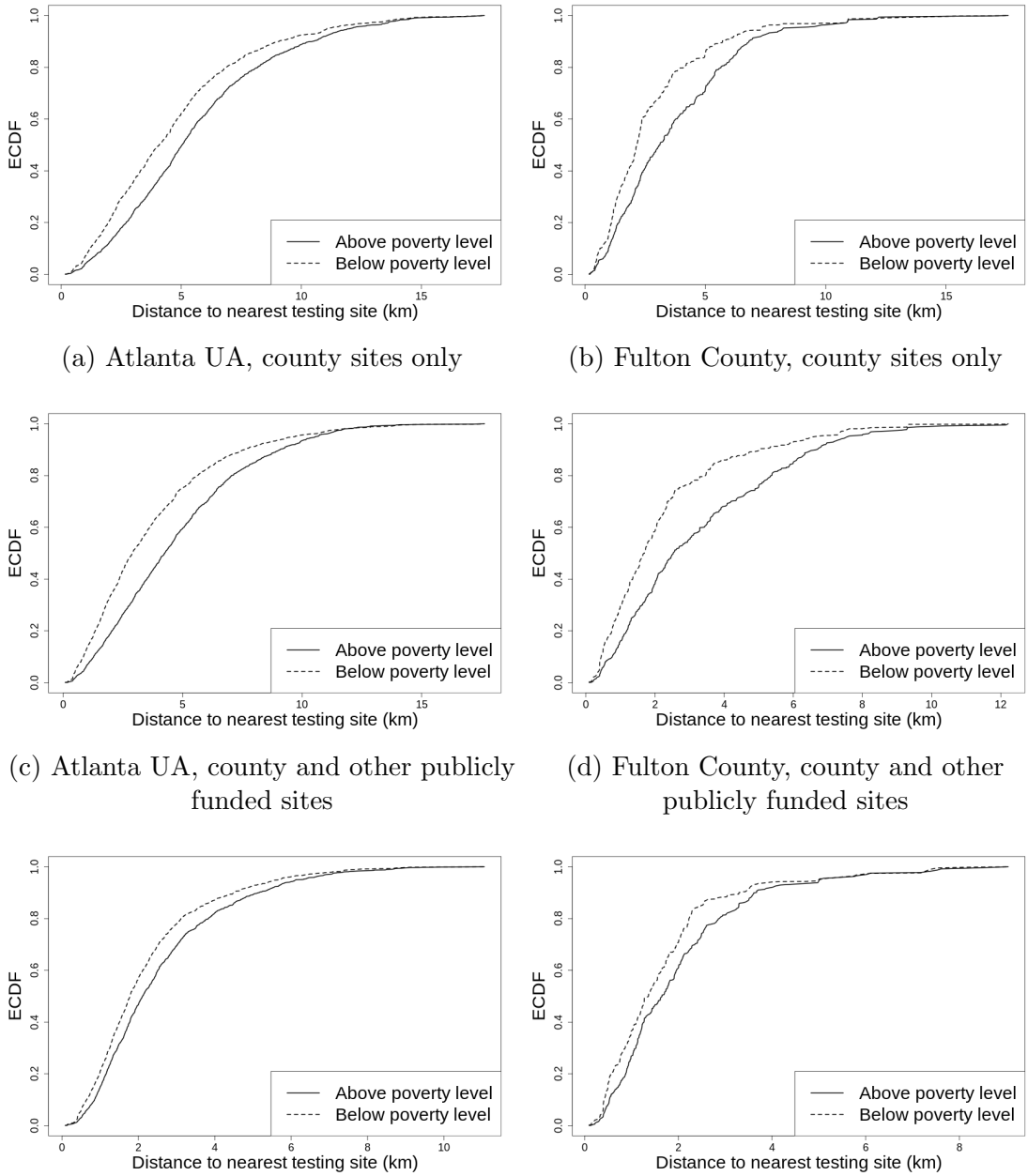
		Median Validation MSE		Median OOB MSE		Number of days where smaller p has better MSE	
mtry	**Radius**	**Larger p**	**Smaller p**	**Larger p**	**Smaller p**	**Validation**	**OOB**
4	100	0.33	0.33	0.06	0.11	14	0
12	100	0.35	0.33	0.05	0.10	23	0
4	250	0.71	0.72	0.06	0.11	15	0
12	250	0.82	0.71	0.05	0.10	20	0

Table B.6: Comparison between including additional nearest-neighbor features (larger p) vs. not (smaller p) across 2 m (`mtry`) values and 2 validation radius values across 31 days in July 2011.

Appendix C

Supplemental Materials to "A framework for assessing COVID-19 testing site spatial access"

C.1 Additional figures



(a) Atlanta UA, county sites only

(b) Fulton County, county sites only

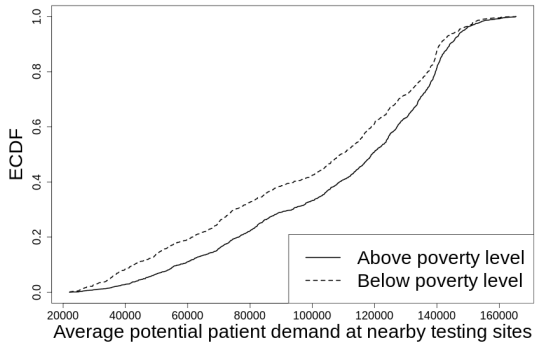
(c) Atlanta UA, county and other publicly funded sites

(d) Fulton County, county and other publicly funded sites

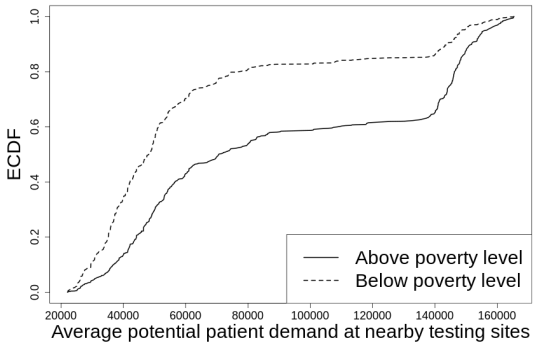
(e) Atlanta UA, all testing sites

(f) Fulton County, all testing sites

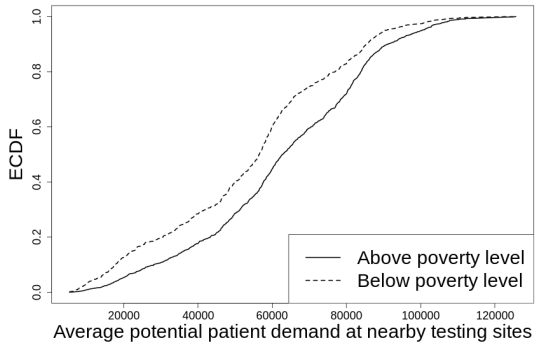
Figure C.1: ECDF comparisons for distance to nearest testing site among persons living below and above the poverty level in the Atlanta UA (left column) and Fulton County (right column), for public sites (top row), public sites together with other community and HRSA health center sites (middle row), and all public and private testing sites together (bottom row).



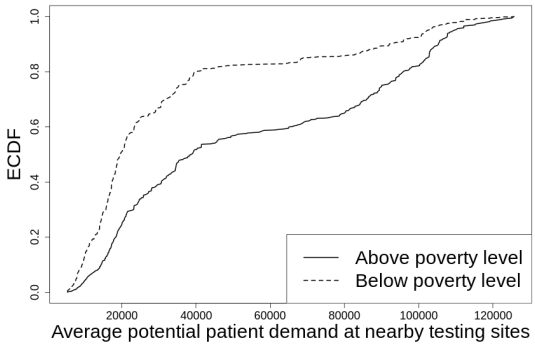
(a) Atlanta UA, county sites only



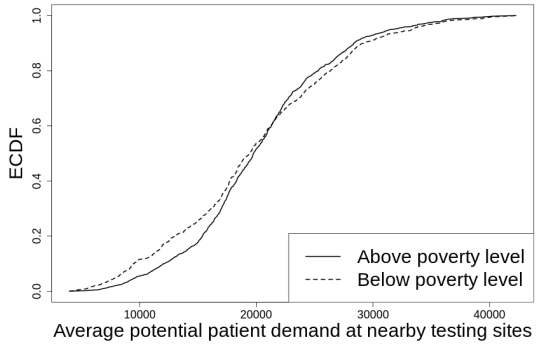
(b) Fulton County, county sites only



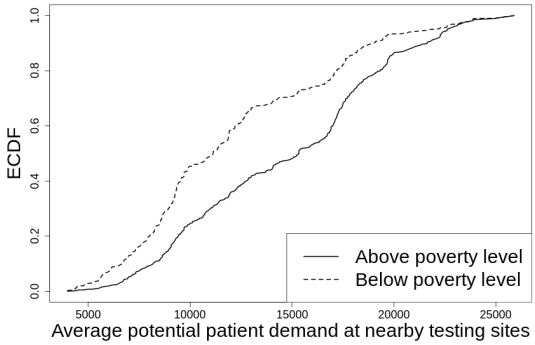
(c) Atlanta UA, county and other publicly funded sites



(d) Fulton County, county and other publicly funded sites

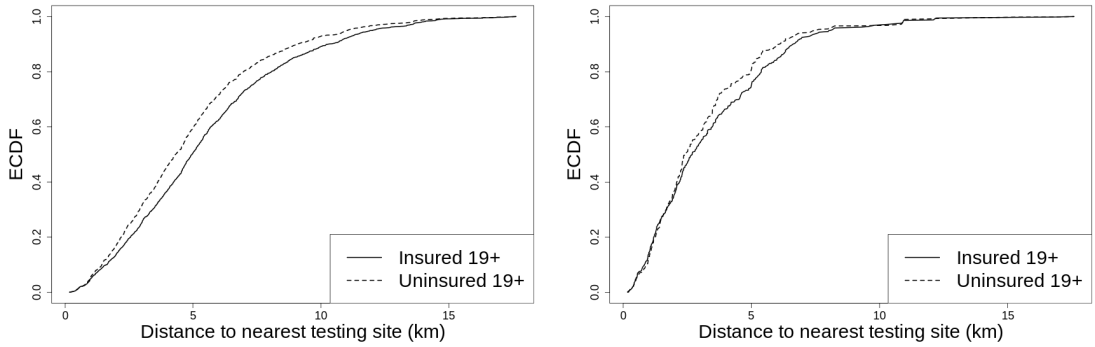


(e) Atlanta UA, all testing sites



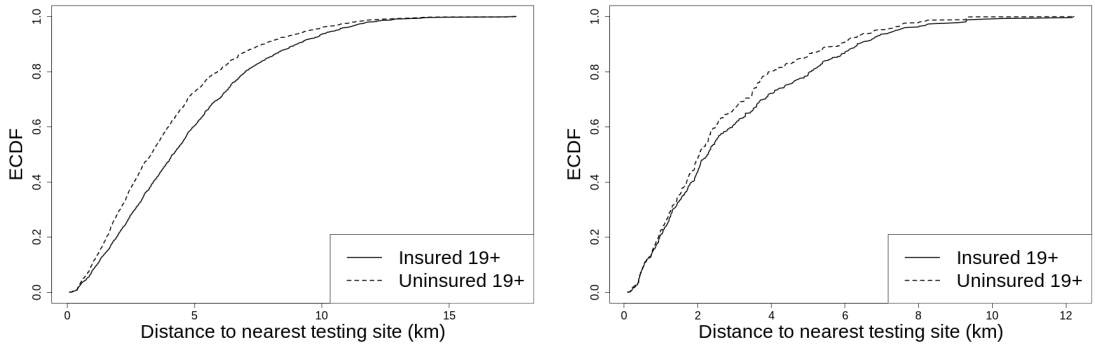
(f) Fulton County, all testing sites

Figure C.2: ECDF comparisons for potential demand at nearby testing sites among persons living below and above the poverty level in the Atlanta UA (left column) and Fulton County (right column), for public sites (top row), public sites together with other community and HRSA health center sites (middle row), and all public and private testing sites together (bottom row).



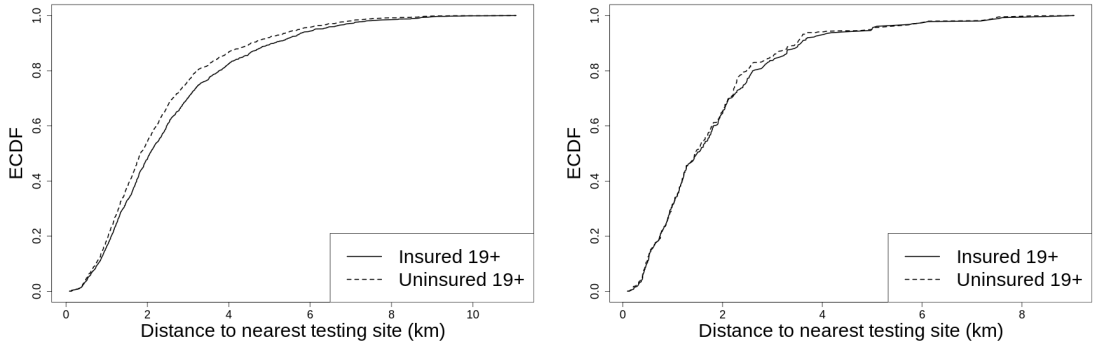
(a) Atlanta UA, county sites only

(b) Fulton County, county sites only



(c) Atlanta UA, county and other publicly funded sites

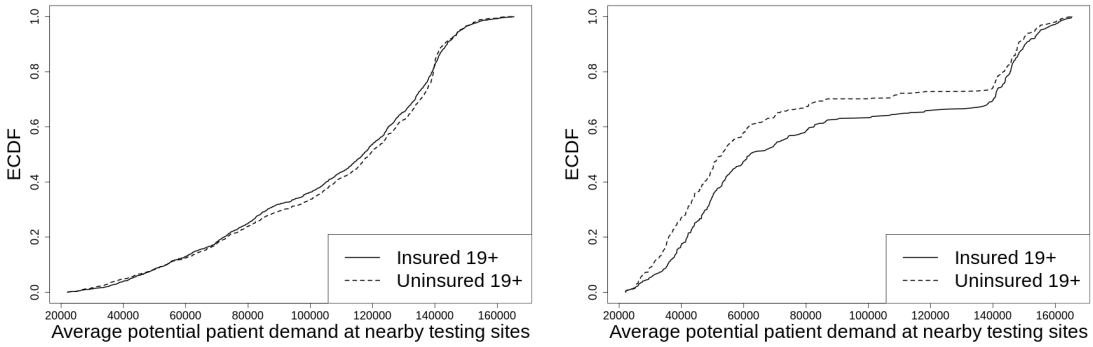
(d) Fulton County, county and other publicly funded sites



(e) Atlanta UA, all testing sites

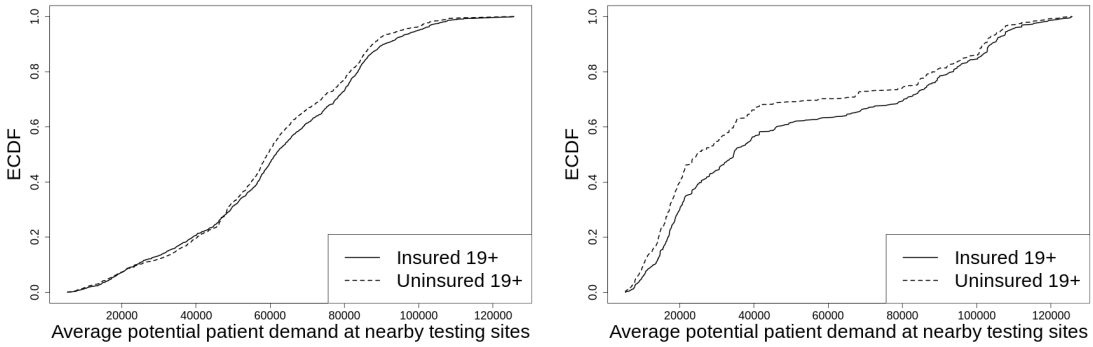
(f) Fulton County, all testing sites

Figure C.3: ECDF comparisons for distance to nearest testing site among uninsured and insured persons 19 and older in the Atlanta UA (left column) and Fulton County (right column), for public sites (top row), public sites together with other community and HRSA health center sites (middle row), and all public and private testing sites together (bottom row).



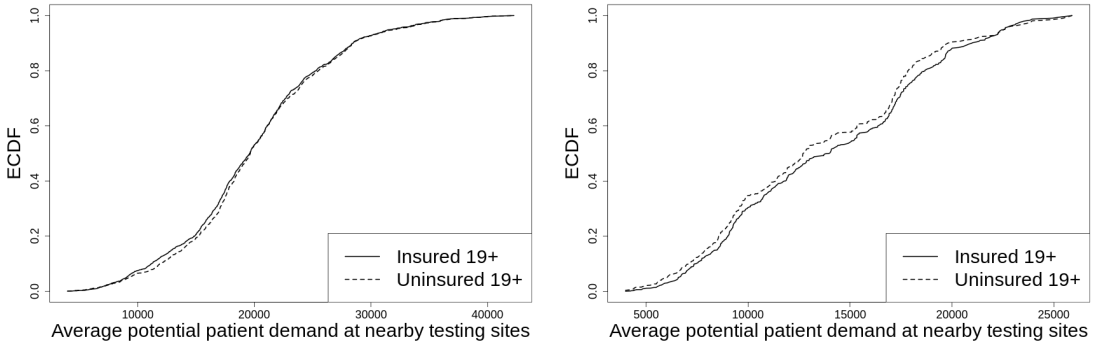
(a) Atlanta UA, county sites only

(b) Fulton County, county sites only



(c) Atlanta UA, county and other publicly funded sites

(d) Fulton County, county and other publicly funded sites



(e) Atlanta UA, all testing sites

(f) Fulton County, all testing sites

Figure C.4: ECDF comparisons for potential demand at nearby testing sites among uninsured and insured persons 19 and older in the Atlanta UA (left column) and Fulton County (right column), for public sites (top row), public sites together with other community and HRSA health center sites (middle row), and all public and private testing sites together (bottom row).

C.2 Additional information about testing sites

Below we provide archived links for county testing sites. Our target area was primarily the Atlanta urbanized area and Fulton County, but we cast a wide net around this area to ensure an appropriate accounting of public testing sites. These links were archived with the Internet Archive's Wayback machine. For some counties, we relied on the Georgia DPH testing map (`https://covid19.dph.ga.gov/en-US/test-location-map/`) to identify the operational testing sites. In some cases the location of county testing sites was ambiguous (e.g., Northwest Health District did not explicitly print the addresses of testing sites on their schedule) so there may be some error in testing site geocoding.

- Fulton County, including the CORE testing sites: Link 1 and Link 2
- Cobb/Douglas, including CORE testing sites
- DeKalb and CORE sites: Main Sites and CORE sites
- Gwinnett, Newton, and Rockdale: Main sites and CORE sites
- North Georgia Health District (1-2)
- Public Health District 2
- Northeast Health District
- Northwest Health District
- Public Health District 4
- North Central District

Bibliography

Philippe Apparicio, Jérémy Gelb, Anne-Sophie Dubé, Simon Kingham, Lise Gauvin, and Éric Robitaille. The approaches to measuring the potential spatial access to urban health services revisited: distance types and aggregation-error issues. *International Journal of Health Geographics*, 16(1):1–24, 2017.

Bruno Arpino and Massimo Cannas. Propensity score matching with clustered data. an application to the estimation of the impact of caesarean section on the apgar score. *Statistics in medicine*, 35(12):2074–2091, 2016.

Bruno Arpino and Fabrizia Mealli. The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, 55(4): 1770–1780, 2011.

Peter C Austin. The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in medicine*, 26(16):3078–3094, 2007.

Peter C Austin. A data-generation process for data with specified risk differences or numbers needed to treat. *Communications in Statistics—Simulation and Computation®*, 39(3):563–577, 2010.

Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3): 399–424, 2011a.

Peter C Austin. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics*, 10(2):150–161, 2011b.

Peter C Austin. A comparison of 12 algorithms for matching on the propensity score. *Statistics in medicine*, 33(6):1057–1069, 2014.

Peter C Austin, Douglas S Lee, and Jason P Fine. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609, 2016.

Jessica H Belle, Howard H Chang, Yujie Wang, Xuefei Hu, Alexei Lyapustin, and Yang Liu. The potential impact of satellite-retrieved cloud parameters on ground-level PM2.5 mass and composition. *International journal of environmental research and public health*, 14(10):1244, 2017.

JH Belle and Yang Liu. Evaluation of Aqua MODIS collection 6 AOD parameters for air quality research over the continental United States. *Remote Sensing*, 8(10): 815, 2016.

Isabelle Bey, Daniel J Jacob, Robert M Yantosca, Jennifer A Logan, Brendan D Field, Arlene M Fiore, Qinbin Li, Honguy Y Liu, Loretta J Mickley, and Martin G Schultz. Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. *Journal of Geophysical Research: Atmospheres*, 106(D19):23073–23095, 2001.

Jianzhao Bi, Jessica H Belle, Yujie Wang, Alexei I Lyapustin, Avani Wildani, and Yang Liu. Impacts of snow and cloud covers on satellite-derived PM2.5 levels. *Remote sensing of environment*, 221:665–674, 2019.

Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

Jonathan R Bradley, Noel Cressie, Tao Shi, et al. A comparison of spatial predictors when datasets could be very large. *Statistics Surveys*, 10:100–131, 2016.

Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Laura Bronner. How we analyzed 7,914 COVID-19 testing sites and found racial disparities. *FiveThirtyEight*, 2020. Accessed August 8, 2020; see `https://web.archive.org/web/20200821221816/https%3A%2F%2Ffivethirtyeight.com%2Ffeatures%2Fhow-we-analyzed-7914-covid-19-testing-sites-and-found-racial-disparities%2F`.

Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72, 2008.

Zhao-Yue Chen, Tian-Hao Zhang, Rong Zhang, Zhong-Min Zhu, Jun Yang, Ping-Yan Chen, Chun-Quan Ou, and Yuming Guo. Extreme gradient boosting model to estimate PM2.5 concentrations with missing-filled satellite data in China. *Atmospheric environment*, 202:180–189, 2019.

P Chernyavskiy, GM Kendall, R Wakeford, and MP Little. Spatial prediction of naturally occurring gamma radiation in Great Britain. *Journal of environmental radioactivity*, 164:300–311, 2016.

Sundar A Christopher and Pawan Gupta. Satellite remote sensing of particulate matter air quality: The cloud-cover problem. *Journal of the air & waste management association*, 60(5):596–602, 2010.

Yuanyuan Chu, Yisi Liu, Xiangyu Li, Zhiyong Liu, Hanson Lu, Yuanan Lu, Zongfu Mao, Xi Chen, Na Li, Meng Ren, et al. A review on predicting ground PM2.5 concentration using satellite aerosol optical depth. *Atmosphere*, 7(10):129, 2016.

Brian A Cosgrove, Dag Lohmann, Kenneth E Mitchell, Paul R Houser, Eric F Wood, John C Schaake, Alan Robock, Curtis Marshall, Justin Sheffield, Qingyun Duan, et al. Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *Journal of Geophysical Research: Atmospheres*, 108(D22), 2003.

Noel Cressie, Tao Shi, and Emily L Kang. Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics*, 19(3):724–745, 2010.

Molly Margaret Davies and Mark J Van Der Laan. Optimal spatial prediction using ensemble machine learning. *The international journal of biostatistics*, 12(1):179–201, 2016.

Melanie Davis. *Addressing Geographic Confounding through Spatial Propensity Score Analysis for Hierarchical Data*. PhD thesis, Medical University of South Carolina, 2018.

Melanie L Davis, Brian Neelon, Paul J Nietert, Kelly J Hunt, Lane F. Burgette, Andrew B Lawson, and Leonard E Egede. Addressing geographic confounding through spatial propensity scores: a study of racial disparities in diabetes. *Statistical Methods in Medical Research*, 28(3):734–748, 2019. doi: 10.1177/0962280217735700. URL `https://doi.org/10.1177/0962280217735700`. PMID: 29145767.

Qian Di, Heresh Amini, Liuhua Shi, Itai Kloog, Rachel Silvern, James Kelly, M Benjamin Sabath, Christine Choirat, Petros Koutrakis, Alexei Lyapustin, et al. An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment international*, 130:104909, 2019.

Paul W. Eggers. CMS 2728: What good is it? *Clinical Journal of the American Society of Nephrology*, 5(11):1908–1909, 2010. ISSN 1555-9041. doi: 10.2215/CJN.08170910. URL `https://cjasn.asnjournals.org/content/5/11/1908`.

Guannan Geng, Nancy L Murray, Daniel Tong, Joshua S Fu, Xuefei Hu, Pius Lee, Xia Meng, Howard H Chang, and Yang Liu. Satellite-based daily PM2.5 estimates during fire seasons in Colorado. *Journal of Geophysical Research: Atmospheres*, 123(15):8159–8171, 2018.

Daniel L Goldberg, Pawan Gupta, Kai Wang, Chinmay Jena, Yang Zhang, Zifeng Lu, and David G Streets. Using gap-filled MAIAC AOD and WRF-Chem to estimate daily PM2.5 concentrations at 1 km resolution in the eastern United States. *Atmospheric Environment*, 199:443–452, 2019.

Neal S Grantham, Brian J Reich, Yang Liu, and Howard H Chang. Spatial regression with an informatively missing covariate: Application to mapping fine particulate matter. *Environmetrics*, 29(4):e2499, 2018.

Mark F Guagliardo. Spatial accessibility of primary care: concepts, methods and challenges. *International journal of health geographics*, 3(1):3, 2004.

Hua Hao, Brendan P Lovasik, Stephen O Pastan, Howard H Chang, Ritam Chowdhury, and Rachel E Patzer. Geographic variation and neighborhood factors are associated with low rates of pre–end-stage renal disease nephrology care. *Kidney international*, 88(3):614–621, 2015.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Matthew J Heaton, Abhirup Datta, Andrew O Finley, Reinhard Furrer, Joseph Guinness, Rajarshi Guhaniyogi, Florian Gerber, Robert B Gramacy, Dorit Hammerling, Matthias Katzfuss, et al. A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425, 2019.

James J Heckman, Hidehiko Ichimura, and Petra E Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654, 1997.

Tomislav Hengl, Madlene Nussbaum, Marvin N Wright, Gerard BM Heuvelink, and Benedikt Gräler. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6:e5518, 2018.

Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236, 2007.

Sari Hopson, Diane Frankenfield, Michael Rocco, and William McClellan. Variability in reasons for hemodialysis catheter use by race, sex, and geography: findings from the ESRD Clinical Performance Measures Project. *American Journal of Kidney Diseases*, 52(4):753–760, 2008.

Xuefei Hu, Jessica H Belle, Xia Meng, Avani Wildani, Lance A Waller, Matthew J Strickland, and Yang Liu. Estimating PM2.5 concentrations in the conterminous United States using the random forest approach. *Environmental science & technology*, 51(12):6936–6944, 2017.

Keyong Huang, Qingyang Xiao, Xia Meng, Guannan Geng, Yujie Wang, Alexei Lyapustin, Dongfeng Gu, and Yang Liu. Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain. *Environmental pollution*, 242:675–683, 2018.

Keyong Huang, Jianzhao Bi, Xia Meng, Guannan Geng, Alexei Lyapustin, Kevin J Lane, Dongfeng Gu, Patrick L Kinney, and Yang Liu. Estimating daily PM2.5 concentrations in New York City at the neighborhood-scale: Implications for in-

tegrating non-regulatory measurements. *Science of The Total Environment*, 697: 134094, 2019.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

Kirsten L Johansen, Rebecca Zhang, Yijian Huang, Rachel E Patzer, and Nancy G Kutner. Association of race and insurance type with delayed assessment for kidney transplantation among patients initiating dialysis in the United States. *Clinical Journal of the American Society of Nephrology*, 7(9):1490–1497, 2012.

Allan C Just, Margherita M De Carli, Alexandra Shtein, Michael Dorman, Alexei Lyapustin, and Itai Kloog. Correcting measurement error in satellite aerosol optical depth with machine learning for modeling PM2.5 in the northeastern USA. *Remote sensing*, 10(5):803, 2018.

Luke Keele and Rocío Titiunik. Geographic natural experiments with interference: The effect of all-mail voting on turnout in Colorado. *CESifo Economic Studies*, 64 (2):127–149, 2018.

Soo Rin Kim, Matthew Vann, Laura Bronner, and Grace Manthey. Which cities have the biggest racial gaps in COVID-19 testing access? *FiveThirtyEight*, 2020. Accessed August 8, 2020; see `https://web.archive.org/web/20200821222141/https%3A%2F%2Ffivethirtyeight.com%2Ffeatures%2Fwhite-neighborhoods-have-more-access-to-covid-19-testing-sites%2F`.

Itai Kloog, Petros Koutrakis, Brent A Coull, Hyung Joo Lee, and Joel Schwartz. Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmospheric environment*, 45(35):6267–6275, 2011.

Itai Kloog, Francesco Nordio, Brent A Coull, and Joel Schwartz. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM2.5 exposures in the Mid-Atlantic states. *Environmental science & technology*, 46(21):11913–11921, 2012.

Geoffrey M Laslett. Kriging and splines: an empirical comparison of their predictive performance in some applications. *Journal of the American Statistical Association*, 89(426):391–400, 1994.

Mihye Lee, Itai Kloog, Alexandra Chudnovsky, Alexei Lyapustin, Yujie Wang, Steven Melly, Brent Coull, Petros Koutrakis, and Joel Schwartz. Spatiotemporal prediction of fine particulate matter using high-resolution satellite images in the Southeastern US 2003–2011. *Journal of exposure science & environmental epidemiology*, 26(4): 377–384, 2016.

Jos Lelieveld, John S Evans, Mohammed Fnais, Despina Giannadaki, and Andrea Pozzer. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525(7569):367–371, 2015.

R Levy, C Hsu, et al. MODIS Atmosphere L2 Aerosol Product. NASA MODIS Adaptive Processing System. *Goddard Space Flight Center, USA*, 10, 2015. doi: 10.5067/MODIS/MOD04_L2.006. URL `http://dx.doi.org/10.5067/MODIS/MOD04_L2.006`.

RC Levy, S Mattoo, LA Munchak, LA Remer, AM Sayer, F Patadia, and NC Hsu. The Collection 6 MODIS aerosol products over land and ocean. *Atmospheric Measurement Techniques*, 6(11):2989, 2013.

Fan Li, Alan M Zaslavsky, and Mary Beth Landrum. Propensity score weighting with multilevel data. *Statistics in medicine*, 32(19):3373–3387, 2013a.

Shenshen Li, Michael J Garay, Liangfu Chen, Erika Rees, and Yang Liu. Comparison of GEOS-Chem aerosol optical depth with AERONET and MISR data over the contiguous United States. *Journal of Geophysical Research: Atmospheres*, 118(19): 11–228, 2013b.

Fengchao Liang, Qingyang Xiao, Keyong Huang, Xueli Yang, Fangchao Liu, Jianxin Li, Xiangfeng Lu, Yang Liu, and Dongfeng Gu. The 17-y spatiotemporal trend of PM2.5 and its mortality burden in China. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1919641117. URL `https://www.pnas.org/content/early/2020/09/15/1919641117`.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

Stephen S Lim, Theo Vos, Abraham D Flaxman, Goodarz Danaei, Kenji Shibuya, Heather Adair-Rohani, Mohammad A AlMazroa, Markus Amann, H Ross Anderson, Kathryn G Andrews, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010. *The lancet*, 380(9859):2224–2260, 2012.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Wei Luo and Yi Qi. An enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility to primary care physicians. *Health & place*, 15 (4):1100–1107, 2009.

Baolei Lv, Yongtao Hu, Howard H Chang, Armistead G Russell, and Yuqi Bai. Improving the accuracy of daily PM2.5 distributions derived from the fusion of ground-

level measurements with aerosol optical depth observations, a case study in North China. *Environmental science & technology*, 50(9):4752–4759, 2016.

Steven Manson, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 15.0 [Database], 2020. Minneapolis, MN: IPUMS. http://doi.org/10.18128/D050.V15.0.

William M McClellan, Haimanot Wasse, Ann C McClellan, Adam Kipp, Lance A Waller, and Michael V Rocco. Treatment center and geographic variability in pre-ESRD care associate with increased mortality. *Journal of the American Society of Nephrology*, 20(5):1078–1085, 2009.

William M McClellan, Haimanot Wasse, Ann C McClellan, James Holt, Jenna Krisher, and Lance A Waller. Geographic concentration of poverty and arteriovenous fistula use among ESRD patients. *Journal of the American Society of Nephrology*, 21(10):1776–1782, 2010.

Robert W. Mee and Dan Anbar. Confidence bounds for the difference between two probabilities. *Biometrics*, 40(4):1175–1176, 1984. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2531174.

Olli Miettinen and Markku Nurminen. Comparative analysis of two rates. *Statistics in medicine*, 4(2):213–226, 1985.

Kenneth E Mitchell, Dag Lohmann, Paul R Houser, Eric F Wood, John C Schaake, Alan Robock, Brian A Cosgrove, Justin Sheffield, Qingyun Duan, Lifeng Luo, et al. The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research: Atmospheres*, 109(D7), 2004.

Nancy L Murray, Heather A Holmes, Yang Liu, and Howard H Chang. A Bayesian ensemble approach to combine PM2.5 estimates from statistical models using satellite imagery and numerical model simulation. *Environmental research*, 178:108601, 2019.

Ashley I Naimi and Laura B Balzer. Stacked generalization: an introduction to super learning. *European journal of epidemiology*, 33(5):459–464, 2018.

Douglas Nychka, Soutir Bandyopadhyay, Dorit Hammerling, Finn Lindgren, and Stephan Sain. A multiresolution Gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24(2):579–599, 2015.

Douglas Nychka, Dorit Hammerling, Stephan Sain, and Nathan Lenssen. Latticekrig: Multiresolution kriging based on markov random fields, 2016. URL `https://github.com/NCAR/LatticeKrig`. R package version 8.4.

Georgia Papadogeorgou, Christine Choirat, and Corwin M Zigler. Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching. *Biostatistics*, 20(2):256–272, 2019a.

Georgia Papadogeorgou, Fabrizia Mealli, and Corwin M Zigler. Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics*, 75(3):778–787, 2019b.

Rachel E Patzer, Jennifer Gander, Leighann Sauls, M Ahinee Amamoo, Jenna Krisher, Laura L Mulloy, Eric Gibney, Teri Browne, Laura Plantinga, and Stephen O Pastan. The RaDIANT community study protocol: community-based participatory research for reducing disparities in access to kidney transplantation. *BMC nephrology*, 15(1):171, 2014.

Rachel E Patzer, Laura C Plantinga, Sudeshna Paul, Jennifer Gander, Jenna Krisher, Leighann Sauls, Eric M Gibney, Laura Mulloy, and Stephen O Pastan. Variation in dialysis facility referral for kidney transplantation among patients with end-stage renal disease in Georgia. *JAMA*, 314(6):582–594, 2015.

Rachel E Patzer, Sudeshna Paul, Laura Plantinga, Jennifer Gander, Leighann Sauls, Jenna Krisher, Laura L Mulloy, Eric M Gibney, Teri Browne, Carlos F Zayas, et al. A randomized trial to reduce disparities in referral for transplant evaluation. *Journal of the American Society of Nephrology*, 28(3):935–942, 2017.

Eric C Polley and Mark J Van der Laan. Super learner in prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 266.*, 2010. URL `https://biostats.bepress.com/ucbbiostat/paper266`.

Philipp Probst, Marvin N Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1301, 2019.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Jordan H Rickles and Michael Seltzer. A two-stage propensity score matching strategy for treatment effect estimation in a multisite observational study. *Journal of Educational and Behavioral Statistics*, 39(6):612–636, 2014.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371): 591–593, 1980.

Megan L Salter, Babak Orandi, Mara A McAdams-DeMarco, Andrew Law, Lucy A Meoni, Bernard G Jaar, Stephen M Sozio, Wen Hong Linda Kao, Rulan S Parekh, and Dorry L Segev. Patient-and provider-reported information about transplantation and subsequent waitlisting. *Journal of the American Society of Nephrology*, 25 (12):2871–2877, 2014.

Ron Sarafian, Itai Kloog, Allan C Just, and Johnathan D Rosenblatt. Gaussian markov random fields versus linear mixed models for satellite-based PM2.5 assessment: Evidence from the northeastern USA. *Atmospheric environment*, 205:30–35, 2019.

Ralph Scherer. *PropCIs: Various Confidence Interval Methods for Proportions*, 2018. URL `https://CRAN.R-project.org/package=PropCIs`. R package version 0.3-0.

Mark R Segal. Machine learning benchmarks and random forest regression. *UCSF: Center for Bioinformatics and Molecular Biostatistics*, 2004. URL `https://escholarship.org/uc/item/35x3v9t4`.

Yanchuan Shao, Zongwei Ma, Jianghao Wang, and Jun Bi. Estimating daily ground-level PM2.5 in China with random-forest-based spatiotemporal kriging. *Science of The Total Environment*, 740:139761, 2020.

Minso Shin, Yoojin Kang, Seohui Park, Jungho Im, Cheolhee Yoo, and Lindi J Quackenbush. Estimating ground-level particulate matter concentrations using satellite-based data: a review. *GIScience & Remote Sensing*, 57(2):174–189, 2020.

M Sorek-Hamer, AC Just, and I Kloog. Satellite remote sensing in epidemiological studies. *Current opinion in pediatrics*, 28(2):228, 2016.

Massimo Stafoggia, Tom Bellander, Simone Bucci, Marina Davoli, Kees De Hoogh, Francesca De'Donato, Claudio Gariazzo, Alexei Lyapustin, Paola Michelozzi, Matteo Renzi, et al. Estimation of daily PM10 and PM2.5 concentrations in Italy,

2013–2015, using a spatiotemporal land-use random-forest model. *Environment international*, 124:170–179, 2019.

Jessica L Heier Stamm, Nicoleta Serban, Julie Swann, and Pascale Wortley. Quantifying and explaining accessibility with application to the 2009 H1N1 vaccination campaign. *Health care management science*, 20(1):76–93, 2017.

Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

Elizabeth A Stuart and Donald B Rubin. Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics*, 33(3):279–306, 2008.

United States Renal Data System. 2017 USRDS annual data report: Epidemiology of kidney disease in the United States, 2017. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD.

URISA's GISCorps, Coders Against COVID (findcovidtesting.com), and Esri. COVID-19 testing locations in the United States. Spatial dataset., 2020. Accessed September 29, 2020; `https://covid-19-giscorps.hub.arcgis.com/datasets/giscorps-covid-19-testing-locations-in-the-united-states-symbolized-by-test-type/data`.

Roozbeh Valavi, Jane Elith, José J. Lahoz-Monfort, and Gurutzeta Guillera-Arroita. blockcv: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2):225–232, 2019.

Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

Natalya Verbitsky-Savitz and Stephen W Raudenbush. Causal inference under interference in spatial settings: A case study evaluating community policing program in chicago. *Epidemiologic Methods*, 1(1):107–130, 2012.

Lance A Waller, Thomas A Louis, and Bradley P Carlin. Bayes methods for combining disease and exposure data in assessing environmental justice. *Environmental and Ecological Statistics*, 4(4):267–281, 1997.

Lance A Waller, Thomas A Louis, and Bradley P Carlin. Environmental justice and statistical summaries of differences in exposure distributions. *Journal of Exposure Science & Environmental Epidemiology*, 9(1):56–65, 1999.

Holger Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematics*, 4(1): 389–396, 1995.

WHO. Ambient (outdoor) air pollution, 2018. URL `https://web.archive.org/web/20200824220508/https%3A%2F%2Fwww.who.int%2Fnews-room%2Ffact-sheets%2Fdetail%2Fambient-%2528outdoor%2529-air-quality-and-health`. Accessed on 24 August 2020.

Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(i01), 2017.

Qingyang Xiao, Yujie Wang, Howard H Chang, Xia Meng, Guannan Geng, Alexei Lyapustin, and Yang Liu. Full-coverage high-resolution daily PM2.5 estimation using MAIAC AOD in the Yangtze River Delta of China. *Remote Sensing of Environment*, 199:437–446, 2017.

Qingyang Xiao, Howard H Chang, Guannan Geng, and Yang Liu. An ensemble

machine-learning model to predict historical PM2.5 concentrations in China from satellite data. *Environmental science & technology*, 52(22):13260–13269, 2018.

Xin Xu, Sheng Nie, Hanying Ding, and Fan Fan Hou. Environmental pollution and kidney diseases. *Nature Reviews Nephrology*, 2018.

Guofen Yan, Alfred K Cheung, Jennie Z Ma, J Yu Alison, Tom Greene, M Norman Oliver, Wei Yu, and Keith C Norris. The associations between race and geographic area and quality-of-care indicators in patients approaching ESRD. *Clinical Journal of the American Society of Nephrology*, 8(4):610–618, 2013.

Michael T Young, Matthew J Bechle, Paul D Sampson, Adam A Szpiro, Julian D Marshall, Lianne Sheppard, and Joel D Kaufman. 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, 2016.

Haozhe Zhang, Joshua Zimmerman, Dan Nettleton, and Daniel J Nordman. Random forest prediction intervals. *The American Statistician*, pages 1–15, 2019.

Ruixin Zhang, Baofeng Di, Yuzhou Luo, Xunfei Deng, Michael L Grieneisen, Zhigao Wang, Gang Yao, and Yu Zhan. A nonparametric approach to filling gaps in satellite-retrieved aerosol optical depth for estimating ambient PM2.5 levels. *Environmental Pollution*, 243:998–1007, 2018.

Corwin M Zigler, Francesca Dominici, and Yun Wang. Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes. *Biostatistics*, 13(2):289–302, 2012.

José R Zubizarreta, Mark Neuman, Jeffrey H Silber, and Paul R Rosenbaum. Contrasting evidence within and between institutions that provide treatment in an observational study of alternate forms of anesthesia. *Journal of the American Statistical Association*, 107(499):901–915, 2012.