Inferring Heterogeneous Causal Effects in Presence of Spatial Confounding

Muhammad Osama [1] **Dave Zachariah** [1] **Thomas B. Schön** [1]

Abstract

We address the problem of inferring the causal effect of an exposure on an outcome across space, using observational data. The data is possibly subject to unmeasured confounding variables which, in a standard approach, must be adjusted for by estimating a nuisance function. Here we develop a method that eliminates the nuisance function, while mitigating the resulting errors-in-variables. The result is a robust and accurate inference method for spatially varying heterogeneous causal effects. The properties of the method are demonstrated on synthetic as well as real data from Germany and the US.

1. Introduction

In a wide range of scientific analyses, the quantity of interest is the causal effect of an exposure z on an outcome y, which may vary across space. Consider, for instance, the causal effect of fertilizer usage z on crop yield y. The effect, denoted τ, quantifies how the average crop yield changes per unit of change of fertilizer quantity. Due to variations in the soil, τ will vary as a function of spatial location s.

The goal in this paper is to infer the effect τ from a finite dataset $\mathcal{D}_n = \{(y_i, z_i, s_i)\}_{i=1}^{n}$. Unmeasured confounding variables, however, render this problem non-trivial. Consider a case illustrated in Figure 1a, where there is a positive association between exposure and outcome yet the causal effect is zero, i.e., $\tau \equiv 0$. The causal structure underlying the data-generating process can be expressed using structural equations and summarized by a directed acyclic graph (DAG) (Peters et al., 2017; Pearl, 2009; Hernán & Robins, 2019). The structure is illustrated in Figure 1b, where c denotes an unmeasured confounding variable which gives

---
*Equal contribution [1]Division of System and Control, Department of Information Technology, Uppsala University. Correspondence to: Muhammad Osama <muhammad.osama@it.uu.se>, Dave Zachariah <dave.zachariah@it.uu.se>.
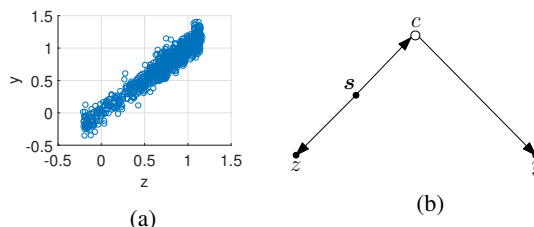
Figure 1. Illustration of spatial confounding. (a) Data exhibiting a clear association between exposure z and outcome y. (b) Causal structure underlying the data-generating process, summarized as a DAG where s denotes spatial location and c is an unmeasured variable. Note that the causal effect of the exposure is zero.

rise to a strong association between z and y despite there being no causal effect. This situation is referred to as *spatial confounding*.

The bulk of the existing literature considers a fixed effect τ across space and, moreover, views c as a residual spatial variation using a spatial random effect (SRE) model (Cressie, 1992). For example, Reich et al. (2006) estimate the effect of socioeconomic factors on stomach cancer incidence ratio using a Gaussian prior model for $c(s)$. Such approaches end up biasing the effect estimates and various restricted spatial regression techniques attempt to mitigate these problems, cf. (Hodges & Reich, 2010; Hughes & Haran, 2013; Hanks et al., 2015; Paciorek, 2010; Page et al., 2017). The geoadditive structural equation model (GSEM) in Thaden & Kneib (2018) eliminates the bias problem by exploiting the causal structure underlying the data generating process. Specifically, they employ a linear Gaussian structural equation model and implicitly remove the spatial variation in z and y due to c. While it is capable of removing systematic errors, the GSEM method is only applicable to discrete space and homogeneous or fixed causal effects. The effect may, however, be *heterogeneous* with respect to spatial location. For example, when applying the method to study the effect of work experience on monthly income (Thaden & Kneib, 2018), it is quite possible that experience will have greater effect in regions with dense industries and high-skilled jobs, as compared to areas dominated by low-skilled jobs.

Existing methods for heterogeneous causal effect build upon statistical machine learning methods but do not ad-

dress spatially varying effects specifically and are, moreover, mainly restricted to binary or categorical exposures z (Imai et al., 2013; Wager & Athey, 2017; Künzel et al., 2017; Nie & Wager, 2017). The standard formulation of the inference problem requires estimating a nuisance function, which leads to significant biases when inferring τ even in the large-sample regime. The orthogonalized formulation introduced by Chernozhukov et al. (2016) circumvents this problem but results in multiplicative errors that are not addressed in the finite sample regime.

In this paper, we develop an inference method based on the orthogonalized formulation that

- can infer heterogeneous causal effects with respect to spatial location,

- is operational in discretized as well as continuous space,

- provides robust and conservative estimates in the finite sample regime.

The outline of the paper is as follows: In Section 2, we present the problem setup and define the heterogeneous effect. In Section 3, we introduce a model of the effect and develop a robust orthogonalized formulation. The resulting method is demonstrated on both simulated as well as real data in Section 4.

Notation: $\mathrm{col}\{a_1, a_2\}$ stacks both elements into a single column vector. $\otimes$ denotes the Kronecker product. $\mathrm{E}_n[a] = \frac{1}{n} \sum_{i=1}^{n} a_i$ denotes the sample mean of a.

2. Problem Formulation

We consider continuous exposures z and outcomes y. Spatial location or region is indexed by $\boldsymbol{s} \in \mathcal{S}$, where $\mathcal{S} \subset \mathbb{R}^d$, if space is continuous, or $\mathcal{S} = \{1, \ldots, d\}$, when space is partitioned into d discrete regions.

Let $y(\widetilde{z})$ be the counterfactual outcome had the exposure been assigned to $\widetilde{z}$, contrary to the observed value z. The (average) causal effect of the exposure at location $\boldsymbol{s}$ is then defined as

$$\tau \triangleq \frac{d}{d\widetilde{z}} \mathrm{E}\left[\, y(\widetilde{z}) \mid \boldsymbol{s} \,\right], \tag{1}$$

We consider scenarios in which the conditional independence

$$z \perp y(\widetilde{z}) \mid \boldsymbol{s} \tag{2}$$

holds. Figure 2 shows examples of causal structures for which (2) is valid. To illustrate the effect of an unmeasured confounding variable, we consider a simple example based on the structure in Fig. 2a.
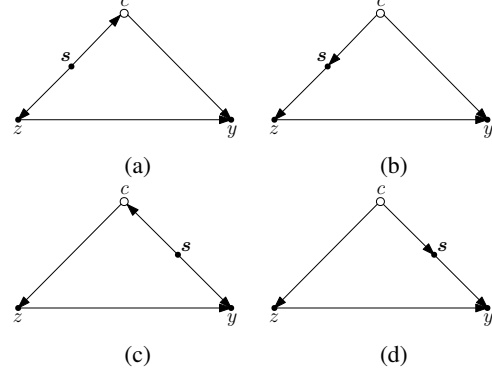

Figure 2. Examples of different DAG configurations that result in spatial confounding, where c is an unmeasured confounding variable. The proposed method can address scenarios that include (a)-(d).

Example:
Let space $\boldsymbol{s}$ be continuous one dimensional. We generate data $\mathcal{D}_n$ shown in Fig. 1a based on the structure

$$y = \tau(\boldsymbol{s})z + c(\boldsymbol{s}),$$
$$z = c(\boldsymbol{s}) + \epsilon,$$

which corresponds to Fig. 2a and where $c(\boldsymbol{s})$ is an unmeasured confounding variable described as a spatial Gaussian Process

$$c(\boldsymbol{s}) \sim \mathcal{GP}(0, k(\boldsymbol{s}, \boldsymbol{s}'))$$

with a Matérn plus white noise covariance function (Williams & Rasmussen, 2006) and ϵ is white Gaussian noise. We consider two cases of effects:

Case 1: Fixed effect $\tau(\boldsymbol{s}) \equiv 0$. A standard approach is to parameterize the unknown effect by a constant and adjust or control for the unmeasured $c(\boldsymbol{s})$ by considering it as spatially structured noise (Cressie, 1992). The effect is then readily estimated using a generalized least squares method that weights the residuals based on the covariance structure of $c(\boldsymbol{s})$. Fig. 3a shows the resulting estimate $\widehat{\tau}$ and its 95% confidence interval (CI) along with $\tau(\boldsymbol{s})$. As can be seen, the estimate is biased and provides misleading inferences. By contrast, the method we propose in this work produces an estimate with 95% CI that clearly contains the true effect, see Fig. 3b.

Case 2: Heterogeneous effect $\tau(\boldsymbol{s})$. In some applications, the effect may vary significantly across space and even change sign. Unlike the standard approaches, the proposed method can also estimate spatially varying $\tau(\boldsymbol{s})$, as shown in Fig. 3c. The estimate closely follows the true effect, whose sign changes across space.
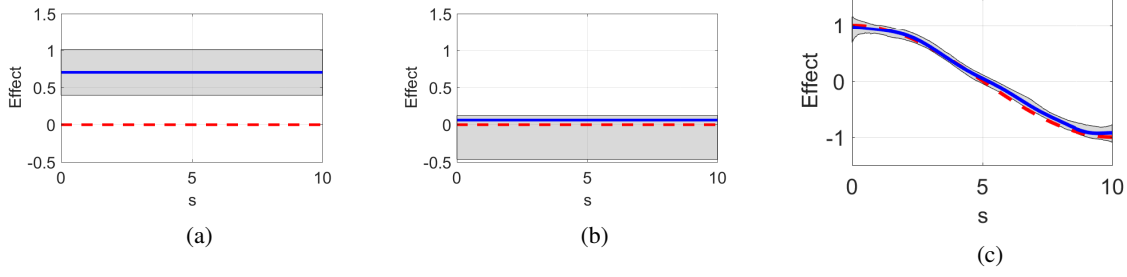
The inability to control for unmeasured $c(\boldsymbol{s})$ by exploiting

Figure 3. Illustration of example. The effect $\tau(\boldsymbol{s})$ of the exposure (dashed line) and its estimates $\widehat{\tau}(\boldsymbol{s})$ (solid line) with 95% CI (shaded regions). Case 1: using (a) the SRE model and (b) The proposed method. Case 2: proposed method in (c). Note that the SRE model is not applicable in this case.

its spatial correlation structure arises because the variable is taken to be independent of the exposure z (Paciorek, 2010). As we see in the example, however, z is highly correlated with $c(\boldsymbol{s})$.

We now formulate the problem in the general case, building upon (2) and modularity (Pearl, 2009; Richardson & Robins, 2013; Hernán & Robins, 2019) we obtain the identity

$$\mathrm{E}\left[y(\widetilde{z})|\boldsymbol{s}\right] \equiv \mathrm{E}\left[y|z = \widetilde{z}, \boldsymbol{s}\right]. \quad (3)$$

Locally at $\boldsymbol{s}$, we assume that the expected outcome (3) is well approximated by an affine function with respect to the exposure $\widetilde{z}$. Then it follows from (1) that the effect is a function of space alone, that is, $\tau(\boldsymbol{s})$. By combining (1) and (3), we get

$$\tau(\boldsymbol{s}) = \frac{d}{dz} \mathrm{E}[y|z, \boldsymbol{s}]\bigg|_{z=\widetilde{z}}$$

After integrating both sides with respect to z, this yields

$$\mathrm{E}[y|z, \boldsymbol{s}] = \tau(\boldsymbol{s})z + \beta(\boldsymbol{s}) \quad (4)$$

which implies the following form of the outcome

$$y = \tau(\boldsymbol{s})z + \beta(\boldsymbol{s}) + \varepsilon, \quad (5)$$

where $\beta(\boldsymbol{s})$ is a nuisance function and ε is a zero-mean error term that is uncorrelated with any function of z and $\boldsymbol{s}$. We see that in the general case, $\beta(\boldsymbol{s})$ will be correlated with a spatially varying exposure z. Therefore omitting it, as in the example above, gives rise to bias due to spatial confounding. On the other hand, taking this term directly into account requires the estimation of a high-dimensional nuisance function $\beta(\boldsymbol{s})$ jointly with the effect $\tau(\boldsymbol{s})$. This leads to non-negligible biases in the finite sample case as pointed out by Chernozhukov et al. (2016) (Illustrated in Fig 4 below). These limitations motivate an alternative approach to infer $\widehat{\tau}(\boldsymbol{s})$ from $\mathcal{D}_n$.

3. Method

In this section, we begin by deriving an orthogonalized formulation of the problem that circumvents the need to directly estimate the nuisance function $\beta(\boldsymbol{s})$ in (5). Then we develop a robust errors-in-variables method that yields heterogeneous effect estimates $\widehat{\tau}(\boldsymbol{s})$ as a function of the spatial location $\boldsymbol{s}$. Finally, we specify a spatial predictive model that we employ in the subsequent experimental section.

3.1. Orthogonalized formulation

Consider residualized versions of the outcome and exposure, $w = y - \mathrm{E}[y|\boldsymbol{s}]$ and $v = z - \mathrm{E}[z|\boldsymbol{s}]$, respectively. These residuals are orthogonal to $\boldsymbol{s}$ and in combination with (4) and (5), we obtain an orthogonalized formulation

$$w = y - \int \mathrm{E}[y|z, \boldsymbol{s}]p(z|\boldsymbol{s})\, dz$$
$$= \tau(\boldsymbol{s})v + \varepsilon, \quad (6)$$

which eliminates the nuisance function $\beta(\boldsymbol{s})$.

To estimate the effect, which is a general function of $\boldsymbol{s}$, we consider a flexible parametric model class

$$\tau_\theta(\boldsymbol{s}) \in \left\{f(\boldsymbol{s}) : f = \boldsymbol{\phi}^\top(\boldsymbol{s})\boldsymbol{\theta}\right\}, \quad (7)$$

where $\boldsymbol{\phi}(\boldsymbol{s})$ is a spatial basis vector described below and $\boldsymbol{\theta}$ is a $d_\theta \times 1$ parameter vector. Thus the effect is directly identifiable from the residuals given in (6).

3.2. Robust errors-in-variables approach

In practice, however, the conditional means $\mathrm{E}[y|\boldsymbol{s}]$ and $\mathrm{E}[z|\boldsymbol{s}]$ are unknown and they are replaced by empirical predictors. Thus the residuals can be written as

$$w = \underbrace{\left(y - \widehat{\mathrm{E}[y|\boldsymbol{s}]}\right)}_{\widehat{w}} + \underbrace{\left(\widehat{\mathrm{E}[y|\boldsymbol{s}]} - \mathrm{E}[y|\boldsymbol{s}]\right)}_{\widetilde{w}} \quad (8a)$$

$$v = \underbrace{\left(z - \widehat{\mathrm{E}[z|\boldsymbol{s}]}\right)}_{\widehat{v}} + \underbrace{\left(\widehat{\mathrm{E}[z|\boldsymbol{s}]} - \mathrm{E}[z|\boldsymbol{s}]\right)}_{\widetilde{v}}, \quad (8b)$$

where $\widetilde{w}$ and $\widetilde{v}$ denote errors. Let

$$\boldsymbol{\delta}(\boldsymbol{s}) = \widetilde{v}\boldsymbol{\phi}(\boldsymbol{s}) \qquad (9)$$

denote an unobservable random vector. Plugging (8) and the parametric model from (7) into (6), we obtain the following relation

$$\widehat{w} = \left(\widehat{v}\boldsymbol{\phi}(\boldsymbol{s}) + \boldsymbol{\delta}(\boldsymbol{s})\right)^{\top}\boldsymbol{\theta} + \widetilde{\varepsilon}, \qquad (10)$$

where $\widetilde{\varepsilon} = \varepsilon - \widetilde{w}$ is an unknown error. Hence using (10), it is therefore possible to estimate the effect as

$$\widehat{\tau}(\boldsymbol{s}) = \boldsymbol{\phi}^{\top}(\boldsymbol{s})\widehat{\boldsymbol{\theta}}, \qquad (11)$$

where the parameter vector $\widehat{\boldsymbol{\theta}}$ is learned from $\widetilde{\mathcal{D}}_n = \left\{ \left(\widehat{w}_i, \widehat{v}_i, \boldsymbol{\phi}(\boldsymbol{s}_i) \right) \right\}_{i=1}^{n}$. This can be done using the least-squares (LS) method, which ignores the unknown deviation $\boldsymbol{\delta}(\boldsymbol{s})$ in (10). It was shown by Chernozhukov et al. (2016) that LS nevertheless has good asymptotic properties in the case of fixed effects, that is when $\tau_\theta(\boldsymbol{s}) \equiv \theta$. The deviation $\boldsymbol{\delta}(\boldsymbol{s})$, however, yields errors that are correlated with $\boldsymbol{\theta}$ so that the LS-based estimate exhibits notable finite-sample biases.

To tackle this problem, we propose using a robust learning method that takes into account the errors-in-variables arising from the unknown deviations. Specifically, we formulate a convex minimax problem

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ \max_{\boldsymbol{\delta} \in \Delta} \sqrt{\mathrm{E}_n \left[\left| \widehat{w} - \left(\widehat{v}\boldsymbol{\phi}(\boldsymbol{s}) + \boldsymbol{\delta} \right)^{\top}\boldsymbol{\theta} \right|^2 \right]} \right\}, \qquad (12)$$

where Δ denotes an uncertainty set for a deviation $\boldsymbol{\delta}$. In this way the method mitigates the worst case deviations in the set, which we now determine on conservative statistical grounds.

When using an empirical predictor of the exposure, it is reasonable to assume that the error term $\widetilde{v}$ in (8) tends to be smaller than $\widehat{v}$, in terms of their second-order moments. Considering the unknown deviation (9), this assumption on $\widetilde{v}$ motivates a conservative upper bound on $\boldsymbol{\delta}$ as specified by the uncertainty set

$$\Delta = \left\{ \boldsymbol{\delta} \ : \ \mathrm{E}_n \left[|\delta_k|^2 \right] \ \leq \ \frac{1}{n} \mathrm{E}_n \left[|\widehat{v}\phi_k(\boldsymbol{s})|^2 \right], \ \forall k \right\}.$$

Using (12), we ensure that the estimated effect is robust against the worst deviations in the set.

To provide further insight to (12), we may reformulate the minimax problem as

$$\min_{\boldsymbol{\theta}} \sqrt{\mathrm{E}_n \left[\left| \widehat{w} - \widehat{v}\boldsymbol{\phi}^{\top}(\boldsymbol{s})\boldsymbol{\theta} \right|^2 \right]} + \sum_{k=1}^{d_\theta} \sqrt{\frac{1}{n} \mathrm{E}_n \left[|\widehat{v}\phi_k(\boldsymbol{s})|^2 \right]} |\theta_i|,$$

using Theorem 1 in Xu et al. (2009). Note that the first term is the square root of the LS criterion and the second term results in a data-adaptive regularization that mitigates overfitting and biases the estimator against spurious effects. Moreover, the problem can be solved in runtime $\mathcal{O}(nd_\theta^2)$ using the algorithm by Zachariah & Stoica (2015). This property enables the efficient computation of bootstrap confidence intervals for $\widehat{\tau}(\boldsymbol{s})$. We use the pivotal bootstrap method (Wasserman, 2013) to obtain CIs below.

3.3. Estimating residuals

The conditional means in (8) are MSE-optimal predictors of the exposure and the outcome. In lieu of these, we use predictors $\widehat{\mathrm{E}[y|\boldsymbol{s}]}$ and $\widehat{\mathrm{E}[z|\boldsymbol{s}]}$ as approximations.

Any reasonable prediction method is applicable since (12) mitigates both of the errors $\widetilde{w}$ and $\widetilde{v}$. To match the form of the spatial model in (7), we consider here predictors that are linear in the parameters so that

$$\widehat{w}(\boldsymbol{\lambda}) = y - \boldsymbol{\psi}^{\top}(\boldsymbol{s})\boldsymbol{\lambda}, \qquad (13a)$$
$$\widehat{v}(\boldsymbol{\lambda}) = z - \boldsymbol{\psi}^{\top}(\boldsymbol{s})\boldsymbol{\lambda}, \qquad (13b)$$

where $\boldsymbol{\lambda}$ denote the parameters and $\boldsymbol{\psi}(\boldsymbol{s}) = \mathrm{col}\{1, \boldsymbol{\phi}(\boldsymbol{s})\}$ for simplicity. To learn the parameters, we use here the method in Zachariah et al. (2017) for which learned parameters are given as

$$\boldsymbol{\lambda}_w = \arg\min_{\boldsymbol{\lambda}} \sqrt{\mathrm{E}_n \left[|\widehat{w}(\boldsymbol{\lambda})|^2 \right]} + \sum_{k=2}^{d_\theta+1} \sqrt{\frac{1}{n} \mathrm{E}_n \left[|\psi_k(\boldsymbol{s})|^2 \right]} |\lambda_k|,$$

$$\boldsymbol{\lambda}_v = \arg\min_{\boldsymbol{\lambda}} \sqrt{\mathrm{E}_n \left[|\widehat{v}(\boldsymbol{\lambda})|^2 \right]} + \sum_{k=2}^{d_\theta+1} \sqrt{\frac{1}{n} \mathrm{E}_n \left[|\psi_k(\boldsymbol{s})|^2 \right]} |\lambda_k|,$$

where weighted penalty term yields tuning-free regularization that mitigates overfitting in a data-adaptive manner. In addition, $\boldsymbol{\lambda}_w$ and $\boldsymbol{\lambda}_v$ can be obtained in a computationally efficient manner. Plugging in $\boldsymbol{\lambda}_w$ and $\boldsymbol{\lambda}_v$ in (13) yields $\widehat{w}$ and $\widehat{v}$, respectively.

Combining the predictors with (12) leads to a *r*obust *o*rthogonalized method for *s*patially varying *c*ausal *e*ffect estimation (ROSCE).

3.4. Spatial basis function

The form of the spatial basis vector $\boldsymbol{\phi}(\boldsymbol{s})$ in (7) depends on whether the space $\boldsymbol{s}$ is continuous or discrete.

Continuous space

In the continuous case, we consider $\mathcal{S} \subset \mathbb{R}^d$. For clarity of explanation we restrict our description here to the case when $d = 2$, but modifications to one or three dimensional space are straightforward. In two dimensions, let

$\boldsymbol{s} = \{s_1, s_2\}$ and

$$\boldsymbol{\phi}(\boldsymbol{s}) = \boldsymbol{\phi}_1(s_1) \otimes \boldsymbol{\phi}_2(s_2), \qquad (14)$$

where

$$\boldsymbol{\phi}_1(s_1) = \text{col}\{\phi_{i,1}(s_1), \dots, \phi_{i,N_s}(s_1)\},$$
$$\boldsymbol{\phi}_2(s_2) = \text{col}\{\phi_{i,1}(s_2), \dots, \phi_{i,N_s}(s_2)\}.$$

Each $\boldsymbol{\phi}_i(s_i)$ is composed of N_s components. Based on their attractive approximation and computational properties, we use the cubic b-spline functions in (15), where c determines the location of each component and L determines the function support (Wasserman, 2013).

Thus the effect $\tau(\boldsymbol{s})$ is approximated as a linear combination of b-splines placed at different points across space via the model class (7). The maximum value of N_s is practically limited by the number of data points n and the available computational resources. For 2-dimensional space with N_s components the dimension of parameter $\boldsymbol{\theta}$ in (7) is $d_\theta = N_s^2$.

The approximation accuracy of the model class introduced above is readily refined by incorporating b-splines with multiple supports. For instance,

$$\boldsymbol{\phi}(\boldsymbol{s}) = \begin{bmatrix} \boldsymbol{\phi}_{L_1}(\boldsymbol{s}) \\ \boldsymbol{\phi}_{L_2}(\boldsymbol{s}) \end{bmatrix} \qquad (16)$$

accommodates two different support sizes L_1 and L_2. This enables inferring the effects at different spatial resolutions.

Discrete space

In the discrete case, the space consists of d disjoint regions $\mathcal{S} = \{1, \dots, d\}$ and the spatial basis vector is simply

$$\boldsymbol{\phi}(\boldsymbol{s}) = \text{col}\{I(\boldsymbol{s} = 1), \dots, I(\boldsymbol{s} = d)\}, \qquad (17)$$

where $I(\boldsymbol{s} = k)$ denotes the indicator function. The above choice of $\boldsymbol{\phi}(\boldsymbol{s})$ yields a model class for with an average effect $\tau(\boldsymbol{s})$ for every discrete region $\boldsymbol{s}$.

4. Experimental Results

In this section, we demonstrate the proposed ROSCE method using simulated data for both continuous and discrete space. We subsequently apply the method to real data.

4.1. Continuous two-dimensional space

First, we illustrate ROSCE in continuous two-dimensional space $\mathcal{S} = [0, 10]^2$. The outcome is generated according to

$$y = \tau(\boldsymbol{s})z + \beta(\boldsymbol{s}) + \varepsilon,$$

where the effect of the exposure varies according to

$$\tau(\boldsymbol{s}) = \cos\left(\frac{2\pi s_1}{20} + \frac{2\pi s_2}{20}\right),$$

as shown in Figure 4a. Note that $\tau(\boldsymbol{s})$ does not belong to the model class (7), but can nevertheless be well approximated using the b-splines basis. The nuisance function is $\beta(\boldsymbol{s}) = \boldsymbol{\phi}_0^\top(\boldsymbol{s})\boldsymbol{\eta}$, where $\boldsymbol{\phi}_0(\boldsymbol{s})$ is a b-spline basis with $N_s = 10$ and support $L = 0.2 \times 10$ (10 being the range of $\boldsymbol{s}$ in each dimension), see Sec. 3.4. The unknown coefficients were drawn as $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I})$. The exposure and error are distributed as $z \sim \mathcal{N}(0.5 \times \beta(\boldsymbol{s}), 1)$ and $\varepsilon \sim \mathcal{N}(0, 0.2^2)$, respectively. We generate a dataset $\mathcal{D}_n = \{(y_i, z_i, \boldsymbol{s}_i)\}_{i=1}^n$, where $n = 676$. The samples of $\boldsymbol{s}$ were drawn uniformly across $\mathcal{S}$.

For estimation, we use a basis vector $\boldsymbol{\phi}(\boldsymbol{s})$ with $N_s = 10$. To capture multiple resolutions, we use three levels of supports $L_1 = 0.2 \times 10$, $L_2 = 0.4 \times 10$ and $L_3 = 0.85 \times 10$, cf. (16). For the sake of comparison, we consider a direct approach to estimate $\tau(\boldsymbol{s})$ using (5), which requires the joint estimation of the nuisance function $\beta(\boldsymbol{s})$. For the latter, we use the well-specified model $\boldsymbol{\phi}^\top(\boldsymbol{s})\boldsymbol{\lambda}$, which contains $\boldsymbol{\phi}_0(\boldsymbol{s})$ since $L_1 = L$. The effect and nuisance functions are jointly estimated using the LS method.

Figure 4b and 4c show the estimate $\widehat{\tau}(\boldsymbol{s})$ obtained using ROSCE and the direct approach, respectively. It can be clearly seen that the latter results in very poor effect estimates, due to the high-dimensionality of $\beta(\boldsymbol{s})$. In contrast to this, the ROSCE method approximates the true effect well. Moreover, we also evaluate 95% confidence interval (CI) using the pivotal bootstrap method (Wasserman, 2013) and 3000 bootstrap iterations for both methods. Figure 4d shows the effect $\tau(\boldsymbol{s})$ along the left diagonal with the 95% CIs of the two methods. The direct approach infers insignificant effects, both where the true effect is actually positive and as well where the true effect is negative. The inference is also inconsistent, reporting a significant effect with an opposite sign to the true effect at around $s = 6$. The CI of ROSCE, on the other hand, is consistent with the true effect.

4.2. Discrete spatial regions

Next, we illustrate ROSCE in the simple case of discrete space $\mathcal{S} = \{1, \dots, d\}$ and compare it with a naive LS approach. The outcome is generated according to

$$y = \tau(\boldsymbol{s})z + \beta(\boldsymbol{s}) + \varepsilon,$$

where the effect of the exposure varies across $d = 5$ regions as

$$\tau(\boldsymbol{s}) = \cos\left(\frac{2\pi}{2d}\boldsymbol{s}\right)$$

and the nuisance function is $\beta(\boldsymbol{s}) = 2 - \boldsymbol{s}$. The exposure and error are distributed as $z \sim \mathcal{N}(\cos\left(\frac{2\pi}{2d}\boldsymbol{s}\right), 1)$ and $\varepsilon \sim \mathcal{N}(0, 0.2^2)$, respectively. We generate a dataset $\mathcal{D}_n$, where $n = 500$ and the samples are drawn uniformly across space.

$$\phi(s) = \begin{cases} \frac{1}{6}f(s)^3 & \frac{(c-2)L}{4} \leq s < \frac{(c-1)L}{4} \\ \frac{-1}{2}f(s)^3 + 2f(s)^2 - 2f(s) + \frac{2}{3} & \frac{(c-1)L}{4} \leq s < \frac{Lc}{4} \\ \frac{1}{2}f(s)^3 - 4f(s)^2 + 10f(s) - \frac{22}{3} & \frac{Lc}{4} \leq s < \frac{(c+1)L}{4} \\ \frac{-1}{6}f(s)^3 + 2f(s)^2 - 8f(s) + \frac{32}{3} & \frac{(c+1)L}{4} \leq s \leq \frac{(c+2)L}{4} \\ 0 & \text{otherwise} \end{cases} \quad \text{where} \quad f(s) = \frac{4s}{L} - c + 2$$
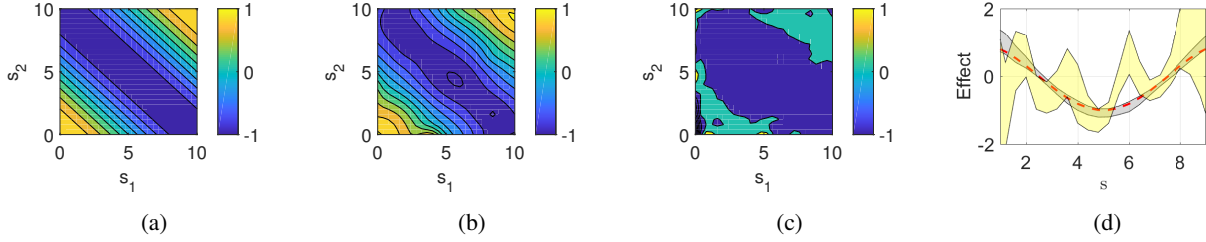
$$(15)$$

(a) (b) (c) (d)

Figure 4. (a) Exposure effect $\tau(s)$ across two-dimensional space $\mathcal{S}$ and estimated effects $\widehat{\tau}(s)$ from $\mathcal{D}_n$ using (b) ROSCE and (c) direct approach. (d) $\tau(s)$ along left diagonal (dashed line) along with 95%-CI for ROSCE (grey shaded) and direct approach (yellow shaded).

The vector $\phi(s)$ is given by (17). For the sake of comparison, we consider a standard LS regression approach, in which the predictive effect of the exposure in each region is estimated separately. Figures 5a and 5b illustrate the difference between ROSCE, which takes the causal structure into account, and a standard regression approach. The latter leads to nonnegligible biases and misleading inferences, see for instance the inferred effect in region 5.

4.3. Robustness to errors-in-variables

Finally, we illustrate the robustness of the method against the unobserved deviations $\delta(s)$. We consider one-dimensional continuous space $\mathcal{S} = [0, 10]$ with an effect parameterized as $\tau(s) = \phi^\top(s)\theta_0$, where θ_0 has dimension $d_\theta = 10$. We generate $n = 41$ samples and, for clarity, let $\widetilde{\varepsilon}$ in (10) be zero, so that the only estimation error arises via $\widetilde{v}$. The residual $v = \widehat{v} + \widetilde{v}$ in (8) is generated with independent variables

$$\widehat{v}, \widetilde{v} \sim \mathcal{N}(0, 1),$$

which yields large errors.

We compare ROSCE with an LS estimator applied to (10), which assumes $\delta \equiv 0$. An example realization of the estimates $\widehat{\tau}(s)$ is shown in Figures 6a and 6b. We see that ROSCE biases the estimates towards zero and reduces the risk of reporting spurious effects. Correspondingly, this method reports larger regions in which effects are significant (via 95%-CIs) as compared to the LS method. We repeat the experiment using Monte Carlo simulations and confirm in Figure 6c that the proposed method is less prone to report spurious effects under errors-in-variables than an
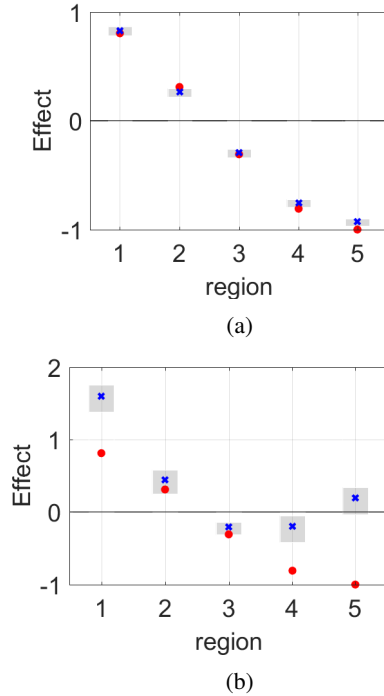

(a)

(b)

Figure 5. Effect $\tau(s)$ (dots), estimated effect $\widehat{\tau}(s)$ (crosses) and the 95%-CI (shaded) across five spatial regions using (a) ROSCE (b) standard regression approach.
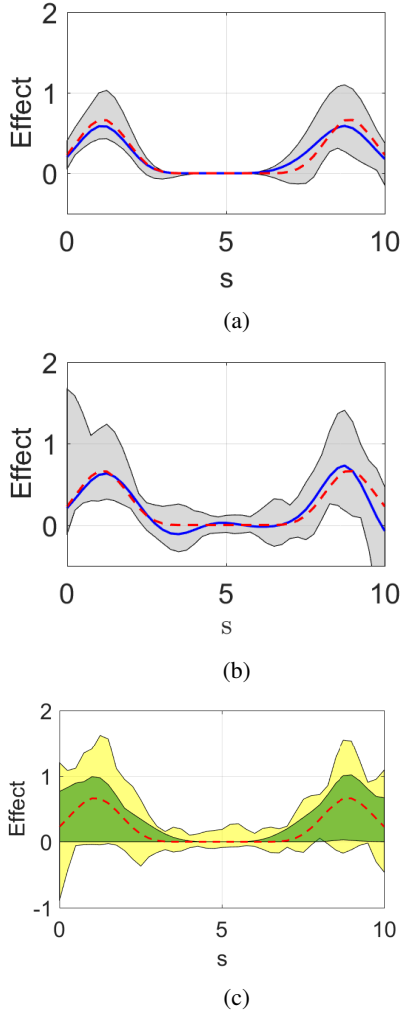
LS-based approach.

(a)

(b)

(c)

Figure 6. Illustrating robustness to errors-in-variables. True effect $\tau(s)$ (dashed line), estimate $\widehat{\tau}(s)$ (solid line) and 95%-CI (grey shaded region) using (a) ROSCE and (b) LS-based method for single dataset. (c) Dispersion of estimates using ROSCE (green shaded) and LS (yellow shaded) in terms of 5th and 95th quantiles $[Q_5, Q_{95}]$, based on 100 Monte Carlo simulations. Note where the estimates are positive, ROSCE concentrates the estimates on the positive side for all simulations in contrast to the LS method.

4.4. German income data

We proceed to illustrate the method with real datasets. In these examples, the space is partitioned into discrete regions. First, we consider the average household income in Euros at county level in Germany for the year 2012 (INKAR, 2012).

Suppose we are interested in studying the effect of age z on income y across Germany's federal states. The observed data is $\mathcal{D}_n = \{(y_i, z_i, \boldsymbol{s}_i)\}_{i=1}^{n=366}$ which spans $d = 11$ states. The data is standardized to have zero mean and unit standard deviation. For the sake of illustration, suppose

that different regions have different age demographics but also different economic structures with varying levels of unemployment. Then the unemployment level, denoted c, acts as a spatial confounder on the relation between age and income (Thaden & Kneib, 2018), see Figure 7.

Assuming that c is unknown, we use ROSCE to infer the average change in income per year of age, i.e., $\tau(\boldsymbol{s})$. The result is shown in Figure 8a, where we do not infer any significant causal effects across states. By contrast, when using a standard regression approach to infer the predictive effect in each state, we find significant negative effects of age in four states, see Figure 8b. For comparison with ROSCE, we control for county-level unemployment rates in each state by partialing out c from both z and y, and then estimating the effect from the residuals. The results in Figure 8c corroborate those reported in Figure 8a in that the effects remain insignificant when the unobserved variables are taken into account.

4.5. US Crime data

Next, we are interested in studying the effect of poverty on crime, which has been a longstanding research topic in criminology. Based on a review of 273 different studies, Ellis & McDonald (2001) concluded that there is an inverse association between an individual's socioeconomic status and criminal behavior. We use number of crimes and number of poor families data on county level from US census of year 2000 (Census Bureau, 2000) to test if our method is able infer such a causal effect. The outcome y considered is the total number of crimes due to theft, larceny and burglary. The exposure z considered is the number of families that are classified as poor. The observed data is $\mathcal{D}_n = \{(y_i, z_i, \boldsymbol{s}_i)\}_{i=1}^{n=2786}$ which spans over $d = 50$ states.

Figure 9 shows the estimated effect of number of poor families on crimes due to theft, larceny and burglary $\widehat{\tau}(\boldsymbol{s})$. It is estimated to be positive for all states, where the estimates are significant in all but three states as shown in Figure 9b. The results are thus consistent with previous findings. We find that across most states, eliminating poverty among ten families, reduces on average three crimes (of either theft, larceny and burglary).

We also consider the effect of education on income using percentage of people getting high school education and further and median annual household income data at county level from US census of year 2000 (Census Bureau, 2000). Figure $S1$ in the supplementary material shows the estimated effect $\widehat{\tau}(\boldsymbol{s})$ of education attainment percentage on annual median household income. Our method gives a positive effect which one would intuitively expect between these two quantities.
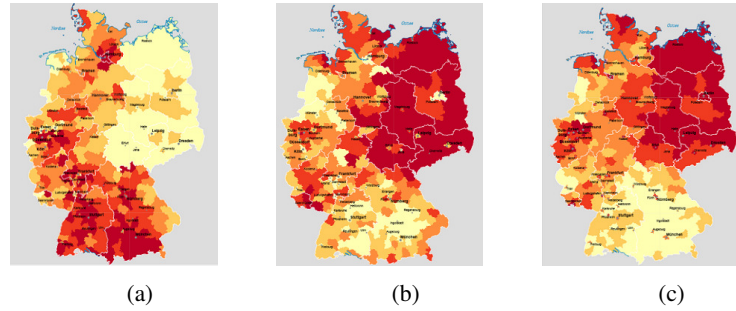
(a)   (b)   (c)

Figure 7. German data (a) Income [Euro]. (b) Age [years]. (c) Unemployment rate [%]. The dark color denotes high values while the light color denotes small values.
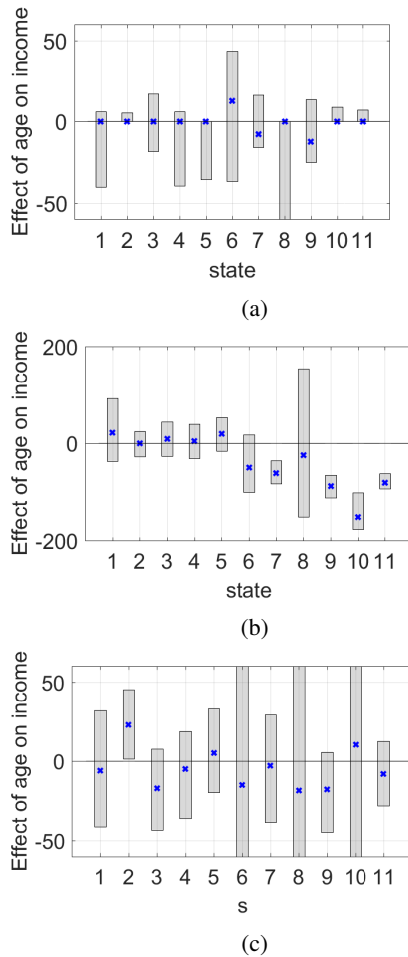

(a)

(b)

(c)

Figure 8. Inferred effect of age on median income $\hat{\tau}(s)$ (blue crosses) along with 95%-CI (shaded) for 11 different federal states in Germany. (a) ROSCE, (b) standard regression approach, (c) regression after controlling for unemployment.
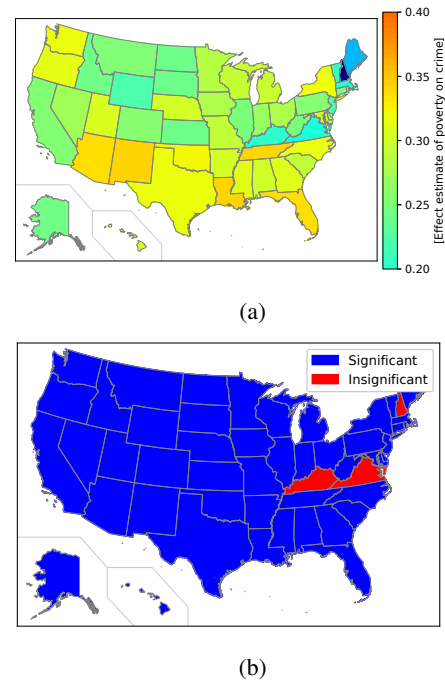

(a)

(b)

Figure 9. USA data: (a) The estimated effect $\hat{\tau}(s)$ of number of poor families on average number of crimes due to theft, larceny and burglary. (b) Significance of estimated effect $\hat{\tau}(s)$.

of possible unmeasured spatially varying confounding variables. The method is applicable to both continuous and discrete space s. The orthogonalized approach circumvents joint estimation of a possibly high-dimensional nuisance function while the robust estimation approach mitigates the resulting errors-in-variables. We demonstrated the properties of the method on synthetic data and illustrated its potential use in real-world applications. Link to code: github.

5. Conclusion

We have proposed a method for estimating heterogeneous effects of an exposure on an outcome variable in presence

References

Census Bureau. USA data, 2000. URL https://www.census.gov/prod/www/decennial.html.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. K. Double machine learning for treatment and causal parameters. Technical report, cemmap working paper, Centre for Microdata Methods and Practice, 2016.

Cressie, N. Statistics for spatial data. *Terra Nova*, 4(5): 613–617, 1992.

Ellis, L. and McDonald, J. N. Crime, delinquency, and social status: A reconsideration. *Journal of Offender Rehabilitation*, 32(3):23–52, 2001.

Hanks, E. M., Schliep, E. M., Hooten, M. B., and Hoeting, J. A. Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, 26(4):243–254, 2015.

Hernán, M. A. and Robins, J. M. *Causal inference*. Boca Raton: Chapman Hall/CRC, forthcoming, 2019.

Hodges, J. S. and Reich, B. J. Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4):325–334, 2010.

Hughes, J. and Haran, M. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):139–159, 2013.

Imai, K., Ratkovic, M., et al. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

INKAR. German income, age and unemployment data, 2012. URL http://http://www.inkar.de/.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv preprint arXiv:1706.03461*, 2017.

Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects, 2017.

Paciorek, C. J. 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):107, 2010.

Page, G. L., Liu, Y., He, Z., and Sun, D. Estimation and prediction in the presence of spatial confounding for spatial linear models. *Scandinavian Journal of Statistics*, 44 (3):780–797, 2017.

Pearl, J. *Causality*. Cambridge university press, 2009.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Reich, B. J., Hodges, J. S., and Zadnik, V. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62(4):1197–1206, 2006.

Richardson, T. S. and Robins, J. M. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

Thaden, H. and Kneib, T. Structural equation models for dealing with spatial confounding. *The American Statistician*, 72(3):239–252, 2018.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

Wasserman, L. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

Williams, C. K. and Rasmussen, C. E. Gaussian processes for machine learning. *the MIT Press*, 2(3):4, 2006.

Xu, H., Caramanis, C., and Mannor, S. Robust regression and lasso. In *Advances in Neural Information Processing Systems*, pp. 1801–1808, 2009.

Zachariah, D. and Stoica, P. Online hyperparameter-free sparse estimation method. *IEEE Trans. Signal Processing*, 63(13):3348–3359, 2015.

Zachariah, D., Stoica, P., and Schön, T. B. Online learning for distribution-free prediction. *arXiv preprint arXiv:1703.05060*, 2017.

Acknowledgements