TOBLER MEETS ROBINSON: SEMIPARAMETRIC METHODS FOR COVARIATE ADJUSTMENT USING SPATIAL DATA

APOORVA LAL

ABSTRACT. Social scientists frequently study causal questions with georeferenced data, wherein observations correspond to areal units or coordinates. Empirical researchers frequently incorporate geographic information like any other covariate and risk producing biased estimates by failing to flexibly incorporate the spatial dependence structure in unobserved confounders. Motivated by the implication from Tobler's first law of geography that units that are close together in space are more comparable to each other on unobserved confounders, we propose a set of sufficient identification conditions such that unobserved spatial confounders can be partialled out by conditioning flexibly on geographic location. This allows us to repurpose well-known estimators in the unconfoundedness literature that involve fitting the outcome and propensity models using flexible nonparametric regressions of the outcome and treatment on smooth functions of location and plugging them into a doubly-robust score function to target the Average Treatment Effect (ATE) or regressing residuals on residuals for an overlap-weighted Treatment Effect (OTE). We find that this semiparametric covariate adjustment approach outperforms conventional covariate adjustment strategies and yields unbiased and precise estimates in a wide variety of spatial confounding simulations, and conclude with two empirical applications.

STANFORD UNIVERSITY, CA

E-mail address: apoorval@stanford.edu.

Date: May 22, 2023.

1. Introduction

Social scientists frequently attempt to study causal effects with geo-referenced data, wherein observations either correspond to areal units, such as counties or districts, or spatial point locations, such as residences, or event coordinates. In many such settings, the assignment mechanism is unknown, but researchers have a loose intuition that closer units are more comparable to each other along unobserved confounders, memorably articulated by Tobler as "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Accordingly, researchers frequently invoke unconfoundedness 'conditional on geography', thereby yielding a 'Geographic Natural Experiment'(Keele and Titiunik, 2016). Current practice in applied social science work treats location as just another covariate, and therefore incorporates it highly parametrically (typically as a regressor in a linear regression), or by restricting comparisons to within larger administrative units by incorporating fixed-effects.

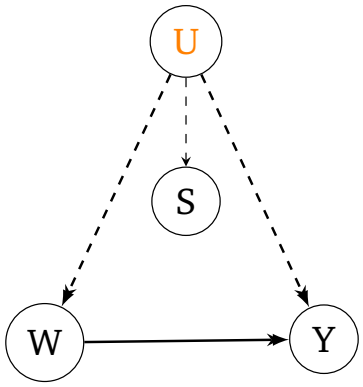
This paper proposes sufficient identification conditions such that unobserved spatial confounders can be partialled out by conditioning flexibly on geographic location in both a partially linear or fully nonparametric potential outcomes framework. Next, we re-purpose well-known estimators in the unconfoundedness literature that involve fitting the outcome and propensity models using flexible nonparametric regressions of the outcome and treatment on smooth functions of location, followed by regressing residuals outcomes on residual treatment in a partially linear model for an overlap-weighted Treatment Effect (OTE) (Robinson, 1988), or plugging them into a doubly-robust score function to target the Average Treatment Effect (ATE) (J. M. Robins, Rotnitzky, and Zhao, 1994; Chernozhukov, Chetverikov, Demirer, et al., 2018). This approach relies on a perturbation of the standard confounding directed acyclic graph (DAG) in fig 1a, wherein instead of adjusting for an unobserved confounder U directly, the researcher adjusts for a proxy S, which is the unit's geographic location, which amounts to various forms of 'local differencing'. Spatial

smoothing is performed using nearest-neighbour smoothers[1], Gaussian Markov Random Field smoothers (Havard Rue and Held, 2005; Wood, 2006) for areal units, spline, kernel sieve, and regularized kriging estimators for point-reference units (Cressie and Johannesson, 2008; E. L. Kang and Cressie, 2011), but in-principle is adaptable to any nonparametric regression method. Inference can be performed using the Bayesian (weighted) bootstrap (D. B. Rubin, 1981; Mason and Newton, 1992; Belloni et al., 2017).

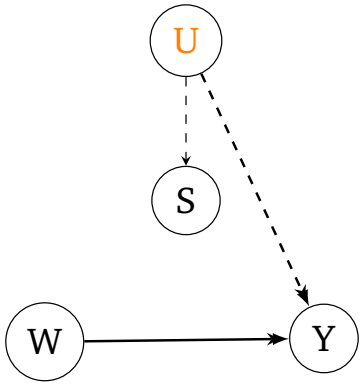
Finally, we conduct simulation studies with spatial confounding and varying degrees of smoothness and find that semiparametric adjustment yields substantially more unbiased and precise estimates relative to conventional strategies. In particular, we find that for smooth to moderate confounding, partialling out estimators yield unbiased estimates, while when confounding is very noisy, they yield biased estimates like all other strategies, but their bias is the smallest among all estimators considered.

The current paper contributes to linking modern causal inference with spatial statistics and spatial econometrics. The standard approach in spatial statistics and applied work public health is fit random-effects or kriging estimators that allow for spatial correlation in the random effects but are assumed to be orthogonal to the treatment by construction (Christensen, 2001; Cressie, 2015; Blangiardo and Cameletti, 2015), thereby typically assuming away confounders and potentially drawing incorrect inferences. This approach invokes figure 1b, where U only affects Y and therefore naive OLS estimates are inefficient but not biased. However, in most social scientific settings, this is an unrealistic assumption, since unobserved confounders are often correlated with both the treatment and the outcome (1a). In early work in this area, Paciorek (2010) analyses the importance of the scale of unobserved confounders versus the treatment on bias and precision of estimates, which can be interpreted as an overlap condition in causal methods. Recent work in statistics and biostatistics has made attempts to bridge the two literatures: Schnell and Papadogeorgou

[1]Also known as *Spatial Autoregressive Models* in the spatial econometrics literature,(H. H. Kelejian and Prucha, 2010)



(A) Unobserved spatial confounding



(B) No bias, spatial adjustment yields only precision gains

FIGURE 1. Directed-Acyclic Graphs of Spatial Confounding. Y is the outcome, W is the treatment, U is an unobserved confounder, and S is the location

(2020) connect the mixed-models approach and provide identification conditions for continuous exposures, while Gilbert, Datta, and Ogburn (2021) adapt double-machine learning for continuous treatments. Spatial econometric methods, on the other hand, explicitly focus on studying interference and decomposing effects into 'direct' and 'indirect' components (H. H. Kelejian and Prucha, 1999; H. H. Kelejian and Prucha, 2010), which is a separate and considerably more challenging exercise that has also been the focus of considerable recent interest (Sävje, Aronow, and Hudgens, 2021; Hu, S. Li, and Wager, 2021). Contemporary causal inference methods in econometrics and political methodology literature thus far treats location as a regular covariate as part of a selection-on-observables design Keele and Titiunik (2016) and Baum-Snow and Ferreira (2015), or focus on narrower assignment mechanisms, such geographic regression discontinuity (G. W. Imbens and Zajonc, 2009; Dell, 2010; Keele and Titiunik, 2015; G. W. Imbens and Wager, 2019)).

The rest of the paper is organised as follows: section 2 introduces the methodology, 3 reports results from a set of simulation studies benchmarking the semiparametric estimators against those commonly used in applied research, 4 illustrates the use of semiparametric estimators on two empirical examples, and section 5 concludes.

2. Model

The data $\mathcal{D} := \{\mathcal{D}_i\}_{i=1}^N := \{Y_i, W_i, S_i\}_{i=1}^N$ where each observation $\mathcal{D}_i$ is comprised of an outcome $Y_i \in \mathbb{R}$, a binary treatment $W_i \in \{0, 1\}$, and spatial location $S_i \in \mathcal{S}$. Location S_i is either *areal data*, such as districts, where units are located on an irregular lattice $\mathcal{S}$ (i.e. a map of administrative units), and *point data*, where each point possesses a location (latitude and longitude) $s \in \mathcal{S} \subseteq \mathbb{R}^2$. Each observation also has unobserved confounder $U_i \in \mathbb{R}$ associated with it, which is correlated with both the outcome Y_i and W_i, thereby making naive comparisons biased for the causal effect of W on Y. We may also have access to other covariates $\mathbf{V}_i$, but will suppress them for notational convenience and without loss of generality work with partialled out outcome and treatment[2]. The outcome is generated as $Y_i = W_i Y_i^1 - (1 - W_i) Y_i^0$, where Y^1, Y^0 correspond with potential outcomes under treatment and control respectively. This switching equation representation rules out cross-sectional interference wherein a unit's treatment status affects another unit's outcome.

2.1. Identification Assumptions.

The first two assumptions stipulate that the standard unconfoundedness and overlap assumptions hold conditional on U. The latter two assumptions stipulate that the unobserved confounder can be learned from the observed location data for areal and point-referenced data respectively. These are related to conditions set out for the continuous-exposure setting analysed by Schnell and Papadogeorgou (2020) and Gilbert, Datta, and Ogburn (2021).

[2]The precise nature of the partialling out can be arbitrarily flexible, and for some estimators additional covariates $\mathbf{V}_i$ can be included alongside location S_i in semiparametric estimators.

A1. Causal Consistency: $Y_i = Y_i(W_i)$. A unit's outcome is generated by its own treatment status alone.

A2. Latent Unconfoundedness: $Y_i(W_i) \perp\!\!\!\perp W_i | U_i$.

A3. Positivity: $\mathbf{Pr}\,(W = 1|U) \in (0, 1)$

The above assumptions assert the standard selection-on-observables assumptions of treatment ignorability and positivity *conditional on the unobserved confounder* U.

A4. Learnability of U: $U = g(S)$ for a fixed, measurable function $g(\cdot)$.

This is an smoothness assumption that requires that the confounder be 'nearly continuous' function of location. This rules out sharp jumps in the level of the unobserved confounder U such that a smoothing method used to estimate $g()$ will fail to pick it up. While this is untestable and therefore fails to guarantee that simply conditioning on location will eliminate all bias in all cases, is typically the case that conditioning on location flexibly even in cases where U is very noisy will yield less biased estimates than parametric or no adjustment (sec 3).

A5. Conditioning on S **doesn't induce confounding**: $Y_i(w_i) \perp\!\!\!\perp W_i | S_i, U_i$. This requires that additionally conditioning on location S does not induce confounding, i.e. location is not a collider.

Proposition 2.1 (Unconfoundedness given location).
Assumptions A4 and A5 imply

$$Y(w) \perp\!\!\!\perp W | S$$

Proof in appdx A.1. Informally, this characterises when conditioning on S as a proxy for the spatial confounder U suffices for unconfoundedness.

Finally, to proceed with estimation, we also need the amended overlap assumption

A6. Positivity: $\mathbf{Pr}\,(W = 1|S) \in (0, 1)$

This stipulates that the probability of treatment is non-zero everywhere in $\mathcal{S}$. If the estimand of interest is the Average Treatment effect on the Treated (ATT) $\mathbb{E}\left[Y^1 - Y^0|W = 1\right]$, this can be weakened to $\mathbf{Pr}\,(W = 1|S) < 1$.

Proposition 2.2 (Identification of Counterfactual mean).
Given assumptions 1:6, We can identify counterfactual mean $\mathbb{E}\left[Y^{(w)}\right]$ from the observed data.

$$\mathbb{E}\left[Y^{(w)}\right] = \mathbb{E}\left[Y|W = w, S = s\right] \mathsf{d}P(S)$$

Proof in appendix A.2. Proposition 2.2 allows us to proceed with the standard suite of estimators for selection on observables, where now the 'observable' is location, which is used as a proxy for the confounder. Since the independence statement $Y(w) \perp\!\!\!\perp W|S$ is a non-parametric identification result, this alone does not motivate a specific estimation strategy, and substantive knowledge motivates the choice of partially linear regression, matching, weighting, or hybrid approaches. We turn to estimation next.

2.2. **Partially Linear Regression Estimator.**

In the partially linear model (PLM), we stipulate that the outcome is generated as a flexible function of the unobserved confounder $m(U)$ plus an additive treatment effect.

$$Y_i = \tau W_i + m(U_i) + \varepsilon_i$$

where $Y^0 = m(U_i)$ and $Y^1 = Y^0 + W_i \cdot \tau$. Since U unobserved, the 'long' regression is infeasible, and therefore the regression coefficient τ is generically biased. The 'short' regression estimate of $\widehat{\tau}^s$ suffers from omitted variables bias $\mathbb{E}\left[\tau - \widehat{\tau}^s\right] = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbb{E}\left[U|\mathbf{X}\right]$, where $\mathbf{X}_i = (1, W_i)$. $\mathbb{E}\left[U|\mathbf{X}\right] \neq 0$ whenever the treatment correlated with unobserved confounders. This means standard approaches unconfoundedness approaches fail to yield consistent estimates of τ. However, since S_i is available, and we have established conditions for identifications in sec 2.1, we can write the following partially linear regression instead

$$Y_i = \tau W_i + g(S_i) + \varepsilon_i \tag{2.1}$$

Robinson (1988) establishes that equation 2.1 can equivalently be represented as

$$(Y_i - \underbrace{\mathbb{E}\left[Y_i|S_i\right]}_{m(s_i)}) = \tau \cdot (W_i - \underbrace{\mathbb{E}\left[W_i|S_i\right]}_{e(s_i)}) + \varepsilon_i \tag{2.2}$$

where the outcome model $m(s_i) := \mathbb{E}\left[Y_i|S_i\right]$ and treatment model $e(s_i) := \mathbb{E}\left[W_i|S_i\right]$ are estimated non-parametrically. Robinson (1988) shows that $\sqrt{n}-$ consistent estimation of τ is possible using the residuals on residuals regression 2.2.

Robinson advocates the use of kernel-regressions with higher-order kernels, which severely limited the practical application of this result. Higher order kernels use negative weights that complicate interpretation, its bandwidth is difficult to tune in realistic data settings struggles in the presence of high dimensional covariates (if s_i contains additional covariates). However, this approach has been rejuvenated by recent work on 'double machine learning'(Chernozhukov, Chetverikov, Demirer, et al., 2018), where the nonparametric regressions are fit with generic learners with sufficiently fast rate of convergence for the estimation errors in $m(\cdot)$ and $e(\cdot)$ cancel out. Popular machine learning regressions such as boosted trees, random forests, or tailored neural networks perform well in this setting

(Chernozhukov, Chetverikov, Demirer, et al., 2018). We outline three methods for fitting spatial non-parametric regressions below, but emphasise that any curve-fitting method that can be tailored for spatial predictions can be used in the first step. This then results in the following estimator

Defn 2.1 (Partially Linear Regression Coefficient).
The linear regression 2.2 of residuals on residuals estimates the following Partially linear regression coefficient

$$\widehat{\tau}^{\text{PLR}} = \frac{\sum_{i=1}^{n}(Y_i - \widehat{m}(s_i))(W_i - \widehat{e}(s_i))}{\sum_{i=1}^{n}(W_i - \widehat{e}(s_i))^2}$$

This is consistent for the constant treatment effect in regression 2.1. Proof in Chernozhukov, Chetverikov, Demirer, et al. (2018, Thm 4.1), which employs the corresponding Neyman-orthogonal score function $\psi(\mathcal{D}_i, \tau, \eta) = (Y - W\tau - m(S))(W - e(S))$ where $\eta = (m(\cdot), e(\cdot))$ collates the nuisance functions.

Under treatment effect heterogeneity, we can write the potential-outcome models $Y^0 = m(S_i)$ and $Y^1 = Y^0 + W\tau(S)$, which can be combined into the outcome model $Y_i = m(s_i) + W\tau(S) + \varepsilon_i$ where $\tau(S)$ is the treatment effect *function* as S varies. In the presence of such heterogeneity, it is well known that the partially linear regression coefficient $\widehat{\tau}^{\text{PLR}}$ isn't consistent for the average treatment effect $\mathbb{E}\left[Y^{(1)} - Y^{(0)}\right]$. Instead, the partially linear regression coefficient estimates a conditional-variance weighted average of strata-specific treatment effects (Angrist, 1998; Aronow and Samii, 2016), which may be substantially different from the ATE or ATT. However, 'moving the goalposts' to an overlap-weighted effect is often argued to be reasonable in the presence of potential overlap violations (Crump et al., 2009; J. Robins et al., 2008). Specifically, $\widehat{\tau}$ from the residuals on residuals regression is consistent for the a non-negative weighted average of the conditional average treatment effect function $\tau(S)$, which focuses on parts of the feature space with reasonable overlap, i.e. where $e(\cdot)$ is closest to 0.5.

Proposition 2.3 (Consistency of the Partially Linear Regression Coefficient Convex-weighted average of Conditional Average Treatment Effects).
$\widehat{\tau}^{\text{PLR}}$ is consistent for the following overlap-weighted estimand in the presence of effect heterogeneity

$$\widehat{\tau}^{\text{PLR}} \xrightarrow{p} \tau^{\text{ATO}} = \mathbb{E}\left[\frac{e(s_i)(1-e(s_i))}{\mathbb{E}\left[e(s_i)(1-e(s_i))\right]}\tau(S)\right]$$

Proof in A.3. This result validates the use of partially linear regression to estimate the treatment effect when effect heterogeneity is minimal, and a convex-weighted average of heterogeneous treatment effects when effects are heterogeneous (but overlap may be challenging for certain values of S). We revisit the latter point in the inference section 2.4.

Next, we outline two spatial regression/smoothing techniques to fit m and e that perform well with spatial data.

2.2.1. **Nearest-Neighbours Differencing.** The first approach is arguably the most straightforward, where location information is used solely to construct the adjacency matrix. This corresponds to the nonparametric regression estimator where the conditional expectation of X at location i is the average of the neighbouring units' value of X. Since location is relatively low-dimensional ($k = 2$ or $k = 3$) and is naturally ordered, constructing the nearest neighbours estimate / spatial lag relies on pre-multiplying by a weight matrix **W**, where

$$w_{ij} = \begin{cases} 1/\left|\mathcal{N}_i\right| & \text{if } \; j \in \mathcal{N}_i \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{N}_i$ is the set of i's neighbours. The normalisation by the number of neighbours ensures that the spatial lag is the average of all neighbouring units' values. Neighbours

may be defined based on substantive knowledge of the problem, but are typically defined to be 'queen' neighbours (where common edges or vertices qualify i to be j's neighbour and vice versa) (H. Kelejian and Piras, 2017).

This leads to the following outcome and propensity models for unit i

$$m(s_i) = \sum_{j \in \mathcal{N}_i} \omega_{ij} Y_j \;\; ; \qquad e(s_i) = \sum_{j \in \mathcal{N}_i} \omega_{ij} W_j$$

Letting $\widetilde{A}$ denote spatial lags for a random variable A, the partial linear regression 2.2 can now be written as

$$(Y_i - \widetilde{Y_i}) = \tau(W_i - \widetilde{W_i}) + \overbrace{U_i - \widetilde{U_i}}^{g(S) + g(S+\varepsilon) \to 0} + \eta_i$$

This approach consistently estimates τ as long as U is sufficiently smooth that local differencing eliminates it from the regression, thereby obviating the need to estimate $g(\cdot)$ at all. Intuitively, as long as the confounder is smooth in space, its value will be locally constant, and therefore its contribution can be differenced out by residualising on spatial lags. This is related to the claim in Paciorek (2010), where he studies the bias of covariate estimates and concludes that bias can be reduced by fitting a model when the spatial scale of the covariate is smaller than that of the confounder. This approach is closely related to random forests (in low/moderate dimensions), since the latter is an adaptive kernel/nearest-neighbours estimator (Athey, Tibshirani, and Wager, 2019).

The local differencing approach is closely related to the 'differencing' estimator of Yatchew (1997) adapted for spatial regressions by Druckenmiller and Hsiang (2018) as 'spatial first differences'. The differencing estimator requires regularly spaced lattices and requires

that one proceed in one direction (say, north to south) at a time, and average over directions. Yatchew shows that the differencing estimator is consistent and asymptotically normal as long as the derivative of the semiparametric part $g'(\cdot)$ is bounded, with variance $\mathcal{N}\left(\tau, 1.5\sigma_\eta^2/N\sigma_v^2\right)$, with σ_η^2 and σ_v^2 denoting the conditional variances of $Y|U$ and $W|U$ respectively. The convergence rate of the differencing estimator is $\frac{2}{3}$ that of the Robinson (1988) and Chernozhukov, Chetverikov, Demirer, et al. (2018) approach (which is $\sqrt{n}$ consistent), but on the other hand sidesteps the need to estimate the preliminary regressions $m(s_i), e(s_i)$.

With moderate-to-large quantities of data, the neighbourhood differencing is particularly appealing as it sidesteps the need to estimate nuisance functions and instead can be computed by simply differencing each observation's treatment and outcome value on their corresponding spatial lags and subsequently estimating residuals on residuals regression.

2.2.2. **Markov Random Field Models.** The use of Markov Random Fields models for spatial smoothing is motivated by a Bayesian prior that the truth is more likely to be smooth than wiggly $\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \|\mathbf{y} - f(\mathbf{x})\|_2^2 + \lambda J(\boldsymbol{\beta})$. In conventional spline problems, this implies a penalty term $J(\cdot)$ that penalises the second derivative $f''(\mathbf{x})$. However, since areal units are discrete and non-uniformly sized, the roughness penalty approach must be amended for unit effects.

To this end, the spatial statistics literature uses a correlated random effects approach where unit effects $\boldsymbol{\gamma}_j$ are assumed to be distributed as a Gaussian Markov Random Field (Havard Rue and Held (2005) and Wood (2006), defined in appdx A.4). Letting $\mathcal{N}_j$ denotes the set of neighbours of unit j, the penalty function can be written

$$J(\boldsymbol{\gamma}) = \sum_{j=1}^{n} \sum_{i \in \mathcal{N}_j, i>j} (\gamma_j - \gamma_i)^2$$

This ensures that unit effects for adjacent units are penalised to be close to each other. The degree of regularization λ dictates the sparsity of this approximation through its precision matrix. This approach can also be interpreted as a feasible implementation of Gaussian process regression via Vechhia approximations (Vecchia, 1988; Cressie and Davidson, 1998). MRFs can be fit using Regularised Maximum Likelihood or via Bayesian methods. In particular, the Integrated Nested Laplace Approximation (Lindgren, Håvard Rue, and Lindström, 2011) is tailored to fit MRFs and is widely used in the geostatistics literature to construct basis representations for large Gaussian process models (E. L. Kang and Cressie, 2011; Cressie and Johannesson, 2008).

To guard against over-fitting and guarantee $\sqrt{n}$ consistency, nuisance models $m(\cdot)$ and $e(\cdot)$ need to be guard against own-observation bias, i.e. the model used to predict a given unit's value of $\widehat{m}(s_i)$ should not be trained on observation i; this is typically performed using 'cross-fitting' by splitting the sample into K folds and predicting nuisance functions for fold k using models $\widehat{m}^{-k}(\cdot), \widehat{e}^{-k}(\cdot)$ omitting fold k. The approach in the present paper is to recognise that S is a *spatial* covariate, and use specialised nearest-neighbours smoothers to fit $\mathbb{E}[Y|S]$ and $\mathbb{E}[W|S]$.

2.3. **Fully Nonparametric Formulation.**

In the presence of treatment effect heterogeneity, the OLS coefficient τ from a partially linear regression uncovers a conditional-variance weighted average of strata-specific treatment effects (Angrist, 1998) where units with the highest conditional variance in treatment (i.e. propensity score $e(S) \approx 0.5$) receive the highest weight, while units with more extreme propensity scores close to 0 or 1 receive low weight. In observational studies, this is often justified as the average effect for observations for whom overlap might plausibly hold.

When effect heterogeneity is potentially high, however, one might want to use nonparametric estimators that target the ATE directly. Many estimators exist for estimation under

unconfoundedness including subclassification, weighting, matching, and regression imputation (see G. W. Imbens (2004) for a review). We outline three estimators with attractive properties in the spatial setting.

2.3.1. **Augmented Inverse-Propensity Weighting Estimator.** The Augmented inverse-propensity weighting (AIPW) estimator is the most popular among hybrid methods that combine modelling the two potential outcomes $m^0(s_i) := \mathbb{E}\left[Y \mid W_i = 1, s_i = s_i\right], m^1(s_i) := \mathbb{E}\left[Y|W_i = 0, s_i = s_i\right]$ and the propensity score $e(s_i) := \textbf{Pr}\left(W_i = 1|s_i = s_i\right)$, which possesses the desirable 'double-robustness' property that ensures consistency as long as either the outcome model or propensity score are correctly estimated. J. M. Robins, Rotnitzky, and Zhao (1994) proposed parametric models for nuisance functions, while recent work in double machine learning (Chernozhukov, Chetverikov, Demirer, et al., 2018) and targeted machine learning Van Der Laan and D. Rubin (2006) allow the use of flexible nonparametric estimators.

The augmented inverse-propensity weighting estimator (AIPW) is characterised by its efficient score (Hahn, 1998). The *uncentered* score for the average treatment effect (ATE) and average treatment effect on the treated (ATT) are

$$\xi_i^{\text{ATE}} = \underbrace{\widehat{m}^1(\mathbf{s}_i) + \frac{W_i(Y_i - \widehat{m}^1(\mathbf{s}_i))}{\widehat{e}(\mathbf{s}_i)}}_{\widehat{\mathbb{E}}[Y^{(1)}]} - \underbrace{\widehat{m}^0(\mathbf{s}_i) + \frac{(1-W_i)(Y_i - \widehat{m}^0(\mathbf{s}_i))}{1-\widehat{e}(\mathbf{s}_i)}}_{\widehat{\mathbb{E}}[Y^{(0)}]}$$

$$\xi_i^{\text{ATT}} = \underbrace{\frac{1}{\widehat{\rho}}\sum_i W_i Y_i}_{\widehat{\mathbb{E}}[Y^{(1)}|W=1]} - \underbrace{\frac{1}{\widehat{\rho}}\sum_i \left[W_i\widehat{m}^{(0)}(s_i)\right] + (1-W_i)\frac{\widehat{e}(s_i)}{1-\widehat{e}(s_i)}\left\{Y_i - \widehat{m}^{(0)}(s_i)\right\}}_{\widehat{\mathbb{E}}[Y^{(0)}|W=1]}$$

where $\widehat{\rho} = 1/n\sum_i W_i$ is the unconditional probability of treatment. With these scores, an estimator for the ATE and ATT can be constructed by averaging them over the sample

$\widehat{\tau}^j = \frac{1}{n}\sum_{i=1}^n \xi_i^j$ for $j \in$ {ATE, ATT}. The counterfactual potential outcome $\widehat{Y}^{(w)}$ is constructed as a prediction from an outcome model $\widehat{m}^{(w)}$ plus an inverse-propensity weighted residual term. This means that when the outcome model is correctly specified, the residuals are noise and achieves consistency via outcome modelling, while if the propensity score is correctly specified, the estimator achieves consistency via inverse propensity weighting and the outcome models cancel out. This estimator is also semiparametrically efficient, which implies that no regular estimator can improve upon its asymptotic risk (Hahn, 1998).

The presence of the propensity score in the denominator also immediately suggests that the AIPW estimator relies heavily on overlap and well calibrated propensity scores. Its performance can degrade quickly when propensity scores are extreme (J. D. Kang and Schafer, 2007) or under 'mild' misspecification, which gets magnified by the inversion of the propensity score. We now turn to alternative estimators that sidestep the need to estimate and invert the propensity score.

2.3.2. **Regression Imputation Methods.** The augmented IPW approach in the previous section was developed for a completely general setting, where the propensity score is performing the role of dimension-reduction by conveying how units are arranged in covariate space $\mathcal{S}$, such that one compares like-for-like units. When the dimension of covariates is not too large (for example in the spatial setting), nearest-neighbours matching based on location s_i and related methods provide an alternative route to estimate treatment effects without the potentially fraught estimation of the propensity score.

A hybrid method with attractive properties in the spatial setting is the Bias-corrected Matching (BCM) or imputation estimator proposed by Abadie and G. W. Imbens (2011), which combines matching and regression to impute missing potential outcomes. Matching estimators impute $Y^{(0)}$ and $Y^{(1)}$ with a nearest-neighbour estimate of $m^1(s_i), m^0(s_i)$ respectively.

The imputation estimator augments the matching estimator with a debiasing term that regression-adjusts for differences in covariate values. Lin and Han (2022) show that this approach can be generalized to a family of estimators involving linear-smoothers with a $n \times n$ smoother matrix Ω with elements $[\omega_{i \to j}]_{i,j}$ contains weights for units j matched with unit i. Ω is learnt from covariates s_i for the treatment and control group.

$$\widetilde{Y}_i^0 = \begin{cases} Y_i & \text{if } W_i = 0 \\ \frac{1}{M} \sum_{j:W_i=0} \omega_{i \leftarrow j} \left(Y_j + \widehat{m}^0(s_i) - \widehat{m}^0(S_j) \right) & \text{if } W_i = 1 \end{cases}$$

$$\widetilde{Y}_i^1 = \begin{cases} \frac{1}{M} \sum_{j:W_i=0} \omega_{i \leftarrow j} \left(Y_j + \widehat{m}^1(s_i) - \widehat{m}^1(S_j) \right) & \text{if } W_i = 0 \\ Y_i & \text{if } W_i = 1 \end{cases}$$

In the case of 1:M nearest-neighbours matching with bias-correction (Abadie and G. W. Imbens, 2011), $\omega_{i,j}$ is $1/M$ for units in the matched set for unit $i, \mathcal{M}_i$, which comprises units from the opposite treatment group with similar values of s_i, and 0 everywhere else. Lin and Han (2022) show that a wide set of methods including kernel regression, local linear regression, nearest neighbours matching, and random forests can be expressed as in the above linear smoother framework. This sidesteps the need to estimate the propensity score $e(s_i)$, and instead involves first fitting nonparametric regressions to estimate $\widehat{m}^{(w)}$ followed by a multiplication with a smoother matrix Ω, which is easily constructed for matching estimators but may otherwise require computation of distances (e.g. for kernel estimators).

Lin and Han (2022) show that imputation methods also attain the semiparametric efficiency bound, and as such are competitive with augmented IPW estimators without the potentially error-magnifying step of propensity score inversion.

2.3.3. **Balancing Weights for Geographic Treatments.** The augmented inverse propensity weighting estimator above requires smoothness in the outcome and propensity models,

which may be infeasible in many settings where treatment assignment is 'patch' wherein blocks of units get treated while others don't, which is akin to a geographic regression discontinuity (Geo-RD) design. In such settings, the propensity score is zero or one in large segments of the map, and therefore an alternate estimand and/or estimator are needed.

Focussing on the effect close to the boundary and invoking smoothness in the outcome model typically motivates a geographic-RD approach (Keele and Titiunik, 2015). Most applied work[3] transports ideas from uni-dimensional regression discontinuity designs where treatment assignment is a deterministic function of a scalar 'running variable' x, and proceeds with estimation using local-linear regression. This approach has obvious shortcomings in geographical settings for two reasons: (1) multivariate running variables potentially induce improper comparisons across units that are quite far apart in space[4], and (2) regression discontinuity designs rely on kernel-weighted comparisons within a narrow bandwidth close to the discontinuity, which is challenging to calibrate in geographic settings since conventional bandwidth-selection approaches (Calonico, Cattaneo, and Titiunik, 2014; G. Imbens and Kalyanaraman, 2012) select bandwidth to minimize mean-squared-error under the assumption of independent and identically distributed observations close to the threshold, which is implausible in the geographic setting thanks to spatial dependence. Furthermore, it is unclear why a single global bandwidth would be appropriate when treatment and control regions are interspersed, and the outcome model is smooth and heteroskedastic in space, as is often the case.

An alternative to the seemingly intractable problem of choosing a multivariate kernel and optimal bandwidth for regression discontinuity type approaches is to rely on smoothness

[3] e.g. Dell (2010), who popularized the geographic regression discontinuity design in development economics and advocates for global polynomials of location, which may have undesirable effects as illustrated by Gelman and G. Imbens (2019)

[4] Applied researchers typically incorporate boundary-segment fixed-effects in their estimation strategies to account for this, which restricts the regression comparison to either side of the boundary (Keele and Titiunik, 2015). However, this requires an ad-hoc partition of the boundary between treatment and control regions, and further complicates inference.

of the outcome model (which can be operationalized as a convex function class of a pre-specified smoothness e.g. $\mathcal{F} : \{\|\nabla^2 \mu(S)\| \leq B)\}$, which is heuristically invoked for regression discontinuity approaches), and using weights that directly minimise worst-case regression error, as proposed by G. W. Imbens and Wager (2019) in the context of generic regression discontinuity problems. For a given bound on the second-derivative of the outcome model μ_0 (which is typically invoked by researchers in geographical settings and is closely related to our assumption of $U-$smoothness in the previous section), they propose solving the following optimization problem

$$\widehat{\tau} = \sum_{i=1}^{n} \widehat{\gamma}_i Y_i, \quad \widehat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \gamma_i^2 \sigma_i^2 + \left\{ \sup_{\|\nabla^2 \mu_0(s)\| \leq B} \left(\sum_{i=1}^{n} \gamma_i \mu_0(s_i) \right) \right\}^2 : \sum_{i=1}^{n} \gamma_i W_i = 1 \right\}$$

This approach solves for the *minimax-optimal* linear estimator (i.e. minimax among all estimators of the form $\widehat{\tau} = \sum_{i=1}^{n} \widehat{\gamma}_i Y_i$ conditional on X). This approach estimates the weighted conditional average treatment effect with weights chosen to make the inference as precise as possible[5], which is closely related to the practice of propensity score trimming (Crump et al., 2009) or overlap weighting (F. Li, Morgan, and Zaslavsky, 2018) to target feasible estimands in the presence of limited overlap. The above approach can be implemented using most modern optimization packages.

2.4. **Inference.**

The partially linear and augmented inverse propensity weighting estimators proposed in the present paper are semiparametrically efficient, in that they are

[5]this is a constant-effects special case of the general formulation in G. W. Imbens and Wager (2019), where the bias term incorporates both outcome models and is therefore

$$\sup_{\mu_0(\cdot),\mu_1(\cdot)} \left\{ \sum_{i=1}^{n} \gamma \mu_{w_i}(s_i) - (\mu_1(c) - \mu_0(c)) : |\mu''_w(s)| \leq B \; \forall \; w, s \right\}$$

The constant effects approximation yields a conditional-variance weighted average of heterogeneous effects, as with our 2.3

the most efficient regular estimators of their respective estimands and attain their respective efficiency bounds (Newey, 1994; Hahn, 1998). Indeed, precision considerations may inform a researcher's choice between them: the semiparametric efficiency bound for the Average Treatment Effect (ATE) derived by Hahn (1998) takes the following form $V = \mathbb{V}\left[\tau(S)\right] + \mathbb{E}\left[\frac{\sigma_1^2}{e(X)} + \frac{\sigma_0^2}{(1-e(X))}\right]$. The second term divides by the pscore, and therefore can blow up when overlap is poor.

Each can be characterised using its influence function (Chernozhukov, Chetverikov, Demirer, et al., 2018)

$$\psi^{\text{PLR}} = (Y - \widehat{m}(S) - \tau(W - \widehat{e}(S)))(W - \widehat{e}(S))$$

$$\psi^{\text{ATE}} = \widehat{m}^1(\mathbf{s}) + \frac{W(Y - \widehat{m}^1(\mathbf{s}))}{\widehat{e}(\mathbf{s})} \quad - \quad \widehat{m}^0(\mathbf{s}) + \frac{(1-W)(Y - \widehat{m}^0(\mathbf{s}))}{1 - \widehat{e}(\mathbf{s})} - \tau$$

$$\psi^{\text{ATT}} = \frac{1}{\widehat{\rho}}\sum WY \quad - \quad \frac{1}{\widehat{\rho}}\sum \left[W\widehat{m}^{(0)}(S)\right] + (1-W)\frac{\widehat{e}(S)}{1-\widehat{e}(S)}\left\{Y - \widehat{m}^{(0)}(S)\right\} - \frac{W}{\widehat{\rho}}\tau$$

These functions are averaged for a point estimate (as in the previous subsection) and its standard deviation $\sqrt{\widehat{\psi}_i/n}$ can be used to construct asymptotic confidence intervals. Inference on parameters involving first-step estimation of nuisance functions (such as m and e above) is generally a challenging problem. Classical theory of semiparametric regression suggests that we can ignore estimation error in the semiparametric steps (Robinson, 1988; Andrews, 1994) under the implausibly strong assumption of the orthogonality between the treatment and confounder (which is the reason we are performing covariate adjustment in the first place). More recent work on two-step estimation for nuisance parameters typically requires the nonparametric regressions to belong to restricted model classes (satisfying Donsker conditions) such as sparse regularized regression (Belloni et al., 2017).

With arbitrarily flexible model classes, the asymptotic distribution of partially linear regression (PLM) or augmented IPW estimator (AIPW) suffers from 'own-observation' bias, which

can be remedied by require cross-fitting, wherein the prediction of $\widehat{m}^0, \widehat{m}^1, \widehat{e}$ for observation i is based on models that exclude observation i. This then permits $\sqrt{n}$ consistent inference on the treatment effects. However, this procedure may be data-inefficient in medium-scale data settings, where withholding some fraction of the data considerably worsens model fit, thereby increasing the bias in treatment effect estimation.

Chen, Syrgkanis, and Austern (2022) show that for a wide class of regression techniques that satisfy the *leave-out stability* property (wherein replacement of any individual data point with an independent copy from the same distribution does not change model fit substantially), $\sqrt{n}$ consistency is feasible without sample splitting. Leave-out stability is considerably weaker than the Donsker conditions required in the prior literature and therefore encompasses a wide variety of nonparametric regressions, including regularized regressions, generalized additive models (GAMs), and ensemble bagging estimators built via subsampling without replacement (which is analysed by Chen, Syrgkanis, and Austern (2022) as a leading case). This implies that for a variety of regression methods used to fit outcome and propensity models for the partially linear or augmented IPW estimators described in subsections 2.2, sample-splitting is not necessary for inference. For the PLM, residuals on residuals regression with robust standard errors provides valid confidence intervals, while for AIPW with GAMs, splines, or bagging estimators[6], the nuisance models $m^{(w)}(\cdot)$ and $e(\cdot)$ can be fit using the entire dataset as with simpler parametric models.

An alternative approach is to use the nonparametric bootstrap. However, a naive implementation of the bootstrap is challenging in this setting, as any given bootstrap replication with replacement might result in draws where a unit's of its neighbours may have dropped out completely, or repetitions of units' neighbours such that the adjacency matrix is degenerate. This leads the high-level Hadamard differentiability conditions that ensure the validity of the Bootstrap to break down, heuristically in the same manner as with matching

[6]This class notably does not include random forests or neural networks, which do not satisfy the leave-out condition required for Stochastic Equicontinuity to continue to hold without sample splitting (Chen, Syrgkanis, and Austern, 2022).

(Abadie and G. W. Imbens, 2008) and the LASSO (Camponovo, 2015). This means that estimates aren't defined for most bootstrap samples, and therefore a sampling distribution cannot be constructed. Jackknife procedures, however, may continue to work as long as the units have more than 1 neighbour.

The potential failure of the standard bootstrap motivates our recommended use of the Bayesian 'Random weighting' bootstrap (D. B. Rubin, 1981; Mason and Newton, 1992; Chamberlain and G. W. Imbens, 2003) for this problem. In this approach, instead of re-sampling units with replacement, in each iteration, one draws an $n-$vector of Dirichlet weights $\mathbf{W}_j \sim \text{Dirichlet}(n; 1, \ldots, 1)$ (exponential weights for each unit) and recomputes the estimates. Assuming units are exchangeable conditional on location, this yields valid inference on target parameters. This approach is also closely related to the multiplier bootstrap (Belloni et al., 2017; Chernozhukov, Chetverikov, Kato, et al., 2022).

3. Simulation Study

3.1. Setup. To benchmark the performance of the semiparametric estimators proposed in the present paper against standard adjustment strategies, we conduct simulation studies with spatial confounding with varying degrees of smoothness and evaluate their performance in estimating the known true estimates with both constant and heterogeneous effects. In particular, we draw data on a $40 \times 40 = 1600$ cell grid (regular lattice, with cells as 'districts'), where the unobserved confounder U is simulated from a Gaussian Process with Matern covariance possessing the following covariance function

$$R(d; \theta, \nu) = \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{2\sqrt{\nu} d}{\theta} \right)^{\nu} \mathcal{K}_{\nu} \left(\frac{2\sqrt{\nu} d}{\theta} \right)$$

where $K(\cdot)$ is a modified Bessel function of the second kind, d is Euclidean distance between two locations, θ is a range parameter, and most importantly for our purposes, $\nu > 0$ is a

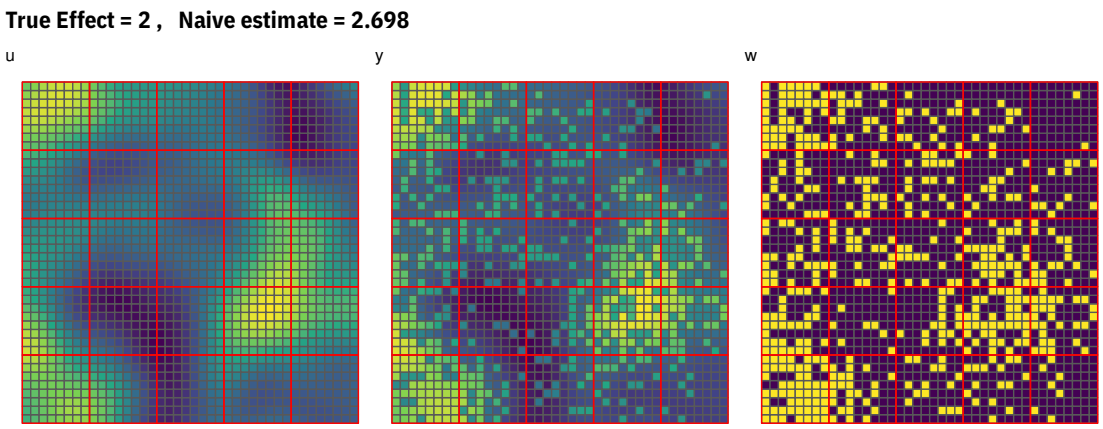


FIGURE 2. Confounder, Outcome, and Treatment Distribution

smoothness parameter that we vary to control the 'measurability' of U via S (A4 in 2.1). Intuitively, the larger the value of ν, the easier it is for smoothing methods to learn the model for U and partial out its effects. We also partition the map into groups of 8×8 'districts' as 'states', which accommodates the standard practice of within-state comparisons. We generate the treatment and outcome as

$$W_i \sim \text{Bernoulli}(\text{logit}(U + \eta_i))) \;\; ; \eta_i \sim \mathcal{N}(0, 1)$$

$$Y_i = \tau W_i + U + \varepsilon_i \;\; ; \varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$$

Since U is present in both the propensity score and outcome model, spatial confounding is present in the DGP.

As a warm-up, we illustrate a single realisation of the DGP with $\nu = 5$ in 2, and residuals and estimates from various partialling out strategies in 3. We find that naive regression substantially over-estimates the effect, as does parametric adjustment and state-fixed effects. The four semiparametric smoothing estimators, on the other hand, get the answer almost exactly right.

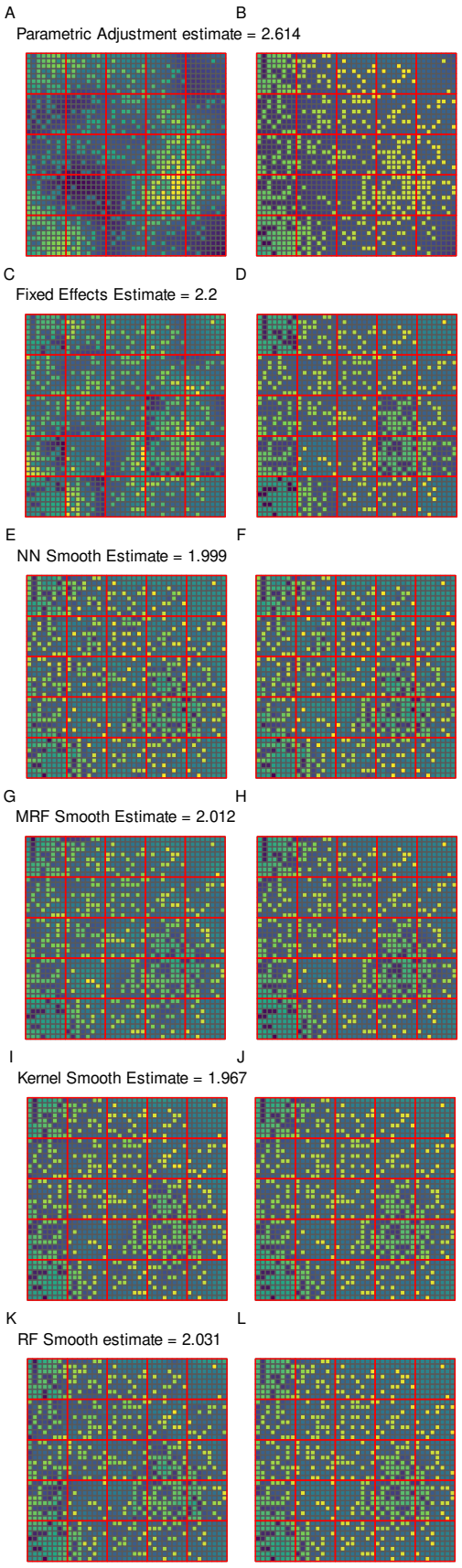
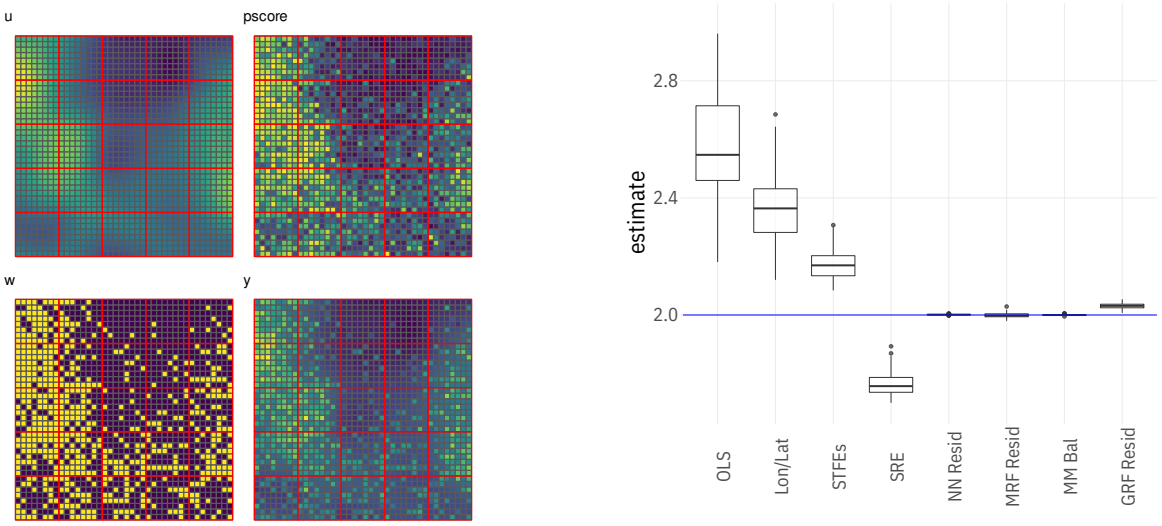


FIGURE 3. Residualised data and estimates from various semiparametric regressions



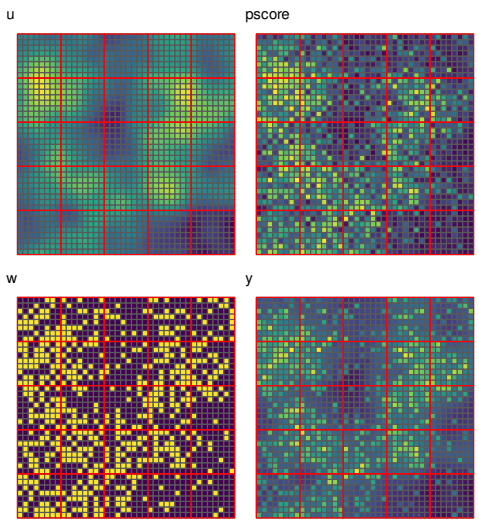
(A) $\nu = 6$: Single Realisation (B) $\nu = 6$: Sampling Distribution

FIGURE 4. Smooth Confounding: $\nu = 6$

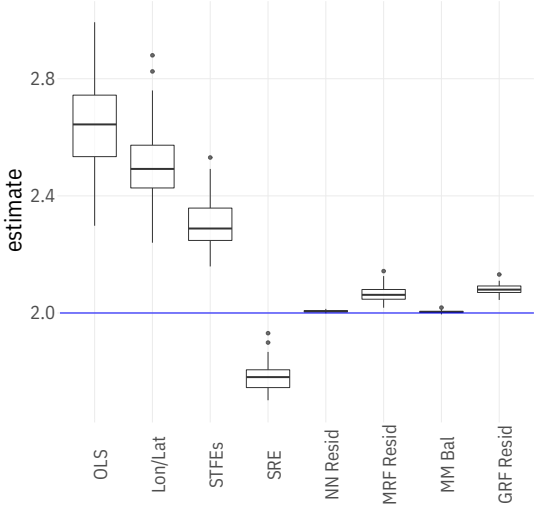
3.2. Results. Next, we compare the performance of the semiparametric estimators with that of conventional approaches. We report simulation results in figs 4, 5, and 6 where a single realisation of the DGP for $\nu = 6, 3, 0.5$ is displayed on the left, and the corresponding sampling distribution around the causal effect estimate of $\tau = 2$ is reported on the right.

The traditional estimators under consideration are (1) naive linear regression (`OLS`) which ignores spatial information entirely, (2) parametric spatial adjustment (`Lon/Lat`) which fits linear and quadratic functions of location, State FEs (`STFEs`) which restricts to *within-state* comparisons, and Spatial random effects (`SRE`) which follows the geo-statistics strategy of fitting an outcome model with spatial random effects assumed to be orthogonal to the treatment. The estimators proposed in the paper are local differencing (`NN Resid`), which residualises on first-degree neighbours' treatment and outcome averages, GMRF residuals on residuals (`MRF Resid`, and generalized random forest residuals on residuals (`GRF Resid`), and minimax balancing (`MM Bal`) which implements the balancing weights algorithm.

We find that when the confounder is smooth ($\nu = 6$ and 3), conventional estimation strategies suffer from considerable bias, while the semiparametric estimators perform very well. This is because the smoothing models are able to learn the representation of the confounder

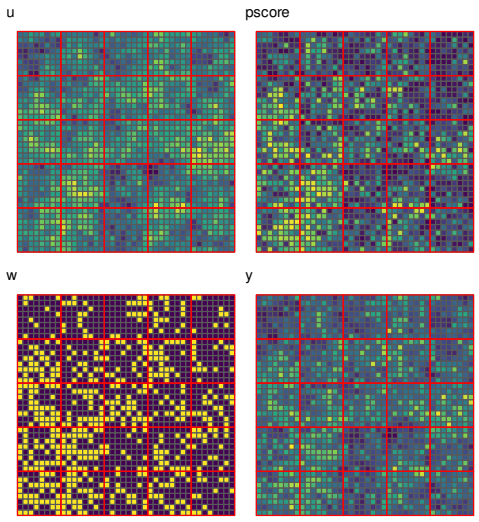


(A) $\nu = 3$: Single Realisation

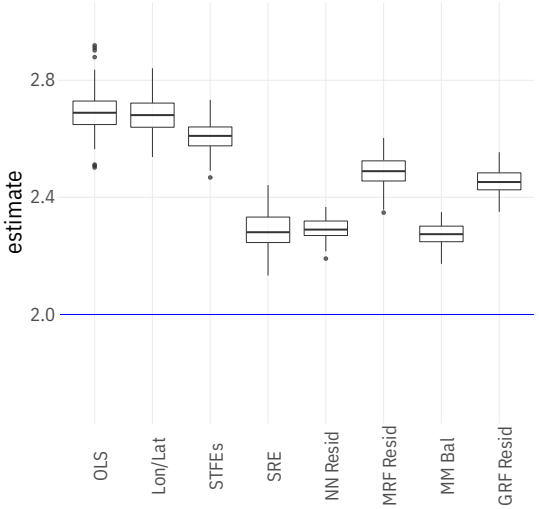


(B) $\nu = 3$: Sampling Distribution

FIGURE 5. Medium-Smooth Confounding: $\nu = 3$



(A) $\nu = 0.5$: Single Realisation



(B) $\nu = 0.5$: Sampling Distribution

FIGURE 6. Noisy Confounding: $\nu = 3$

using location data s. When the confounder is noisy, all estimators are biased, but semiparametric estimators are still considerably less biased than naive methods. When the confounder is very noisy ($\nu = 0.5$), there is very little signal about the confounder in the spatial location of an observation; using very narrow windows of comparison, as in spatial differences (where only adjacent units' values are used to construct a prediction for a given unit's value) or minimax balancin performs best.

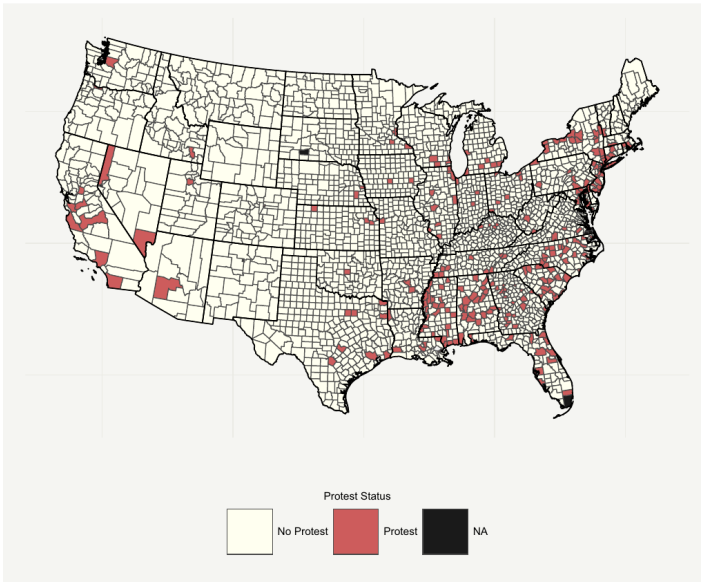
4. Empirical Applications

Next, we implement the methods under consideration on two real data examples. We find that when the unconfoundedness assumption is somewhat plausible based on the treatment distribution, semiparametric estimators produce similar estimates to conventional parametric ones. On the other hand, when the treatment and outcome distributions are highly spatially correlated, semiparametric estimators produce substantially attenuated estimates.

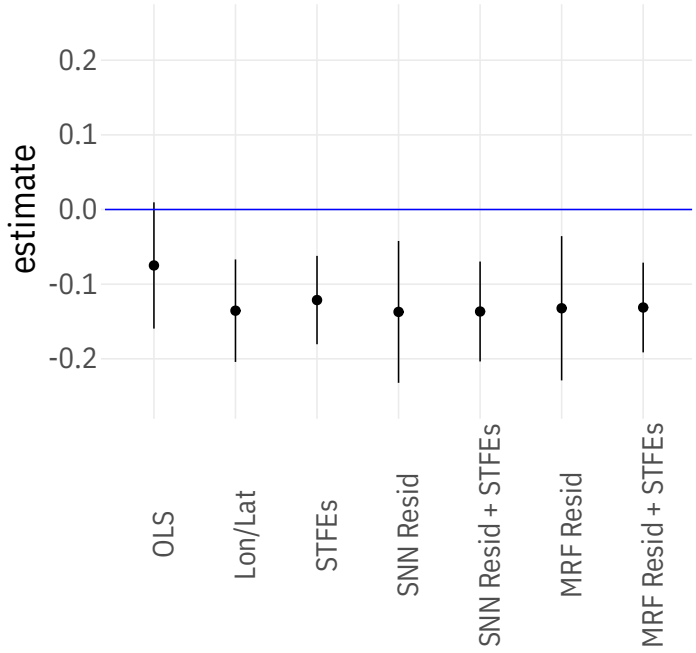
Mazumder (2018) studies whether historical civil rights protest activities in the 1960s is affects contemporary racial attitudes among whites in the US. To do so, he argues that protest activity at the county level (fig 7a) is exogenous conditional on geography, and regresses county-level measures of contemporary racial resentment on a dummy for protest presence, state fixed effects, and various geographic controls in his preferred specification and finds that counties that experienced protest activity in the 1960s have lower levels of contemporary racial resentment measures. We partial controls out linearly and estimate the naive specification, parametric controls for location, state-FEs, spatial nearest-neighbours residuals with and without FEs, and Markov random field residuals with and without FEs, and report effect estimates in fig 7b. With the exception of the naive OLS estimate, which is likely highly confounded, the rest of the estimators yield remarkably similar estimates and intervals. This suggests that the findings in the paper are relatively robust to smooth spatial confounding.

Dincecco et al. (2022) study the pre-colonial roots of local economic development in India. They conjecture that higher levels of pre-colonial conflict between rival states increased state-capacity in locations that experienced them (right panel of 8a, and this manifests in higher levels of local level economic development measured by nighttime luminosity (DMSP) (left panel of 8a. They also argue for unconfoundedness conditional on state fixed effects. Their treatment is continuous with substantial heaping at zero (left panel of 8b), so we discretise it to a binary measure at the median (which effectively codes the treatment

FIGURE 1 The Geographic Distribution of Civil Rights Protests, 1960–65



(A) Protest activity distribution (figure from original paper)



(B) Coefficient estimates

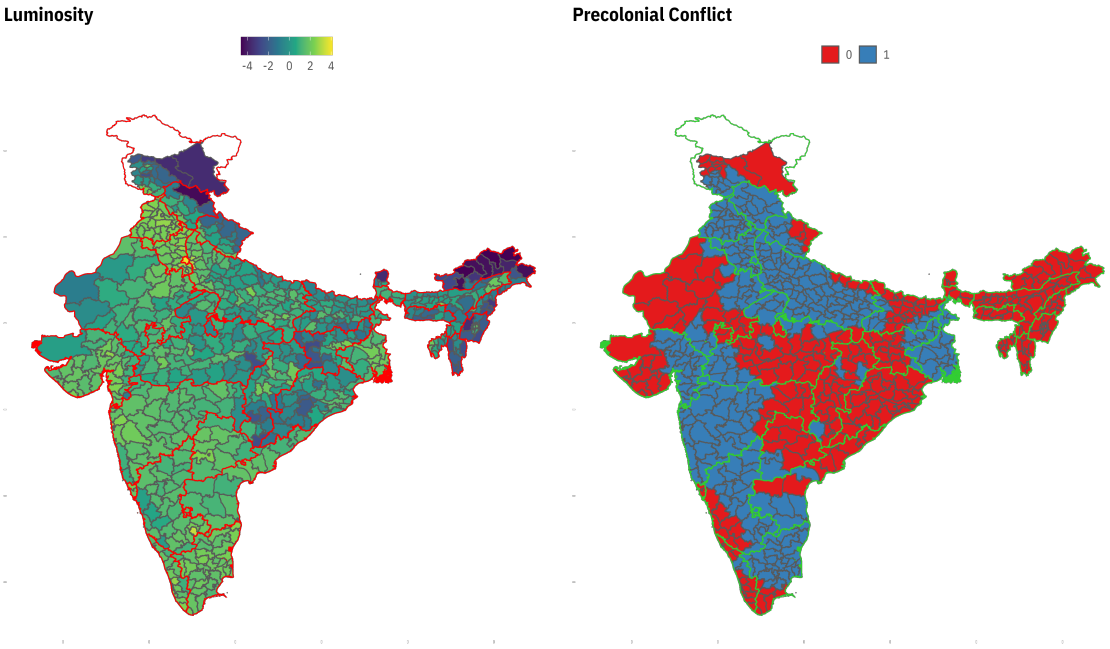
FIGURE 7. Treatment distribution (Top panel) and estimates (Bottom panel) for Mazumder (2018)

as non-zero conflict). As with the previous example, we partial controls out linearly and estimate the naive specification, parametric controls for location, state-FEs, spatial nearest-neighbours residuals with and without FEs, and Markov random field residuals with and

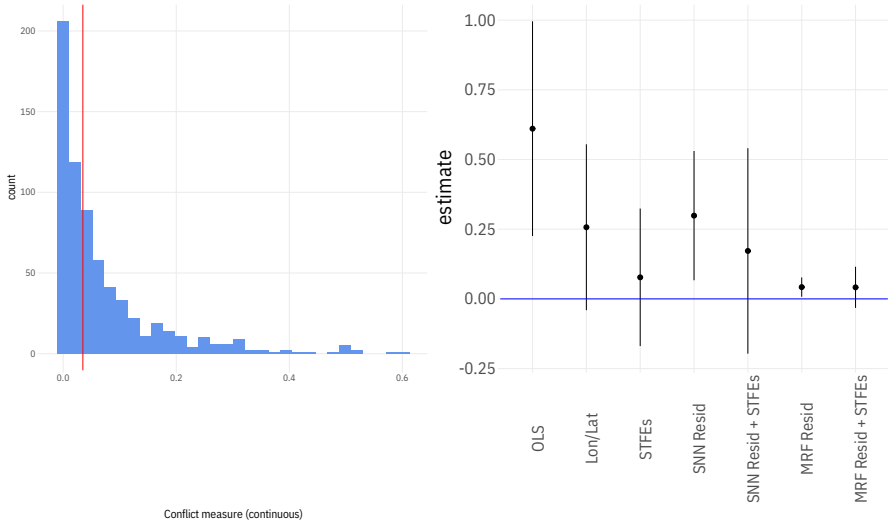
without FEs, and report effect estimates in fig 8a. In contrast to the previous example, the treatment and outcome distribution is highly spatially correlated, and therefore spatial smoothing substantially attenuates estimates. Smoothing using either nearest neighbours or Markov random fields with state fixed effects attenuates the effect down to zero, with the latter being precisely estimated (potentially because of longer trends in the outcome data because it is a satellite measure).

5. Conclusion

In summary, we have proposed a class of semiparametric estimators for covariate adjustment using spatial data under unconfoundedness given location and proximal designs. These are motivated by semiparametric regression and double-machine learning methods, which recognise that consistent causal estimation is possible by first residualising the treatment and outcome on flexible functions of covariates, which in the spatial setting is best performed by spatial smoothing methods. We provide conditions for nonparametric identification, outline several promising spatial regression strategies for the partial out step, and discuss inference using the Bayesian bootstrap. Next, we illustrate the promising performance of semiparametric adjustment relative to conventional strategies in a variety of simulation studies, and show that for smooth to moderate confounding, partialling out estimators yield unbiased estimates, while when confounding is very noisy, they yield biased estimates like all other strategies, but their bias is the smallest among all estimators considered. We conclude by illustrating the use of the estimators in two empirical examples, where they yield similar and substantially attenuated effect estimates depending on the plausibility of the 'geographic natural experiment'.



(A) Precolonial conflict (w) and luminosity (y) with state boundaries overlaid



(B) Treatment distribution with median in red (used to discretize treatment) and coefficient estimates

FIGURE 8. Treatment distribution (Top panel) and estimates (Bottom panel) for Dincecco et al. (2022)

References

ABADIE, Alberto and Guido W IMBENS (2008). "On the failure of the bootstrap for matching estimators". *Econometrica* 76.6, pp. 1537–1557 (cit. on p. 21).

Abadie, Alberto and Guido W Imbens (2011). "Bias-corrected matching estimators for average treatment effects". *Journal of Business & Economic Statistics* 29.1, pp. 1–11 (cit. on pp. 15, 16).

Andrews, Donald W. K. (1994). "Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity". *Econometrica* 62.1, pp. 43–72 (cit. on p. 19).

Angrist, Joshua D (1998). "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants". *Econometrica: journal of the Econometric Society* 66.2, pp. 249–288 (cit. on pp. 9, 13).

Aronow, Peter M and Cyrus Samii (Jan. 2016). "Does Regression Produce Representative Estimates of Causal Effects?" *American journal of political science* 60.1, pp. 250–267 (cit. on p. 9).

Athey, Susan, Julie Tibshirani, and Stefan Wager (Apr. 2019). "Generalized random forests". en. *Annals of statistics* 47.2, pp. 1148–1178 (cit. on p. 11).

Baum-Snow, Nathaniel and Fernando Ferreira (2015). "Chapter 1 - Causal Inference in Urban and Regional Economics". *Handbook of Regional and Urban Economics*. Ed. by Gilles Duranton, J Vernon Henderson, and William C Strange. Elsevier (cit. on p. 4).

Belloni, Alexandre et al. (2017). "Program evaluation and causal inference with high-dimensional data". *Econometrica* 85.1, pp. 233–298 (cit. on pp. 3, 19, 21).

Besag, Julian (1974). "Spatial interaction and the statistical analysis of lattice systems". *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2, pp. 192–225 (cit. on p. 38).

Blangiardo, Marta and Michela Cameletti (2015). *Spatial and spatio-temporal Bayesian models with R-INLA* (cit. on p. 3).

Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik (2014). "Robust nonparametric confidence intervals for regression-discontinuity designs". *Econometrica: journal of the Econometric Society* 82.6, pp. 2295–2326 (cit. on p. 17).

Camponovo, Lorenzo (2015). "On the validity of the pairs bootstrap for lasso estimators". *Biometrika* 102.4, pp. 981–987 (cit. on p. 21).

Chamberlain, Gary and Guido W Imbens (2003). "Nonparametric applications of Bayesian inference". *Journal of Business & Economic Statistics* 21.1, pp. 12–18 (cit. on p. 21).

Chen, Qizhao, Vasilis Syrgkanis, and Morgane Austern (June 2022). "Debiased Machine Learning without Sample-Splitting for Stable Estimators". arXiv: 2206.01825 [econ.EM] (cit. on p. 20).

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, et al. (Feb. 2018). "Double/debiased machine learning for treatment and structural parameters". *The econometrics journal* 21.1, pp. C1–C68 (cit. on pp. 2, 8, 9, 12, 14, 19).

Chernozhukov, Victor, Denis Chetverikov, Kengo Kato, et al. (May 2022). "High-dimensional Data Bootstrap". arXiv: 2205.09691 [math.ST] (cit. on p. 21).

Christensen, Ronald (2001). *Advanced Linear Modeling: Multivariate, Time Series, and Spatial Data; Nonparametric Regression and Response Surface Maximization*. Springer Science & Business Media (cit. on p. 3).

Cressie, Noel (2015). *Statistics for spatial data*. John Wiley & Sons (cit. on p. 3).

Cressie, Noel and Jennifer L Davidson (1998). "Image analysis with partially ordered Markov models". *Computational statistics & data analysis* 29.1, pp. 1–26 (cit. on p. 13).

Cressie, Noel and Gardar Johannesson (2008). "Fixed rank kriging for very large spatial data sets". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.1, pp. 209–226 (cit. on pp. 3, 13).

Crump, Richard K et al. (2009). "Dealing with limited overlap in estimation of average treatment effects". *Biometrika* 96.1, pp. 187–199 (cit. on pp. 9, 18).

Dell, Melissa (2010). "The persistent effects of Peru's mining mita". *Econometrica* 78.6, pp. 1863–1903 (cit. on pp. 4, 17).

Dincecco, Mark et al. (2022). "Pre-colonial warfare and long-run development in India". *The Economic Journal* 132.643, pp. 981–1010 (cit. on pp. 26, 29).

Druckenmiller, H and S Hsiang (2018). "Accounting for Unobservable Heterogeneity in Cross Section Using Spatial First Differences" (cit. on p. 11).

Gelman, Andrew and Guido Imbens (2019). "Why high-order polynomials should not be used in regression discontinuity designs". *Journal of Business & Economic Statistics* 37.3, pp. 447–456 (cit. on p. 17).

Gilbert, Brian, Abhirup Datta, and Elizabeth Ogburn (Dec. 2021). "Approaches to spatial confounding in geostatistics". arXiv: 2112.14946 [stat.ME] (cit. on pp. 4, 5).

Hahn, Jinyoung (1998). "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects". *Econometrica* (cit. on pp. 14, 15, 19).

Hu, Yuchen, Shuangning Li, and Stefan Wager (2021). "Average Direct and Indirect Causal Effects under Interference". *Biometrika* (cit. on p. 4).

Imbens, Guido and Karthik Kalyanaraman (2012). "Optimal Bandwidth Choice for the Regression Discontinuity Estimator". *The Review of economic studies* 79.3, pp. 933–959 (cit. on p. 17).

Imbens, Guido W (Feb. 2004). "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review". *The review of economics and statistics* 86.1, pp. 4–29 (cit. on p. 14).

Imbens, Guido W and Stefan Wager (2019). "Optimized regression discontinuity designs". *Review of Economics and Statistics* 101.2, pp. 264–278 (cit. on pp. 4, 18).

Imbens, Guido W and Tristan Zajonc (2009). "Regression discontinuity design with vector-argument assignment rules". *Unpublished paper* (cit. on p. 4).

Kang, Emily L and Noel Cressie (2011). "Bayesian inference for the spatial random effects model". *Journal of the American Statistical Association* 106.495, pp. 972–983 (cit. on pp. 3, 13).

Kang, Joseph DY and Joseph L Schafer (2007). "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data" (cit. on p. 15).

Keele, Luke and Rocio Titiunik (2015). "Geographic boundaries as regression discontinuities". *Political analysis: an annual publication of the Methodology Section of the American Political Science Association* 23.1, pp. 127–155 (cit. on pp. 4, 17).

– (2016). “Natural experiments based on geography”. *Political Science Research and Methods* 4.1, pp. 65–95 (cit. on pp. 2, 4).

Kelejian, Harry and Gianfranco Piras (2017). *Spatial econometrics*. Academic Press (cit. on p. 11).

Kelejian, Harry H and Ingmar R Prucha (1999). “A generalized moments estimator for the autoregressive parameter in a spatial model”. *International economic review* 40.2, pp. 509–533 (cit. on p. 4).

– (2010). “Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances”. *Journal of econometrics* 157.1, pp. 53–67 (cit. on pp. 3, 4).

Li, Fan, Kari Lock Morgan, and Alan M Zaslavsky (2018). “Balancing covariates via propensity score weighting”. *Journal of the American Statistical Association* 113.521, pp. 390–400 (cit. on p. 18).

Lin, Zhexiao and Fang Han (Dec. 2022). “On regression-adjusted imputation estimators of the average treatment effect”. arXiv: 2212.05424 [math.ST] (cit. on p. 16).

Lindgren, Finn, Håvard Rue, and Johan Lindström (2011). “An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.4, pp. 423–498 (cit. on p. 13).

Mason, David M and Michael A Newton (1992). “A rank statistics approach to the consistency of a general bootstrap”. *The Annals of Statistics*, pp. 1611–1624 (cit. on pp. 3, 21).

Mazumder, Soumyajit (2018). “The persistent effect of US civil rights protests on political attitudes”. *American Journal of Political Science* 62.4, pp. 922–935 (cit. on pp. 26, 27).

Newey, Whitney K (1994). “The Asymptotic Variance of Semiparametric Estimators”. *Econometrica: journal of the Econometric Society* 62.6, pp. 1349–1382 (cit. on p. 19).

Paciorek, Christopher J (2010). "The importance of scale for spatial-confounding bias and precision of spatial regression estimators". *Statistical science: a review journal of the Institute of Mathematical Statistics* 25.1, p. 107 (cit. on pp. 3, 11).

Robins, James et al. (May 2008). "Higher order influence functions and minimax estimation of nonlinear functionals". arXiv: 0805.3040 [math.ST] (cit. on p. 9).

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao (1994). "Estimation of regression coefficients when some regressors are not always observed". *Journal of the American statistical Association* 89.427, pp. 846–866 (cit. on pp. 2, 14).

Robinson, P M (1988). "Root-N-Consistent Semiparametric Regression". *Econometrica: journal of the Econometric Society* 56.4, pp. 931–954 (cit. on pp. 2, 8, 12, 19).

Rubin, Donald B (1981). "The bayesian bootstrap". *The annals of statistics*, pp. 130–134 (cit. on pp. 3, 21).

Rue, Havard and Leonhard Held (2005). *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC (cit. on pp. 3, 12).

Sävje, Fredrik, Peter M Aronow, and Michael G Hudgens (2021). "Average treatment effects in the presence of unknown interference". *The Annals of Statistics* 49.2, pp. 673–701 (cit. on p. 4).

Schnell, Patrick M and Georgia Papadogeorgou (2020). "Mitigating unobserved spatial confounding when estimating the effect of supermarket access on cardiovascular disease deaths". *The Annals of Applied Statistics* 14.4, pp. 2069–2095 (cit. on pp. 3, 5).

Tobler, Waldo R (1970). "A computer movie simulating urban growth in the Detroit region". *Economic geography* 46.sup1, pp. 234–240 (cit. on p. 2).

Van Der Laan, Mark J and Daniel Rubin (2006). "Targeted maximum likelihood learning". *The international journal of biostatistics* 2.1 (cit. on p. 14).

Vecchia, Aldo V (1988). "Estimation and model identification for continuous spatial processes". *Journal of the Royal Statistical Society: Series B (Methodological)* 50.2, pp. 297–312 (cit. on p. 13).

Wood, Simon N (2006). *Generalized additive models: an introduction with R*. chapman and hall/CRC (cit. on pp. 3, 12).

Yatchew, Adonis (1997). "An elementary estimator of the partial linear model". *Economics letters* 57.2, pp. 135–143 (cit. on p. 11).

Appendix A. Proofs

A.1. Proof of Prop 2.1. By A4, $U = g(S)$ is measurable, so is (S, U), which then ensures that there exists a measurable one-to-one mapping between S and S, U. This means S and S, U induce the same $\sigma-$ algebra and therefore give rise the same conditional distribution.

$$Y(w) \perp\!\!\!\perp W|U, S \quad \Longrightarrow \quad Y(w) \perp\!\!\!\perp W|S$$

□

A.2. Proof of Prop 2.2.

$$\begin{aligned}
\mathbb{E}\left[Y^{(w)}\right] &= \int \mathbb{E}\left[Y^{(1)}|U = u\right] \mathsf{d}P(U) \\
&= \int \mathbb{E}\left[Y^{(1)}|S = s\right] \mathsf{d}P(S) && \text{by A4} \\
&= \int \mathbb{E}\left[Y^{(1)}|W = w, S = s\right] \mathsf{d}P(S) && \text{by Prop 2.1} \\
&= \int \mathbb{E}\left[Y|W = w, S = s\right] \mathsf{d}P(S) && \text{by A1}
\end{aligned}$$

where the last quantity involves observable data alone.

A.3. Proof of Prop 2.3. We first work with an oracle version of $\widehat{\tau}^{\text{PLR}}$ that uses true nuisance functions m, e

$$\widehat{\tau}^* = \frac{\sum_{i=1}^n (Y_i - m(s_i))(W_i - e(s_i))}{\sum_{i=1}^n (W_i - e(s_i))^2} \tag{A.1}$$

$$\xrightarrow{p} \frac{\mathbb{E}\left[(Y - m(s))(W - e(S))\right]}{\mathbb{E}\left[(w - e(S))\right]^2} \qquad \text{by LLN to Numerator, Denominator} \tag{A.2}$$

$$= \frac{\mathbb{E}\left[((W - e(S))\tau(S) + \varepsilon)(W - e(S))\right]}{\mathbb{E}\left[(w - e(S))\right]^2} \qquad \text{substitution of } Y - m(S) = (W - e(S))\tau(S) + \varepsilon \tag{A.3}$$

$$= \frac{\mathbb{E}\left[((W - e(S))\tau(S) + \varepsilon)(W - e(S))\right]}{\mathbb{E}\left[e(S)(1 - e(S))\right]} \tag{A.4}$$

$$= \frac{\mathbb{E}\left[((W - e(S))^2\tau(S)\right] + \overbrace{\mathbb{E}\left[\varepsilon(W - e(S))\right]}^{=0 \text{ by orthogonality of } \varepsilon}}{\mathbb{E}\left[e(S)(1 - e(S))\right]} \tag{A.5}$$

$$= \mathbb{E}\left[\frac{e(S)(1 - e(S))}{\mathbb{E}\left[e(S)(1 - e(S))\right]}\tau(S)\right] \tag{A.6}$$

Therefore the probability limit of $\widehat{\tau}^*$ recovers a weighted average of strata-level effects $\tau(S)$ with weights $\frac{e(S)(1-e(S))}{\mathbb{E}[e(S)(1-e(S))]}$, which is nonnegative and integrates to one.

A.4. **Details of Gaussian Markov Random Fields.**

Defn A.1 (Gaussian Markov Random Field).

A Gaussian Markov Random Field is a spatial collection of random variables $\boldsymbol{\gamma} = \{\gamma(s_1), \ldots \gamma(s_n)\}$ for n units that can be specified in terms of a scaled precision $\boldsymbol{\gamma} \sim \mathcal{N}\left(0, \tau^2 \mathbf{Q}^{-1}\right)$.

$\boldsymbol{\gamma}$ is a a GMRF with respect to a labelled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q} > 0$ iff its density has the form

$$\pi(\gamma) = (2\pi)^{-n/2} |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2}(\gamma - \boldsymbol{\mu})^\top \mathbf{Q}(\gamma - \boldsymbol{\mu})\right)$$

and $Q_{ij} \neq 0 \Leftrightarrow \{i, j\} \in \mathcal{E}\ \forall i \neq j$

The precision matrix **Q** is generally sparse, is typically configured to take a **(Conditional) Autoregressive** (CAR) structure (Besag, 1974), such that adjacent units' random effects are correlated.