GeoRDD manuscript

Maxime Rischard

January 4, 2018

Contents

Abstract

Regression discontinuity designs (RDDs) arise in observational studies when the treatment assignment is fully determined by covariates. Most research has focused on one-dimensional cases, where units with a "forcing" variable lying on one side of a threshold value receive a treatment that the rest do not receive. More recently, situations with multiple forcing variables have garnered interest. When these variables are specifically spatial covariates, that is when a treatment is applied to a region but not its neighbor, the resulting natural experiment is termed a geographic regression discontinuity design (GeoRDD). In this paper, we propose a framework for analysing GeoRDDs, which can be encapsulated in three steps: (1) fit a response surface to the outcomes in the adjacent regions, (2) extrapolate the two fitted surfaces to the border, and (3) take the difference of the two extrapolations to obtain an estimate of the treatment effect at each point along the border. We implement these steps by employing modeling tools from the spatial statistics literature, in particular kriging (Gaussian process regression). We present and examine several causal estimands of the average treatment effect along the border, and their corresponding estimators. We also develop methods for hypothesis testing in GeoRDDs. Finally, we illustrate our methodology on a dataset of property sales in New York City, and show evidence of a discontinuity in price between two neighboring school districts.

1 Introduction

The original theory and methods for regression discontinuity designs (RDDs) date from the 1960s, starting with (Thistlethwaite and Campbell, 1960). (Cook, 2008) tells the story of how interest in RDDs subsequently waned, until the late 1990s when the design saw renewed attention, theoretical progress, and popularity in the social sciences.

In particular beginning with (Papay et al., 2011), methods have been recently developed to analyse RDDs with multiple forcing variables. (Imbens and Lemieux, 2008) extend the non-parametric methods based on local linear regression that has become popular for analysing RDDs (Imbens and Lemieux, 2008) from settings with a single forcing variable (1D RDDs) to multiple forcing variables.

Geographical regression disctontinuity designs (GeoRDDs) are such RDDs where the forcing variables are spatial coordinates (latitude and longitude), meaning that units within a certain region are assigned to treatment, while units in a neighboring region are controls. For example, in (MacDonald et al., 2015), a private police force patrols a neighborhood, but stays out of surrounding areas, and a causal effect on crime rates is sought. In (Chen et al., 2013), a policy applies South of the Huai River in China but not in the North, and pollution levels and life expectancies are measured to infer environmental and health impacts of the policy. Practicioners often wish to use the well-established methods and software developed for 1D RDDs for their spatial problem. It is therefore tempting to reduce a GeoRDD problem to a 1D RDD by using the signed distance from the boundary (positive for treatment and negative for control) as the forcing variable, a method that we will refer to as "projected RDD," and is used by both examples cited above. However, this method fails to fully capture the spatial variation in the outcomes, resulting in spatial confoundedness of the estimator. We demonstrate this in a simple example in Appendix A.

More principled methods for analysing GeoRDDs have been proposed, though none have yet been adopted by practicioners as standard. (Keele "An Overview of Geographically Discontinuous Treatment Assignments with an Application to Children's Health Insurance", 2017) build on the methods of (Imbens and Lemieux, 2008), also using local linear regression to estimate the treatment effects at any point along the border. (Keele et al., 2015) develop a methodology to use matching for GeoRDDs, by matching units across the border while minimizing the total sum of geographic distances between them. This requires the selection of a "buffer" width from the border within which units on either side of the border are claimed to be sufficiently similar, so that the observational study on the subset of units nearest the border can be interpreted and analyzed as a randomized experiment.

When treatment assignment is dictated by thresholding a single covariate (above or below the threshold all units are assigned to treatment, and all others to control), then the methodologies developed for

1D RDDs enable the estimation of a causal effect despite the lack of overlap in the covariate. We propose a framework for analysing GeoRDDs that is analogous to their 1D counterpart. Broadly, 1D RDD methodologies are composed of three steps:

1. Fit a smooth **function** to the outcomes against the forcing variable on each side of the discontinuity;
2. Extrapolate the functions to the **discontinuity point**; and
3. Take the difference between the two extrapolations to estimate the treatment effect at the threshold point.

Reusing the same methodological skeleton and applying it to geographical RDDs, our framework proceeds analogously:

1. Fit a smooth **surface** to the outcomes against the geographical covariates in each region;
2. Extrapolate the surfaces to the **border curve**; and
3. Take the pointwise difference between the two extrapolations to estimate the treatment effect along the border.

Recently, (Branson et al., 2017) proposed a Gaussian process regression (GPR) methodology that exhibits promising coverage and MSE properties compared to local linear regression for 1D RDDs. We believe this approach to be particularly suitably to GeoRDDs, as GPR is a well-established tool in spatial statistics (where it is known as krigin) for fitting smoothly varying spatial processes. See (Banerjee et al., 2014) for a textbook introduction to kriging for spatial data, and (Rasmussen and Williams, 2006) for a machine learning perspective on GPR.

A peculiarity of GeoRDDs is the functional estimand, defined everywhere along the border. In Section 2, we use GPR to obtain an estimator for it, by extending the model of (Branson et al., 2017) to geographical settings. Section 3 explores possible estimands for the average treatment effect (ATE), and elucidates their respective pitfalls and advantages. Section 4 turns to hypothesis testing, and proposes methods to test against the null hypothesis of no treatment effect along the border. We also suggest Placebo tests [cite?] to examine the validity of the hypothesis tests. Section 5 applies our methodology to an empirical example, using a publicly available dataset of property sales in NYC to determine whether school districts affect property prices.

2 GeoRDD Modeling with Gaussian Processes

2.1 GeoRDD Setup and Notation

We largely adopt the setup and notation for geographic regression discontinuity designs laid out in (Keele and Titiunik, 2015). The outcomes Y_i of n units with spatial coordinates $\mathbf{s}_i$ are observed within an area $\mathcal{A}$ of 2-dimensional coordinate space. The units are divided into n_T treatment units in area $\mathcal{A}^T \subset \mathcal{A}$ and n_C units in the control area $\mathcal{A}^C$. The defining characteristic of the regression discontinuity design is that the two areas are adjacent but non-overlapping, so they intersect only at the border $\mathcal{B}$ between the two areas. Points on the border are always denoted by $\boldsymbol{b}$. For computational reasons, we will often represent the border as a set $\boldsymbol{b}_{1:R} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_R\}$ of R "sentinel" points along the border. In the potential outcomes framework, we imagine that each unit i has a potential outcome under treatment Y_{iT} and a potential outcome under control Y_{iC}. The unit's treatment indicator Z_i is 1 if the unit is in the treatment group, and 0 in control. Unlike traditional randomized experiments, the treatment assignment is a deterministic function of the unit's geographical coordinates $\mathbf{s}_i$, $Z_i = \mathbb{I}\left\{\mathbf{s}_i \in \mathcal{A}^T\right\}$. The observed outcome can be written as $Y_i = Z_i Y_{iT} + (1 - Z_i) Y_{iC}$.

RDDs can be understood as natural experiments, as near the discontinuity we can reasonably claim that the side of the discontinuity that each unit fell into was largely dictated by random noise in the covariate, which in turn justifies the claim that a natural randomized experiment took place near the border, with treatment and control units coming from the same population. This forms the basis of causal interpretations of RDDs, see for example chapter 3 of (Dunning, 2012). We can extend this interpretation to the spatial setting, by conceiving of multiple correlated experiments taking place all along the border.

Because the treatment and control regions do not overlap, inference on the treatment effect is only measurable near the border. In the 1D RDD with forcing variable X, the estimand is therefore defined at the threshold $X = b$:

$$\tau = \lim_{x \downarrow b} \mathbb{E}\left[Y \mid X = x\right] - \lim_{x \uparrow b} \mathbb{E}\left[Y \mid X = x\right] = \mathbb{E}\left[Y_{iT} \mid X_i = b\right] - \mathbb{E}\left[Y_{iC} \mid X_i = b\right] \tag{1}$$

where the second equality requires the assumption that the conditional regression function $\mathbb{E}\left[Y \mid X = x\right]$ be continuous in x above and below b (see Assumption 2.1 in (Imbens and Lemieux, 2008) and the discussion that follows). Analogously, we focus on the treatment effect at the border $\mathcal{B}$ between the treatment and control regions

$$\tau : \mathcal{B} \to \mathbb{R},\ \tau(\boldsymbol{b}) = \mathbb{E}\left[Y_{iT} - Y_{iC} \mid \mathbf{s}_i = \boldsymbol{b}\right] \tag{2}$$

This is also the estimand defined in (Imbens and Zajonc, 2011) and (Keele et al., 2017). Again, τ can be understood as the difference of the two limits of the expected outcomes, approaching the border from the treatment side or the region side, given the assumption that the conditional regression function $\mathbb{E}\left[Y \mid S = \mathbf{s}\right]$ is continuous in $\mathbf{s}$ within $\mathcal{A}_T$ and within $\mathcal{A}_C$ (see Assumption 2.2.2 in (Imbens and Zajonc, 2011)).

2.2 Model Specification

In this paper, we propose to use Gaussian process regression (GPR), also known as kriging in the spatial statistics literature, to fit the outcomes on either side of the border. GPR is a Bayesian non-parametric method for fitting smooth functions, that was shown by (Branson et al., 2017) to be a promising tool for fitting 1D RDDs. Further inspired by the popularity of GPR in spatial statistics, we extend the model and method of (Branson et al., 2017) to geographical RDDs.

On each side of the border, we model the observed outcomes Y_i at location $\mathbf{s}_i$ as the sum of an intercept m, a spatial Gaussian process $f(\mathbf{s})$, and iid normal noise ϵ. The Gaussian process has zero mean, and its covariance function is a modeling choice. There is a rich literature of possible covariance functions ("kernels" in the machine learning world) **[LukeB: do you have a good reference summarizing popular options?]**, but in this paper, we will use the squared exponential kernel, for its ease of understanding and its prevalence in applied spatial statistics.

$$
\begin{aligned}
Y_{iT} &= \underbrace{m_T + f_T(\mathbf{s}_i)}_{g_T(\mathbf{s}_i)} + \epsilon_i \\
Y_{iC} &= \underbrace{m_C + f_C(\mathbf{s}_i)}_{g_C(\mathbf{s}_i)} + \epsilon_i \\
f_T, f_C &\overset{\perp\!\!\!\perp}{\sim} \mathcal{GP}\left(0, k(\mathbf{s}, \mathbf{s}')\right) \\
k(\mathbf{s}, \mathbf{s}') &= \sigma^2_{GP} \exp\left(-\frac{(\mathbf{s}-\mathbf{s}')^\intercal (\mathbf{s}-\mathbf{s}')}{2\ell^2}\right)
\end{aligned}
\tag{3}
$$

The treatment effect at a location $\boldsymbol{b}$ on the border is derived as the difference between the two (noise-free) surfaces g_T and g_C

$$
\tau(\boldsymbol{b}) = \left[m_T + f_T(\boldsymbol{b})\right] - \left[m_C + f_C(\boldsymbol{b})\right] . \tag{4}
$$

This can be visualized as the height of a cliff separating the two smooth plains of the treatment and control regions.

In this specification, the parameters ℓ, σ_{GP}, and σ_ϵ are the same in the treatment and control regions, so we assume that the spatial smoothness of the responses isn't affected by the treatment. We expect that this assumption will be reasonable in many applications, but it can be easily relaxed, as discussed in (Branson et al., 2017).

2.3 Inference

If m_T and m_C are given normal priors with variance σ_μ, then the model specification (3) can be used to obtain covariances between the observations, the Gaussian processes, and the mean parameters. Given hyperparameters $(\ell, \sigma_{GP}, \sigma_\epsilon, \sigma_\mu)$, $\left(Y_i, f_T(\mathbf{s}_j), f_C(\mathbf{s}_k), m_C, m_T\right)$ is multivariate normal for any set of observations indexed by i and points on the control and treatment Gaussian processes indexed by j and k, and so the distribution of any variables conditioned on the others can be obtained analytically and easily computed.

We proceed by extrapolating both Gaussian processes to the border, and then taking the difference of the predictions to obtain the posterior treatment effect $\tau(\mathcal{B})$ along the border. Computationally, we need to represent this border as a set $\boldsymbol{b}_{1:R} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_R\}$ of R "sentinel" units dotted along $\mathcal{B}$. The extrapolation step then follows mechanically through multivariate normal theory:

$$
\begin{aligned}
g_T(\boldsymbol{b}_{1:R}) \mid Y_T, S_T, \ell, \sigma_{GP}, \sigma_\epsilon &\sim \mathcal{N}\left(\mu_{\boldsymbol{b}_{1:R}|T}, \Sigma_{\boldsymbol{b}_{1:R}|T}\right) \\
\mu_{\boldsymbol{b}_{1:R}|T} &\equiv \operatorname{cov}\left(g_T(\boldsymbol{b}_{1:R}), Y_T\right) \operatorname{cov}(Y_T)^{-1} Y_T \\
\Sigma_{\boldsymbol{b}_{1:R}|T} &\equiv \operatorname{cov}\left(g_T(\boldsymbol{b}_{1:R})\right) - \operatorname{cov}\left(g_T(\boldsymbol{b}_{1:R}), Y_T\right) \operatorname{cov}(Y_T)^{-1} \operatorname{cov}\left(Y_T, g_T(\boldsymbol{b}_{1:R})\right)
\end{aligned}
\tag{5}
$$

with all the covariance matrices derived from the model specification (see Appendix B). Analogously, predictions for $g_C(\boldsymbol{b}_{1:R})$ are obtained using the data in the control region, and their posterior mean and covariance denoted by $\mu_{\boldsymbol{b}_{1:R}|C}$ and $\Sigma_{\boldsymbol{b}_{1:R}|C}$. Since the two surfaces are modeled as independent, the treatment effect $\tau(\boldsymbol{b}_{1:R}) = g_T(\boldsymbol{b}_{1:R}) - g_C(\boldsymbol{b}_{1:R})$ has posterior

$$
\begin{aligned}
\tau(\boldsymbol{b}_{1:R}) \mid Y_T, Y_C &\sim \mathcal{N}\left(\mu_{\boldsymbol{b}_{1:R}|Y}, \Sigma_{\boldsymbol{b}_{1:R}|Y}\right) \\
\mu_{\boldsymbol{b}_{1:R}|Y} &= \mu_{\boldsymbol{b}_{1:R}|T} - \mu_{\boldsymbol{b}_{1:R}|C} \\
\Sigma_{\boldsymbol{b}_{1:R}|Y} &= \Sigma_{\boldsymbol{b}_{1:R}|T} + \Sigma_{\boldsymbol{b}_{1:R}|C}\,.
\end{aligned}
\tag{6}
$$

The posterior mean and covariance of the "cliff height" $\tau(\boldsymbol{b}_{1:R})$ are the primary output of our GeoRDD

analysis, and we refer to them as the "cliff face" estimator.

This leaves the choice of the hyperparameters ℓ, σ_{GP}, σ_ϵ, and σ_μ. For σ_μ, we arbitrarily pick a large number, so that the prior on the mean parameters is weak. The Gaussian process and noise parameters are optimized by maximizing the marginal likelihood of the observations $\mathbb{P}\left(Y_T, Y_C \mid \ell, \sigma_{GP}, \sigma_\epsilon\right)$, which is available analytically and easily computed for GPR. An alternative would be to also specify a prior on the hyperparameters, which is preferable in order to fully account for the uncertainty in the model, but fully Bayesian inference of large Gaussian process models tends to be computationally very expensive.

2.4 Handling Covariates

The Gaussian Process specification also makes it easy to incorporate a linear model on non-spatial covariates, both mathematically and computationally. The models are modified by the addition of the linear regression term $D\beta$ on the $n \times p$ matrix of covariates D. We recommend placing a normal prior $\mathcal{N}(0, \sigma_\beta^2)$ on the regression coefficients, as this preserves the multivariate normality of the model, with the simple addition of a term $\sigma_\beta^2 D^\intercal D$ to the covariance of Y.

Our model becomes

$$\begin{aligned}
Y_{iT} &= \underbrace{m_T + f_T(\mathbf{s}_i)}_{g_T(\mathbf{s}_i)} + \mathbf{d}_i^\intercal \beta + \epsilon_i \\
Y_{iC} &= \underbrace{m_C + f_C(\mathbf{s}_i)}_{g_C(\mathbf{s}_i)} + \mathbf{d}_i^\intercal \beta + \epsilon_i \\
f_T, f_C &\overset{\perp\!\!\!\perp}{\sim} \mathcal{GP}\left(0, k(\mathbf{s}, \mathbf{s}')\right) \\
k(\mathbf{s}, \mathbf{s}') &= \sigma_{GP}^2 \exp\left(-\frac{(\mathbf{s}-\mathbf{s}')^\intercal (\mathbf{s}-\mathbf{s}')}{2\ell^2}\right) \\
\beta_j &\overset{\perp\!\!\!\perp}{\sim} \mathcal{N}\left(0, \sigma_\beta^2\right) \text{ for } j = 1, 2, \ldots, p
\end{aligned} \tag{7}$$

Unfortunately, the linear term induces a covariance between the treatment and control region, which quadruples the computational cost of the analysis. When the two regions are independent, fitting the Gaussian processes required the inversion of an $n_T \times n_T$ covariance matrix, and of an $n_C \times n_C$ matrix. But with the additional covariates, the covariance of Y is no longer block diagonal. Thus the inversion of an $(n_T + n_C) \times (n_T + n_C)$ is now required. Matrix inversion algorithms generally have computational complexity $O(n^3)$. Therefore, if the units are evenly split between the two regions, the overall complexity of the model fitting increases fourfold.

3 Average Treatment Effect

Once we obtain the posterior on the treatment effect function $\tau(\mathcal{B})$, estimating the average treatment effect (ATE) along the border will often be of interest. We consider the class of weighted means of the functional treatment effect $\tau(\boldsymbol{b})$, with weight function $w_{\mathcal{B}}(\boldsymbol{b})$ defined everywhere on the border $\mathcal{B}$. The weighted mean integral can be approximated as a weighted sum at the sentinels $\boldsymbol{b}_{1:R}$:

$$\begin{aligned} \tau^w &= \frac{\oint_{\mathcal{B}} w_{\mathcal{B}}(\boldsymbol{b})\tau(\boldsymbol{b})\,\mathrm{d}\boldsymbol{s}}{\oint_{\mathcal{B}} w_{\mathcal{B}}(\boldsymbol{b})\,\mathrm{d}\boldsymbol{b}}, \\ &\approx \frac{\sum_{r=1}^{R} w_{\mathcal{B}}(\boldsymbol{b}_r)\tau(\boldsymbol{b}_r)}{\sum_{r=1}^{R} w_{\mathcal{B}}(\boldsymbol{b}_r)}. \end{aligned} \tag{8}$$

Note that the approximation assumes that the sentinels are evenly spaced, otherwise each term in the sum needs to be re-weighted by the length of the border that the sentinel occupies. We have shown the posterior distribution of $\tau(\boldsymbol{b}_{1:R})$to be multivariate normal, with mean $\mu_{\boldsymbol{b}_{1:R}|Y}$ and covariance $\Sigma_{\boldsymbol{b}_{1:R}|Y}$ given in (6). Since τ^w is a linear transformation of $\tau(\boldsymbol{b}_{1:R})$, its posterior is also multivariate normal, with mean $\mu_{\tau^w|Y}$ and covariance $\Sigma_{\tau^w|Y}$ given by

$$\begin{aligned} \mu_{\tau^w|Y} &= \frac{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^\intercal \mu_{\boldsymbol{b}_{1:R}|Y}}{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^\intercal \mathbf{1}_R} \\ \Sigma_{\tau^w|Y} &= \frac{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^\intercal \Sigma_{\boldsymbol{b}_{1:R}|Y} w_{\mathcal{B}}(\boldsymbol{b}_{1:R})}{\left(w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^\intercal \mathbf{1}_R\right)^2} \end{aligned} \tag{9}$$

where $w_{\mathcal{B}}(\boldsymbol{b}_{1:R})$ is the R-vector of weights evaluated at the sentinels, and $\mathbf{1}_R$ is an R-vector of ones. For each estimator obtained in (9) as a weighted mean of $\mu_{\boldsymbol{b}_{1:R}|Y}$, we consider the "natural" estimand to be the same weighted mean applied to the truth $\tau(\mathcal{B})$, given by (8).

An alternative perspective on these estimators is given by the weights induced on the observations. Indeed, combining equations (5), (6), and (9), we obtain that the posterior mean of τ^w is a linear combination

$$\mathbb{E}\left(\tau^w \mid Y\right) = w_T^\intercal Y_T + w_C^\intercal Y_C \tag{10}$$

of the observed data, with "unit weights" given by

$$
\begin{aligned}
w_T &= \frac{1}{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^\intercal \mathbf{1}_R} \operatorname{cov}(Y_T)^{-1} \operatorname{cov}\left(Y_T, g_T(\boldsymbol{b}_{1:R})\right) w_{\mathcal{B}}(\boldsymbol{b}_{1:R}), \text{ and} \\
w_C &= -\frac{1}{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^\intercal \mathbf{1}_R} \operatorname{cov}(Y_C)^{-1} \operatorname{cov}\left(Y_C, g_C(\boldsymbol{b}_{1:R})\right) w_{\mathcal{B}}(\boldsymbol{b}_{1:R}),
\end{aligned}
\tag{11}
$$

for treatment and control units respectively.

The question remains: what is the most appropriate choice of weights? In this section, we motivate and consider six possibilities choices of $w_{\mathcal{B}}(\boldsymbol{b})$, and explore interpretations, advantages, and disadvantages. A summary of their properties is provided in Table 2.

3.1 Uniform ATE

The simplest choice is uniform weights $w_{\mathcal{B}}(\boldsymbol{b}) = 1$, a seemingly reasonable and unopinionated decision. We estimate τ^{UNIF}, the uniformly weighted mean of $\tau(\mathcal{B})$, by averaging the entries of the mean posterior at the sentinels. Following (8) and (9):

$$
\begin{aligned}
\tau^{\text{UNIF}} &\equiv \frac{\oint_{\mathcal{B}} \tau(x)\, ds}{\oint_{\mathcal{B}} dx} \\
\tau^{\text{UNIF}} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell &\sim \mathcal{N}\left(\mu_{\tau^{\text{UNIF}}|Y}, \Sigma_{\tau^{\text{UNIF}}|Y}\right) \\
\mu_{\tau^{\text{UNIF}}|Y} &= \left(\mathbf{1}^\intercal \mu_{\boldsymbol{b}_{1:R}|Y}\right)/R \\
\Sigma_{\tau^{\text{UNIF}}|Y} &= \left(\mathbf{1}^\intercal \Sigma_{\boldsymbol{b}_{1:R}|Y} \mathbf{1}\right)/R^2
\end{aligned}
\tag{12}
$$

The uniformly weighted estimand takes on a geometric interpretation: equal-length segments of the border are given equal weight. But the choice of uniform weights suffers from issues that we describe and address in the next two sections.

3.2 Density weighted ATE

With uniform border weights, parts of the border adjoining dense populations are given equal weights to those in sparsely populated areas. But if the border goes through an unpopulated area, like a lake or a public park, then the treatment effect there has little meaning and importance. Furthermore, $\tau(\boldsymbol{b})$ in those empty areas will have large posterior variances, which will dominate the posterior variance of τ^{UNIF}, potentially jeopardizing the successful detection of otherwise strong treatment effects.

We can address this issue by weighting the treatment effect at each sentinel location by the local density.

That is we choose $w_{\mathcal{B}}(\boldsymbol{b}) = \rho(\boldsymbol{b})$, where ρ is the local population density. The resulting estimand τ^{ρ} also has an attractive interpretation as population-based rather than geometry-based. It gives equal weights to units of the superpopulation who live on the border rather than to lengths of the border, and it therefore better captures the "typical" treatment effect received by a unit. This is the estimand used by (Keele and Titiunik, 2015), who themselves follow in the footsteps of (Imbens and Zajonc, 2011).

In practice, the local density needs to be estimated. A simple kernel density estimator can be used, though one could also deploy a more sophisticated spatial point process model. Strictly speaking, the uncertainty of the local density estimate should then be propagated to the estimate of τ^{ρ}, which may therefore no longer have a normally distributed or analytically tractable posterior. These inconveniences certainly reduce the appeal of the density-weighted estimator, but there is a deeper issue affecting this choice of estimand.

3.3 Inverse variance weighted ATE

The unweighted and density-weighted mean treatment estimands are subtly affected by the shape of the border between the treatment and control regions, giving higher weight to wigglier sections of the border. We illustrate this with the border separating two American States: Louisiana and Mississippi. From North to South, the border follows the meandering Mississippi river, then takes a sharp turn to the East and becomes a straight line, until it meets the even more sinuous Pearl river, which it then follows until it reaches the Gulf of Mexico. Sentinels placed at equal distance intervals along this border will therefore be more densely packed along the rivers, and sparsest along the straight segment (see Figure 1). When averaging a function over the border, those sections will therefore be overrepresented. Troublingly, the sinuousness of the border therefore determines the estimand, even though the outcomes of interest will generally have nothing to do with river topologies. In population terms, the result is that units near wigglier segments receive more weight. Worse, the resolution of the map used in the analysis affects the estimated ATE.

This unwelcome dependence of the τ^{UNIF} estimand on the border topology is a symptom of the geometry of the problem: the 1-dimensional treatment function $\tau(\mathcal{B})$ is embedded in a Euclidean 2-dimensional space. The dependencies induced by this geometric fact are reflected in the covariance $\Sigma_{\boldsymbol{b}_{1:R}|Y}$: sentinels in the straight segment of the border will be less strongly correlated than in the sinuous segments. The more correlated sentinels individually carry less information about the local treatment effect. Instead of averaging the posterior treatment effect along the border based on geometry or population, we consider averaging the information contained therein. This motivates the inverse-variance weighted mean τ^{INV}:

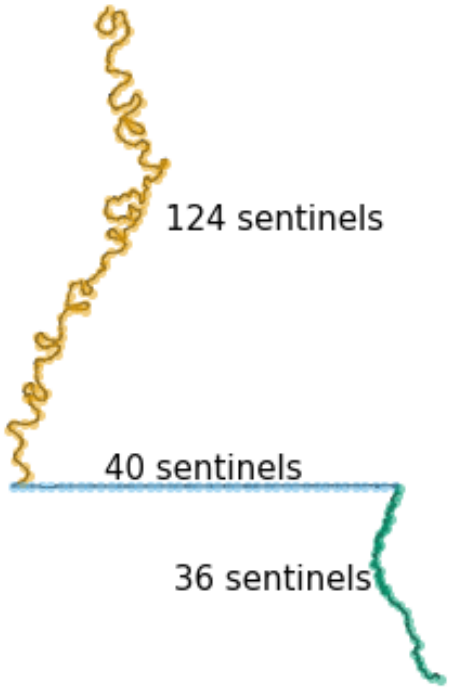

Figure 1: Evenly spaced sentinels along the border between Mississippi and Louisiana.

$$\tau^{\text{INV}} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell \sim \mathcal{N}\left(\mu_{\tau^{\text{INV}}|Y}, \Sigma_{\tau^{\text{INV}}|Y}\right),$$
$$\mu_{\tau^{\text{INV}}|Y} = \left(\mathbf{1}^\intercal \Sigma^{-1}_{\boldsymbol{b}_{1:R}|Y} \mu_{\boldsymbol{b}_{1:R}|Y}\right) / \left(\mathbf{1}^\intercal \Sigma^{-1}_{\boldsymbol{b}_{1:R}|Y} \mathbf{1}\right), \tag{13}$$
$$\Sigma_{\tau^{\text{INV}}|Y} = 1 / \left(\mathbf{1}^\intercal \Sigma^{-1}_{\boldsymbol{b}_{1:R}|Y} \cdot \mathbf{1}\right)$$

This estimator efficiently extracts the information from the posterior treatment effect, as it minimizes the posterior variance amongst weighted averages of the form (8). It automatically gives more weight to sentinels in dense areas (as the variance will be lower there), and to sentinels in straight sections of the border (as the correlations between sentinels will be lower).

The estimand is still a weighted mean, with weights for the sentinels given by $w_{\mathcal{B}}(\boldsymbol{b}_{1:R}) = \Sigma^{-1}_{\boldsymbol{b}_{1:R}|Y} \mathbf{1}$. This can put negative weights on some sentinels, as seen in a simulated example in Figure 3(c), and generally this estimand doesn't lend itself to an intuitive interpretation. This estimand isn't chosen on scientific grounds, but rather it is dictated by the observed data. This is counter to the conventional wisdom in causal inference, that the estimand should be chosen based on substantive grounds, ideally before collecting any data. While perhaps unorthodox, analogous "estimands of convenience" have been proposed in other settings, for example matching methods that exclude some unmatched units from the analysis (discussed in (Crump 2009)), or (Li et al., 2016) in the context of balancing treatment and control populations with little overlap in their covariate distributions. The classical RDD could be said to provide another example, as the estimand (1) focuses on the treatment effect near the threshold not because those units are of particular substantive interest, but because the available data restricts estimation of the treatment effect elsewhere.

3.4 Projected Finite-population ATE

All average treatment effect estimators considered so far presuppose evenly spaced sentinel points, which are then given weights. Alternatively, we can project the positions of treatment and control units onto the border, and use those projected sentinel positions without weights. For any points $\mathbf{s}$, we use the notation $\text{proj}_{\mathcal{B}}(\mathbf{s})$ to give the coordinates of the point on the border $\mathcal{B}$ that is closest to $\mathbf{s}$ (assuming uniqueness). The projected finite-population τ^{PROJ} is then the uniformly weighted mean applied with the projected sentinels instead of the evenly spaced sentinels. We can therefore modify (12), replacing the cliff face mean vector $\mu_{\boldsymbol{b}_{1:R}|Y}$ and covariance matrix $\Sigma_{\boldsymbol{b}_{1:R}|Y}$ with equivalent quantities obtained at the projected sentinels, to obtain the posterior mean and covariance of τ^{PROJ}:

$$\tau^{\text{PROJ}} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell \sim \mathcal{N}\left(\mu_{\tau^{\text{PROJ}}|Y}, \Sigma_{\tau^{\text{PROJ}}|Y}\right),$$

$$\mu_{\tau^{\text{PROJ}}|Y} = \frac{1}{n_C + n_T} \sum_{i=1}^{n_T+n_C} \mathbb{E}\left[\tau\left(\text{proj}_{\mathcal{B}}\left(\mathbf{s}_i\right)\right) \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell\right],$$

$$\Sigma_{\tau^{\text{PROJ}}|Y} = \frac{1}{(n_C + n_T)^2} \sum_{i=1}^{n_T+n_C} \sum_{j=1}^{n_T+n_C} \text{cov}\left[\tau\left(\text{proj}_{\mathcal{B}}\left(\mathbf{s}_i\right)\right), \tau\left(\text{proj}_{\mathcal{B}}\left(\mathbf{s}_j\right)\right) \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell\right]. \tag{14}$$

The posterior expectations and covariances in (14) can be obtained as in (6), but using the projected sentinels. Note that τ^{PROJ} is in the class of weighted mean estimands (8), with weight function $w_{\mathcal{B}}(\boldsymbol{b}) = \sum_{i=1}^{n_T+n_C} \delta\left(\boldsymbol{b} - \text{proj}_{\mathcal{B}}(\mathbf{s}_i)\right)$, where δ is the Dirac delta function.

The resulting estimator has desirable properties: densely populated regions receive proportionately more sentinels, but wigglier segments of the border do not. While it lacks the information efficiency of the inverse-variance estimator, the projected estimand is easier to understand and interpret, and may feel more familiar to practicioners used to finite-population inference. The averaging is over the observed units, although with their locations projected to the border.

If there are units very far away from the border, their location may be deemed irrelevant for the purposes of the analysis of a regression discontinuity design. In that case, only those units within a certain distance of the border (e.g. one or two lengthscales of the fitted Gaussian process) may be projected onto the border. Note that this only affects the location of sentinels on the border, the Gaussian process always gives low unit weights (10) to units far away from the border.

3.5 Projected land ATE

In certain applications, estimands that depend on the position of measurements are undesirable, and geography-weighted estimands are more natural. See (Antonelli 2016) for a discussion of this distinction in the context of preferential sampling. **[LukeB: Do you agree with these two sentences?]** Remember that the "geometry-based" estimand τ^{UNIF} places uniform weights along the border. Instead, the "geography-based" projected land ATE estimand τ^{GEO} begins by placing uniform weights on the treatment and control regions $\mathcal{A}_T$ and $\mathcal{A}_C$, but then projects the regions onto the border $\mathcal{B}$ to derive border weights. In other words, the projection method from τ^{PROJ} is applied to an infinite population of uniform density on both sides of the border, instead of the finite population of observed units.

To estimate τ^{GEO}, a tight grid of evenly spaced points is first generated within $\mathcal{A}^T$ and $\mathcal{A}^C$. Each point

on this grid is then projected onto the border to become a sentinel. The treatment effect at these positions is then estimated as before, yielding a mean vector and covariance matrix akin to (6). The mean of the mean vector then gives an estimate of τ^{GEO}. In other words, τ^{GEO} is estimated by applying the τ^{UNIF} procedure with sentinels obtained by projecting the grid points, instead of equispaced sentinels. τ^{GEO} remains in the category of weighted-mean estimands, with the weight function $w_{\mathcal{B}}(\boldsymbol{b})$ in (8) proportional to the area of $\mathcal{A}^{\text{T}}$ and $\mathcal{A}^{\text{C}}$ that $\mathbf{s}$ is nearest to:

$$w_{\mathcal{B}}(\boldsymbol{b}) = \int_{\mathcal{A}} \mathbb{I}\left\{\boldsymbol{b} = \text{proj}_{\mathcal{B}}\left(\mathbf{s}'\right)\right\} \mathrm{d}\mathbf{s}' \tag{15}$$

[**Note:** this integral is not strictly correct, since the area closest to a point on the border can be infinitesimal (for example in the case of a straight border). Suggestions for better notation welcome.]

Again, if land far away from the border is deemed irrelevant to the analysis, the grid should be restricted to within a certain distance of the border. This can be achieved in GIS software by obtaining a buffer around the border, then intersecting the resulting polygon with the grid points.

3.6 Projected superpopulation ATE

Lastly, the purely geographical estimand τ^{GEO} can be modified by weighing the grid points by the population density at that location. This gives the projected superpopulation ATE τ^{POP}. Similarly to the density-weighted ATE τ^{ρ}, estimating τ^{POP} requires an estimate of the density $\rho(\mathbf{s})$ at every point covered by the grid. Strictly speaking, the uncertaining in the estimate of ρ should be propagated to the estimate of τ^{POP}, which generally will make the posterior distribution of τ^{POP} neither normal nor analytically tractable.

3.7 Wiggly Border Simulation

We illustrate the above ATE estimators with a simulation. 200 units are placed in a square area delimited by spatial coordinates $S_1 \in \{0, 2\}$ and $S_2 \in \{-1, 1\}$. A border at $S_2 = 0$ divides units vertically into a control and treatment region, which are then further divided horizontally at $S_1 = 0.5$ and $S_1 = 1.5$ into three bands:

- The leftmost band $S_1 < 0.5$ has a weak treatment effect.
- The middle band $0.5 \geqslant S_1 < 1.5$ has a much lower population density, and a stronger treatment effect.
- The rightmost band $S_1 \geqslant 1.5$, has a much higher population density, and a very strong treatment effect.

Furthermore, the border in the leftmost band is a triangular wave, to create "wiggliness." We increase the number of wiggles from 0 to 1000 to observe the effect on the estimates. The simulation setting is

summarized in Table 1. We draw a single set of spatial coordinates, shown in Figure 2(a), then draw 10,000 simulations of the outcomes Y from a Gaussian process with squared exponential kernel ($\ell = 0.4$, $\sigma = 0.5$). To units above the border we add a treatment effect $\tau(S_1, S_2) = S_1$.

Table 1: Summary of wiggly border simulation setup.

	Left $s_1 < 0.5$	Middle $0.5 \geqslant s_1 < 1.5$	Right $1.5 \geqslant s_1$
Border	wiggly	straight	straight
Density	low $\rho = 1.0$	very low $\rho = 0.3$	high $\rho = 2.0$
$\boldsymbol{\tau}$	weak	medium	strong

We fit the Gaussian process model (3), using the known hyperparameters of the covariance kernel and a weak prior on the mean parameter of each region, and estimate the average treatment effect using the six methods proposed above. For each estimator, we show in Figure 2(b) the estimand and average posterior mean estimate evolving as the number of border wiggles increases. The behavior of the posterior standard deviation is shown in Figure 2(c), and is the same for every simulation as it does not depend on the outcomes.

As the border is a straight line and $\mathcal{A}^T$ and $\mathcal{A}^C$ are rectangles, and as the treatment effect does not depend on the vertical axis S_2, the density-weighted estimand τ^ρ equals the projected superpopulation estimand τ^{POP}, and they are in fact both equal to the infinite-population average treatment effect. Correspondingly, the posteriors of τ^ρ and τ^{POP} are identical. With 200 units, τ^{POP} and the finite-population projected ATE τ^{PROJ} are also similar, but the latter has the advantage of not require estimating the population density.

The geometry- and geography-based ATE τ^{UNIF} and τ^{GEO} are also equivalent when the border is a straight line. They give equal weight to the sparsely populated middle band, which produces a lower estimate with higher variance than the posteriors of τ^ρ and τ^{POP}.

Lastly, the information-based inverse-variance estimand τ^{INV} does not coincide with any others. The estimand and mean estimate change slightly from 0 to 1 wiggles, but remains stable thereafter, demonstrating the robustness of this estimator to border topology. Weighting by the inverse variance gives the lowest posterior variance within the class of ATEs under consideration, which can indeed be seen in Figure 2(c).

As we introduce wiggles into the leftmost band, τ^ρ and τ^{UNIF} show their susceptibility to the border topology. Proportionally more sentinels are packed into the leftmost section of the border, upweighting the lower treatment effect of that band, and resulting in a drop of the two estimates and estimands. Meanwhile, τ^{INV} remains stable despite the wiggles, because the additional sentinels in the leftmost band get automat-

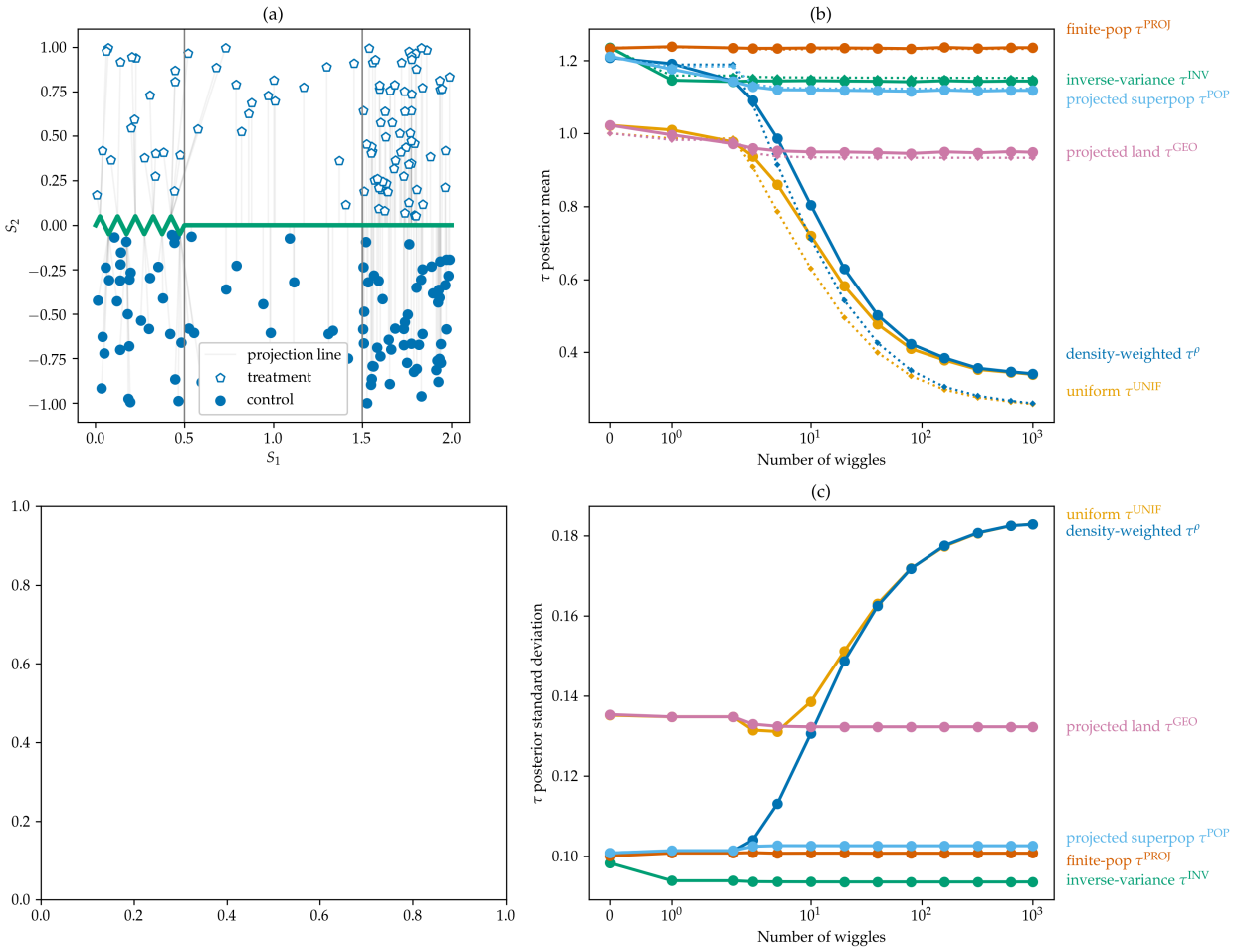

Figure 2: For a single simulation: (a) spatial positions of units and border; (b) estimator (posterior mean) and estimand (dotted lines of same color) behavior as left border gets wigglier; and (c) posterior standard deviation for each estimator as left border gets wigglier. Note that this only shows the results of a single simulation, so the fact that the posterior means are above the estimands is not indicative of bias. **(no plot in bottom left yet, final layout TBD)**

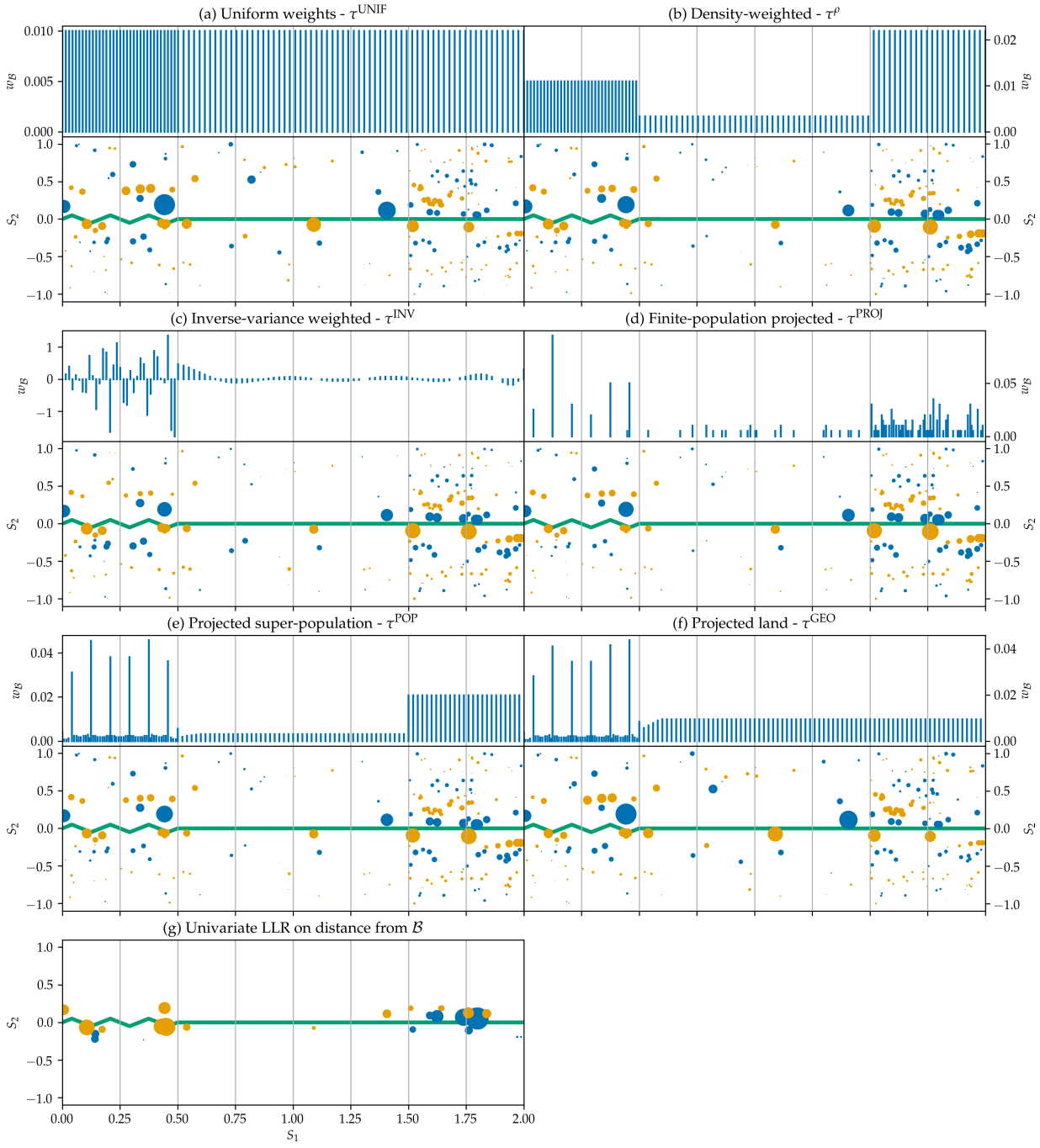

Figure 3: Weight functions and induced weights on the observations for the six weight functions proposed in this paper. The weight function plots show the weight $w_{\mathcal{B}}(\boldsymbol{b})$ against each sentinel's S_1 coordinate. Sentinels with equal S_1 positions were merged and their weights summed. The induced weight plots show a circle for each unit, with the area of the circle proportional to its weight (w_{T} and w_{C}), and colored in blue for positive weights and orange for negative weights.

ically downweighted as their correlation rises. The estimators that rely on projection τ^{PROJ}, τ^{GEO}, and τ^{POP} also remain stable, because the projected sentinels hardly move. These robust estimands show only a slight displacement when the first wiggles are introduced, caused by the presence of some sentinels nearer to the observed units.

In Figure 3, we illustrate the behavior of border weights $w_{\mathcal{B}}(\boldsymbol{b})$ and unit weights (w_T and w_C) in this simulation setting with 3 wiggles. Note that all estimators can give some small negative weights to treatment units, and small positive weights to control units. For Gaussian processes, this can be understood in terms of the negative side-lobes of the equivalent kernel (see (Rasmussen and Williams, 2006) Section 2.6). For local linear regression, it results from the negative influence on the prediction $\widehat{y^*}$ at x^* that univariate linear regression can give to an observation Y_i at X_i sufficiently far away on the opposite side of the mean $\overline{X}$ of all observations. The high variance of τ^{UNIF} and τ^{GEO} manifests itself as large weights given to a small number of units. All other estimators spread the weights more evenly amongst the units near the border, which reduces their variance. We also see how more sentinels are packed into the leftmost area because of the zig-zagging border. The inverse-variance weighted estimator border weights can be seen to respond to this change in the border topology, though it is difficult to interpret their oscillating behavior. While these border-weights look unreasonable and unstable, the induced unit weights are actually well-behaved, and in fact quite similar to those of the projected finite- and infinite-population estimators.

The weights placed on units by the projected RDD are shown in Figure 3(g). A triangular kernel in S_2 was used with bandwidth chosen using the MSE-minimizing method proposed by (Imbens and Kalyanaraman, 2012). By construction, the weights drop to zero outside of the support of the kernel. Strikingly, almost all of the positive weights are given to units in the rightmost treatment area that are closest to the border, and almost all the negative weights are given to units in the leftmost control area. Consequently, any trend in the outcomes across S_1 would confound the estimated treatment effect.

In most applications, we recommend the use of the finite population or inverse-variance-weighted estimators, to prevent the undesirable influence of border topology. The projected finite population method is simplest to understand and interpret in the tradition of finite population estimators, and unlike the projected superpopulation estimator τ^{POP} it does not require estimating population density. Meanwhile, the inverse-variance estimator is the most efficient (lowest posterior variance) weighted mean estimator, and avoids the potential complication of the choice of a distance cutoff for projected units.

Table 2: Summary of average treatment effect estimtor and estimand properties.

Symbol	Description	Border Topology	Sentinels	Principle	Variance
τ^{UNIF}	Uniform ATE	Sensitive	Equispaced	Geometry-based	High
τ^{ρ}	Density-weighted ATE	Sensitive	Equispaced	Population-based	Low
τ^{INV}	Inverse-variance weighted ATE	Robust	Equispaced	Information-based	Lowest
τ^{PROJ}	Projected finite population ATE	Robust	Projected Units	Finite-population	Low
τ^{GEO}	Projected land ATE	Robust	Projected Grid	Geography-based	High
τ^{POP}	Projected superpopulation ATE	Robust	Projected Grid	Population-based	Low

4 Testing for non-zero effect

Once we have obtained the "cliff face" estimate (6) and estimated an average treatment effect, we might also naturally wonder whether we can claim to have detected a significant treatment effect at the border. In the hypothesis testing framework, we have two possible choices of null hypotheses. The **sharp null** specifies that the treatment effect is zero everywhere along the border: $\tau(\mathcal{B}) = 0$, while the **weak null** only requires the average treatment effect to be zero.

4.1 Using the inverse-variance weighted mean treatment effect posterior to test the weak null hypothesis

As we saw in the previous section, the "average" treatment effect can be defined in multiple ways. If we choose the inverse-variance weighted mean, then τ^{INV} has posterior given by (13). While the posterior is a Bayesian object, we can use it heuristically to derive a pseudo-p-value

$$
\begin{aligned}
Z_0 &\sim \mathcal{N}\left(0, \Sigma_{\tau^{\text{INV}}|Y}\right) \\
p^{\text{INV}} &= \mathbb{P}\left(|Z_0| > \left|\mu_{\tau^{\text{INV}}|Y}\right|\right) \\
&= 2\Phi\left(-\frac{\left|\mu_{\tau^{\text{INV}}|Y}\right|}{\sqrt{\Sigma_{\tau^{\text{INV}}|Y}}}\right)
\end{aligned}
\tag{16}
$$

This "p-value" obtained from the Bayesian posterior may not have good frequentist properties. In particular, there is no guarantee that under the null hypothesis, p^{INV} is below 0.05 less than 5% of the time. We elaborate on this below after we demonstrate the behavior of this test in a simulated example.

4.2 Likelihood-based sharp null test

We can also target the sharp null hypothesis. We first create a null model $\mathcal{M}_0$, specified as a single Gaussian process spanning the control and treatment regions, with the same kernel and hyperparameters obtained in the 2GP procedure. $\mathcal{M}_0$ is smooth and continuous at the border, and therefore accords with the sharp null hypothesis. Intuitively, if there is a treatment effect, the likelihood of the observations should be lower under $\mathcal{M}_0$ than under $\mathcal{M}_1$, the $2\mathcal{GP}$ model as specified in equation (3). We therefore choose the difference in log-likelihoods as our test statistic

$$
t = \log \mathbb{P}\left(Y_T, Y_C \mid \mathcal{M}_1\right) - \log \mathbb{P}\left(Y_T, Y_C \mid \mathcal{M}_0\right) \tag{17}
$$

and wish to reject the sharp null hypothesis when its observed value t_{obs} is high.

A parametric bootstrap approach is used to quantify what "high" means. We draw ${Y_T}^*, {Y_C}^*$ from $\mathcal{M}_0$, using the same spatial locations as in the original data, and then fit the two competing models to the simulated data in order to obtain the bootstrapped test statistic

$$
t^* = \log \mathbb{P}\left({Y_T}^*, {Y_C}^* \mid \mathcal{M}_1\right) - \log \mathbb{P}\left({Y_T}^*, {Y_C}^* \mid \mathcal{M}_0\right) \tag{18}
$$

Repeating this procedure, we obtain a distribution of t under $\mathcal{M}_0$, which we can then compare to the observed t. More precisely, we can interpret the proportion of t^* drawn above t_{obs} as a p-value.

$$
p^{\text{lik}} = \mathbb{P}\left(t^* > t_{obs} \mid \mathcal{M}_0\right) \tag{19}
$$

Computationally, because the hyperparameters and locations of the units are held constant during the

bootstrap, we can reuse the Cholesky decomposition of the covariance matrix, allowing the test to be performed in seconds even with hundreds of units and thousands of bootstrap samples.

4.3 "chi-squared" test for the sharp null

The likelihood-based sharp null above is valid and easy to understand. But it may seem odd that the test aims to detect a non-zero treatment effect at the border, without any explicit reference to the border $\mathcal{B}$. The test statistic and p-values can be computed without access to the sentinel positions, using only the treatment and control indicators. If the test is significant, there is no guarantee that this is due to a discontinuity at the border.

To address this oddity, we can derive a test statistic directly from the cliff face estimator (6). We will use μ and Σ as shorthand for the posterior mean $\mu_{b_{1:R}|Y}$ and covariance matrix $\Sigma_{b_{1:R}|Y}$ throughout this section. If a k-vector y is distributed $\mathcal{N}(\mu, \Sigma)$, with mean vector μ unknown and covariance Σ known, then under the null hypothesis that $\mu = 0$, the test statistic $y^\intercal \Sigma^{-1} y$ has distribution χ^2_k. See for example (Rencher, 2003) Section 5.2.2 for a classical derivation of this test. This suggests that we could use $S^2 = \mu^\intercal \Sigma^{-1} \mu$ as a test statistic, and obtain a p-value from a χ^2_R distribution function evaluated at S^2, where R is the number of sentinels. However, we face two problems. Firstly, this test obtained heuristically from a Bayesian posterior, by analogy with the classical multivariate normal result, is not a valid frequentist test. Secondly, while Σ is mathematically full-rank, it is typically numerically rank-deficient. Therefore, R overestimates the true degrees of freedom of the null distribution.

Benavoli and Mangili (2015), developing a test for function equality, address the second problem by trimming the Σ eigenvalues λ_i lower than $\epsilon \sum_{j=1}^{k} \lambda_j$, with ϵ a pre-specified small number (they use 0.01). They address the first problem by showing that the resulting p-value is always conservative in their simulations. However, in our work, we found the resulting p-value to be sensitive to the arbitrarily chosen ϵ tolerance parameter, which makes it difficult to trust its validity.

We therefore again take the parametric bootstrap approach, this time using S^2 as the test statistic instead of the likelihood ratio. With B bootstrap samples, the p-value is obtained as

$$\begin{aligned} p &= \frac{1}{B} \sum_{t=1}^{T} \mathbb{I}\left\{ S^2_{(b)} < S^2 \right\}, \\ S^2_{(b)} &= \left(\mu_{(b)}\right)^\intercal \Sigma^{-1} \mu_{(b)} \end{aligned} \tag{20}$$

where $\mu_{(b)}$ is the result of applying (6) to $Y_T^{(b)}$ and $Y_C^{(b)}$, themselves drawn from $\mathcal{M}_0$ at the same loca-

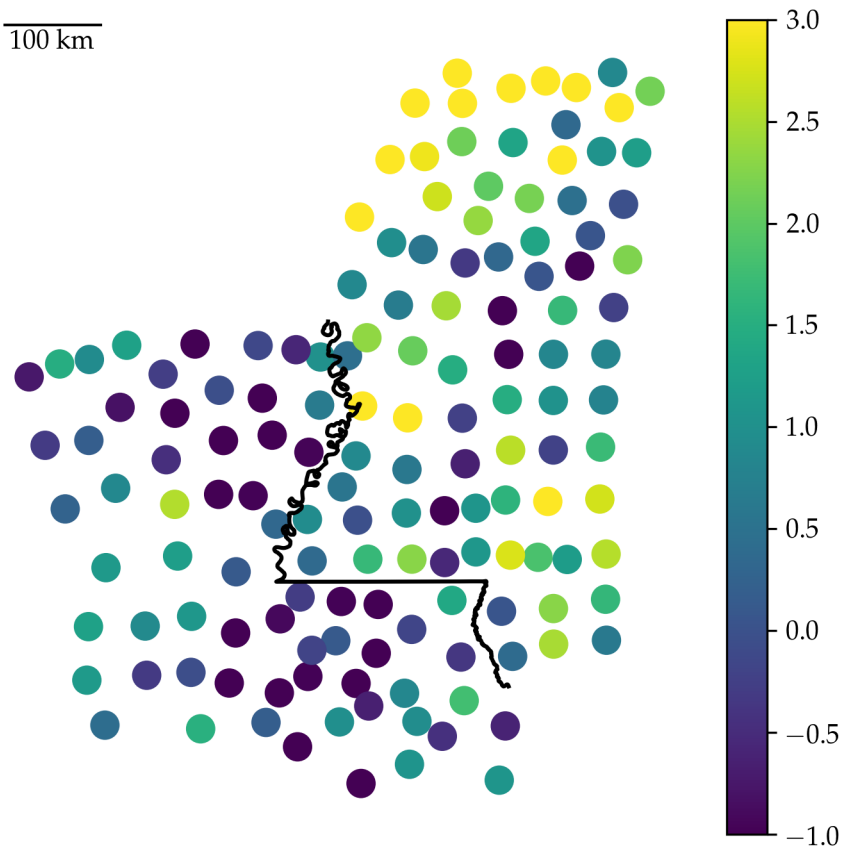

Figure 4: Set-up of the imaginary experiment in Louisiana and Mississippi. Each unit is at the centroid of a county. The colors indicated the observed outcomes in one draw of the simulation under $\tau = 1.5$. In this particular run, the p-values were 0.0016, 0.0018, and 0.0013 for the mLL, χ^2, and inverse-variance test respectively.

tions as the observations Y_T and Y_C.

Because calculating S^2 involves inverting a matrix Σ that is mathematically of full rank, but numerically of low rank, we may worry about the numerical stability of computing S. We verified in simulated examples that regularizing Σ by adding a small constant to its diagonal does not greatly affect the computed S^2. The parametric bootstrap ensures the frequentist validity of the test regardless of the regularization.

4.4 Power in simulated example

The three tests we developed leverage different aspects of the problem, and target two different null hypotheses. One may wonder how their power compares in the presence of a treatment effect. Considering once more the border between Louisiana and Mississippi, we imagine an experiment where the unit of analysis is the county, located at its centroid, as shown in Figure 4. We will simulate outcomes from a sin-

Table 3: Power of marginal likelihood, chi-squared, and inverse-variance tests, with nominal significance of $\alpha = 0.05$, under null and alternative hypothesis for simulated outcomes at the centroids of Louisiana and Mississippi counties.

	Power under	
Test	$\tau = 0$	$\tau = 1.5$
Marginal log-likelihood	0.050	0.935
χ^2	0.051	0.878
inverse-variance Σ^{-1}	0.067	0.971
bootstrap-calibrated Σ^{-1}	0.052	0.962
analytically-calibrated Σ^{-1}	0.051	0.962

gle Gaussian Process covering both states. For simplicity, we fix the hyperparameters to arbitrary values: $\sigma_\epsilon = \sigma_{GP} = 1.0$ and $\ell = 100\,\text{km}$. We then add a constant treatment effect τ to all the outcomes in Louisiana. The results of the three tests proposed so far are shown in the first three rows of Table 3 for $\tau = 0$ (null hypothesis) and $\tau = 1.5$ and significance level $\alpha = 0.05$.

We see that under the null, the χ^2 and likelihood ratio tests are valid (rejection of the null in 5% of simulations up to simulation error). This is enforced by the parametric bootstrap, which draws test statistics from the same null distribution to calibrate the tests. However, the p-values for the inverse-variance test are biased down, so that we will falsely reject the null 6.7% instead of 5% of the time. While unfortunate, this is unsurprising, since the inverse-variance test was derived heuristically rather than from a rigorous frequentist procedure. To make it valid, it can be calibrated using the same parametric bootstrap approach that was used for the likelihood and χ^2 tests. The calibration can also be achieved analytically, since $\mu_{\tau^{INV}|Y}$ is normally distributed under the null hypothesis. We derive the analytical calibration of the inverse-variance test in Appendix D.

After the calibration, the hypothesis test based on the inverse-variance mean still has higher power to detect the constant treatment effect than the mLL and χ^2 tests. This can lead to a paradox: we may reject the weak null hypothesis, but fail to reject the sharp null hypothesis (using the χ^2 or likelihood test), even though rejection of the weak null should logically imply rejection of the sharp null. This paradox isn't specific to this setting, and is discussed in depth in the context of randomization-based inference by (Ding, 2014). To maximize power, we therefore recommend using the calibrated inverse-variance test in studies where the main interest is an overall (average) increase or decrease in outcomes.

4.5 Placebo tests

Gaussian Process models are almost always misspecified. We do not believe that the Gaussian process with stationary squared exponential kernel is the true data-generating process, although we hope that the model

is sufficiently flexible to represent reality well. Under misspecification, we should be skeptical of results that rely on the truth of the model specification. We therefore encourage practicioners to probe the validity of the above hypothesis tests by running a "placebo" test. A placebo test repeatedly applies the hypothesis test on data that are known to have zero treatment effect (a "placebo"), in order to verify that the returned p-values are uniformly distributed. In our spatial setting, we will use the treatment and control regions separately as placebo groups. Within each placebo group, we repeatedly draw an arbitrary geographical border, creating new treatment and control groups. Because the border was chosen arbitrarily by us, we should not expect there to be a discontinuous jump in outcomes at this border. We then apply the bootstrapped likelihood test procedure described above to this arbitrarily divided data, store the results, and hope to obtain a roughly uniform distribution of p-values. In our implementation, we drew lines that split the placebo units in half at a sequence of angles $1°, 2°, 3°, \ldots, 180°$. The resulting p-values will obviously be highly correlated, so we should only expect a very roughly uniform distribution (because of the small effective sample size), but at the very least, this procedure allows us to visually verify that the p-values are not blatantly biased.

5 Example: NYC school districts

We illustrate the analysis of geographical regression discontinuity designs using house sales data from New York City. The city publishes information pertaining to property sales within the city in the last 12 months on a rolling basis. This includes the sale price, building class, and the address of the property. Public schools in the city are all part of the City School District of the City of New York, but the city-wide district is itself divided into 32 sub-districts. Within these districts, schools also have attendance zones, and children living within a zone are guaranteed attendance in their zone school unless the school is full [is this true? insideschools.com gives a more complete picture]. It is commonly held [could cite this article at cityrealty.com] that school districts therefore have an impact on real estate price, as parents are willing to pay more to live in districts with better schools. We therefore ask: can we measure a discontinuous jump in house prices across school district boundaries?

5.1 Preprocessing

In order to model the property sale prices, we need to obtain their locations. We geocode the address of each sale by merging the sales with NYC's Pluto database, which contains X and Y coordinates for each house, identified by its borough, zip code, block and lot. These coordinates are given in the `EPSG:2263` projection in units of feet. We use this projection throughout this example. For addresses that do not find a

match in Pluto, we use Google's geocoding API to obtain a latitude and longitude, which we then project to `EPSG:2263`.

We then filter the sales data, by removing sales (1) without a reported sale price, (2) outside of the residential building class categories (family dwellings, coops and condos); (3) missing the square footage or other covariates; (3) without a location due to failed geocoding; (4) smaller than 100 sq ft, and (5) outliers with log price per square foot less than 3 or more than 8.

5.2 Model for property prices

The outcome of interest is price per square foot. As is often done in the real estate literature, we take its logarithm to reduce the skew of the outcome. The complete model is then a Gaussian Process over the geographical covariates $\mathbf{s}$ super-imposed with a linear regression on the property covariates (building and tax class). Within a school district we could write the model as [suggestions for clearer notation welcome]:

$$
\begin{aligned}
Y_i &= \mu_{\texttt{District}[i]} + \beta_{1,\texttt{TaxClass}[i]} + \beta_{2,\texttt{BuildingClass}[i]} + f_{\texttt{District}[i]}(\mathbf{s}_i) + \epsilon_i \\
\epsilon_i &\sim \mathcal{N}\left(0, \sigma_y^2\right) \\
\mu_j &\sim \mathcal{N}\left(0, \sigma_\mu^2\right) \\
\beta_{1j}, \beta_{2j} &\sim \mathcal{N}\left(0, \sigma_\beta^2\right) \\
f_j &\sim \mathcal{GP}\left(0, k(\mathbf{s}, \mathbf{s}')\right) \\
k(\mathbf{s}, \mathbf{s}') &= \sigma_{GP}^2 \exp\left\{-\frac{(\mathbf{s}-\mathbf{s}')^\intercal(\mathbf{s}-\mathbf{s}')}{2\ell^2}\right\}
\end{aligned}
\tag{21}
$$

A visual inspection of the house sales map in Figure 5 suggests examining the border between districts 19 and 27. Importantly, the border between the two districts is also part of the border between Brooklyn and Queens, so we won't be able to attribute a difference in price solely to the causal effect of the school districts. We are first and foremost *measuring* a discontinuity in the house prices at the district. Attributing the discontinuity to a particular cause (school district or borough) is not directly supported by the data. This can be understood as an instance of *compound* treatments, as discussed by (Keele and Titiunik, 2015)

Figure 6 of Y in both districts also shows that marginally the house prices are very different. Our goal is to establish that this difference is measurable at the border, and not merely an underlying trend that spans both districts.

We fit the hyperparameters σ_β, σ_{GP}, ℓ and σ_ϵ by optimizing the marginal log-likelihood of the data within school districts 18, 19, 23, 24, 25, 26, 27, 28, and 29. We hold σ_μ fixed to 20 to give the district means

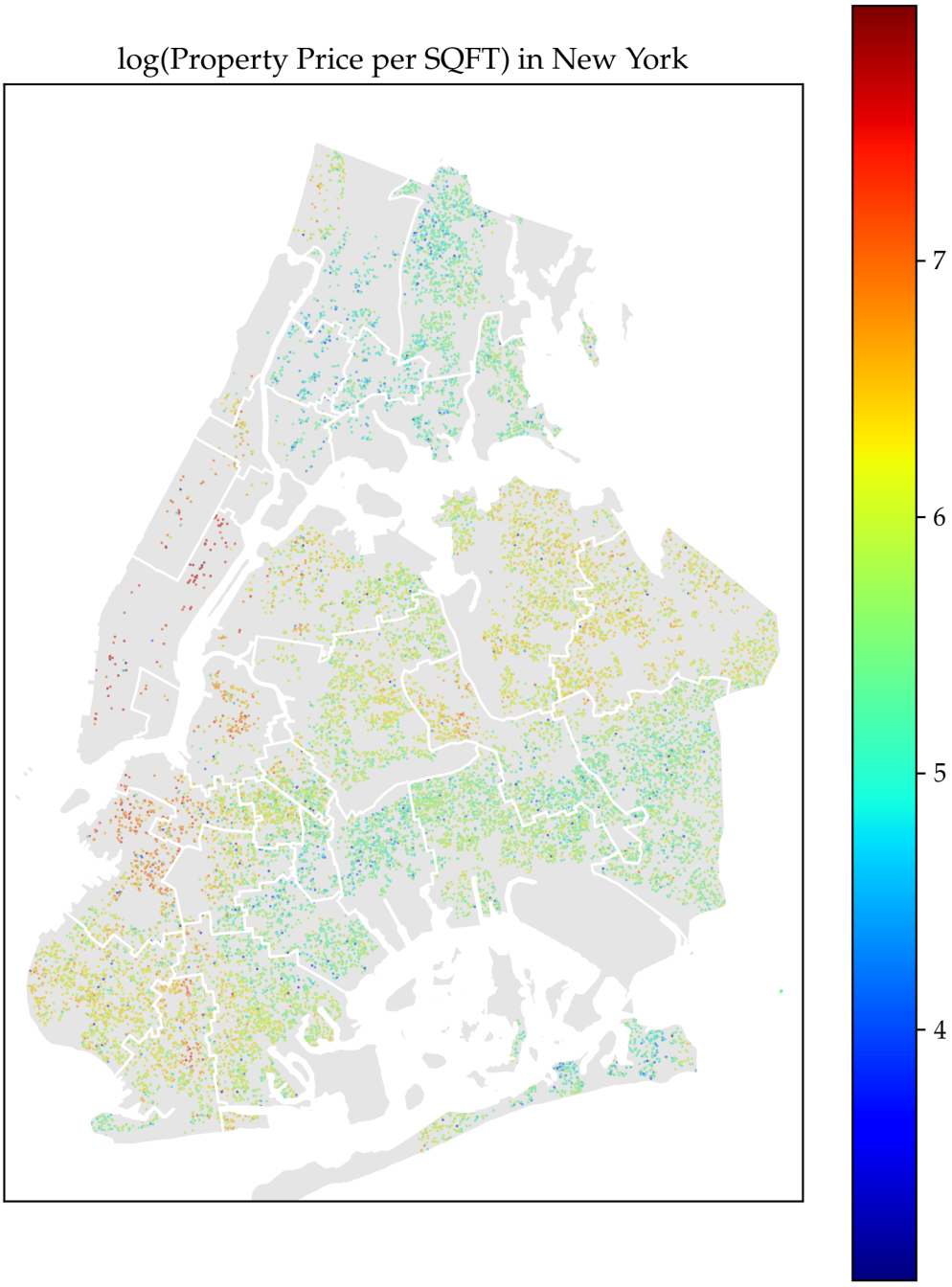

Figure 5: sales map

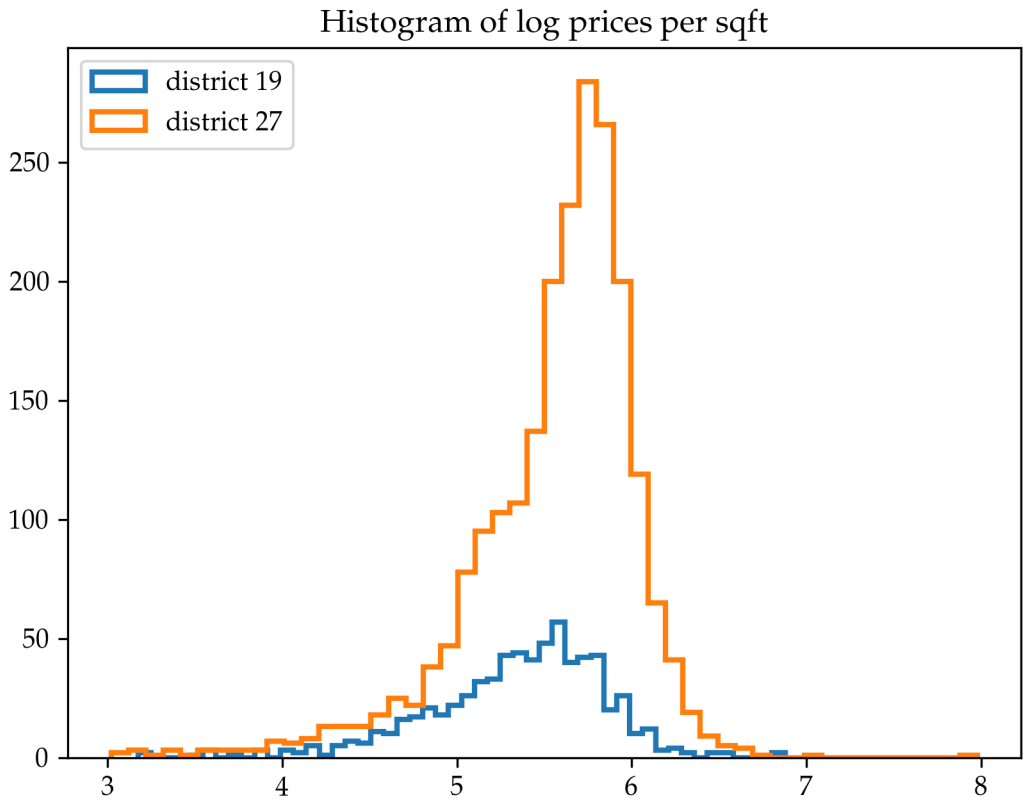

Figure 6: Histogram of log sale prices per square foot in NYC school districts 19 and 27.

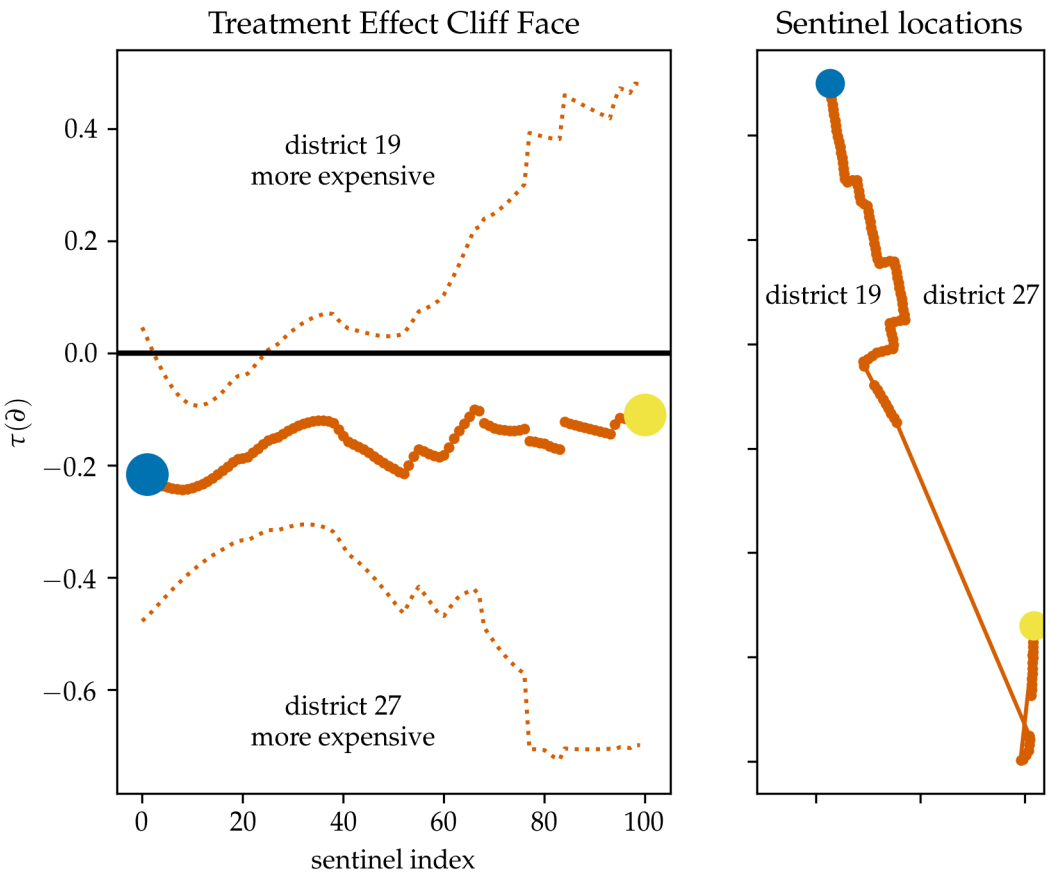

Figure 7: NYC cliff face

μ_j a fairly uninformative prior. T he fitted hyperparameters were $\sigma_\epsilon = 0.4020$, $\sigma_{GP} = 0.1955$, $\sigma_\beta = 0.1465$, and $\ell = 4482$ ft.

5.3 Cliff Face Estimator

We seek the treatment effect function $\tau(\mathcal{B})$ between the two districts. We could proceed by computing the joint predictive distributions $g_T(\boldsymbol{b}_{1:R}), g_C(\boldsymbol{b}_{1:R}) \mid Y_T, Y_C, \sigma_\beta, \sigma_{GP}, \ell, \sigma_\epsilon$, which is a 2R-dimensional multivariate normal distribution. Instead, we obtain the posterior means of the β_{1j} and β_{2j} coefficients, extract the residuals $Y_T - D_T\hat{\beta}$ and $Y_C - D_C\hat{\beta}$. This decorrelates $g_T(\boldsymbol{b}_{1:R})$ and $g_C(\boldsymbol{b}_{1:R})$ so they become independent multivariate normal distributions $g_T(\boldsymbol{b}_{1:R}) \mid Y_T, \hat{\beta}, \sigma_{GP}, \ell, \sigma_\epsilon$ and $g_C(\boldsymbol{b}_{1:R}) \mid Y_C, \hat{\beta}, \sigma_{GP}, \ell, \sigma_\epsilon$. In this example, we find that the posterior variance of β is low, and therefore the two approaches yield very similar results, but conditioning on the estimate of β is computationally convenient. We therefore proceed with this two-step approach.

Equipped with multivariate normal posteriors on $g_C(\boldsymbol{b}_{1:R})$ and $g_T(\boldsymbol{b}_{1:R})$, which are uncorrelated condi-

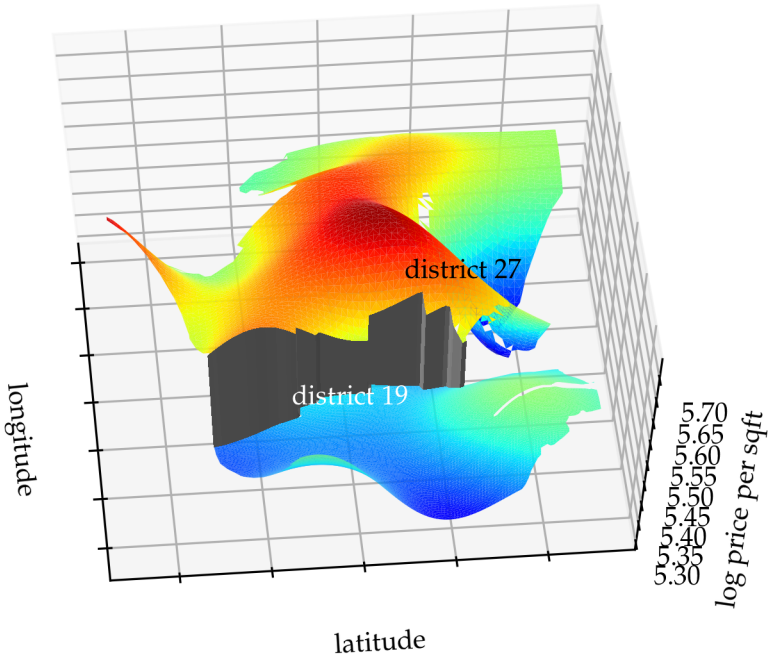

Figure 8: NYC surface plot viewed from the West. The grey cliff face connects the fitted price surfaces of districts 19 and 27, and has height given by (6) and shown in Figure 7.

tional on $\beta = \hat{\beta}$, we can now take their difference according to the procedure outline in Section 2.2, to obtain the posterior distribution of the cliff height $\tau(\mathcal{B})$ obtained at the sentinel locations. The cliff face is shown in Figure 7, and shows that the estimated $\tau(\mathcal{B})$ is negative everywhere along the border, which corresponds to higher property prices in district 27. However, the credible envelope is wide, especially in the southern section of the border, so we cannot visually rule out the null hypothesis that $\tau(\mathcal{B}) = 0$.

The "cliff face" can also be visualized directly in Figure 8 as the difference between the two log-price mean surfaces $g(\mathbf{s})$. The figure also gives a better sense of the spatial variation in prices captured by the model.

5.4 Average Log-Price Increase

The cliff face Figure 7 shows a negative treatment effect everywhere along the border, which can be averaged by the estimators we developed in Section 3. The two estimators we recommend, the inverse-variance

Estimand	Posterior Mean	Posterior Standard Dev.	Tail Prob.
τ^{UNIF}	-0.17	0.09	2.93%
τ^{ρ}	-0.19	0.06	0.04%
τ^{INV}	-0.19	0.06	0.03%
τ^{PROJ}	-0.18	0.08	1.31%
τ^{GEO}	-0.16	0.09	3.99%
τ^{POP}	-0.18	0.06	0.08%

Table 4: Average difference in log price per square foot between school districts 19 and 27. For each ATE estimand, we show the mean and standard deviation of its posterior distribution, and the tail probability $\mathbb{P}(\tau > 0 \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \hat{\beta}, \ell, \sigma_\mu)$ of the average treatment being greater than zero. Negative ATEs correspond to district 27 being more expensive.

weighted ATE and the finite-population yield ATE estimates of -0.19 and -0.18, which corresponds to about a 20% increase in property prices going from district 19 to district 27. All ATE estimators from Section 3 applied to this setting are shown in Table 5.4. In this example the different estimators yield similar answers, as the border is fairly straight and short relative to the fitted lengthscale.

5.5 Significant Difference in Price?

The inverse-variance weighted mean treatment effect hints at a significant treatment effect. But the posterior tail probability cannot be interpreted as a p-value. For this, we turn to the three tests developed in Section 4. In applied settings, running multiple tests should be done with care, but as we are proposing this new methodology, we apply all three tests in order to gain insight into their differences. Their results are found in Table 5.

Table 5: Results of hypothesis tests for New York school district house prices example.

Test	p-value
χ^2 bootstrap	0.012
mLL bootstrap	0.002
τ^{INV} uncalibrated	0.0007
τ^{INV} calibrated	0.002

The three tests reject the null hypothesis that $\tau(\mathcal{B}) = 0$ along the border between districts 19 and 27. This will not always be the case, as the calibrated inverse-variance test has higher power than the other two tests. The χ^2 test had the lowest power in the simulated example of Section 4.4, and here also returns the highest p-value.

To assess the validity of the three tests, we apply the placebo tests devised in Section 4.5. Within each district, we split the data in half by a line at angles $1°, 3°, 5°, 6°, \ldots, 179°$. Because these lines were drawn arbitrarily, we don't expect a discontinuous treatment effect between the two halves, and so we hope to

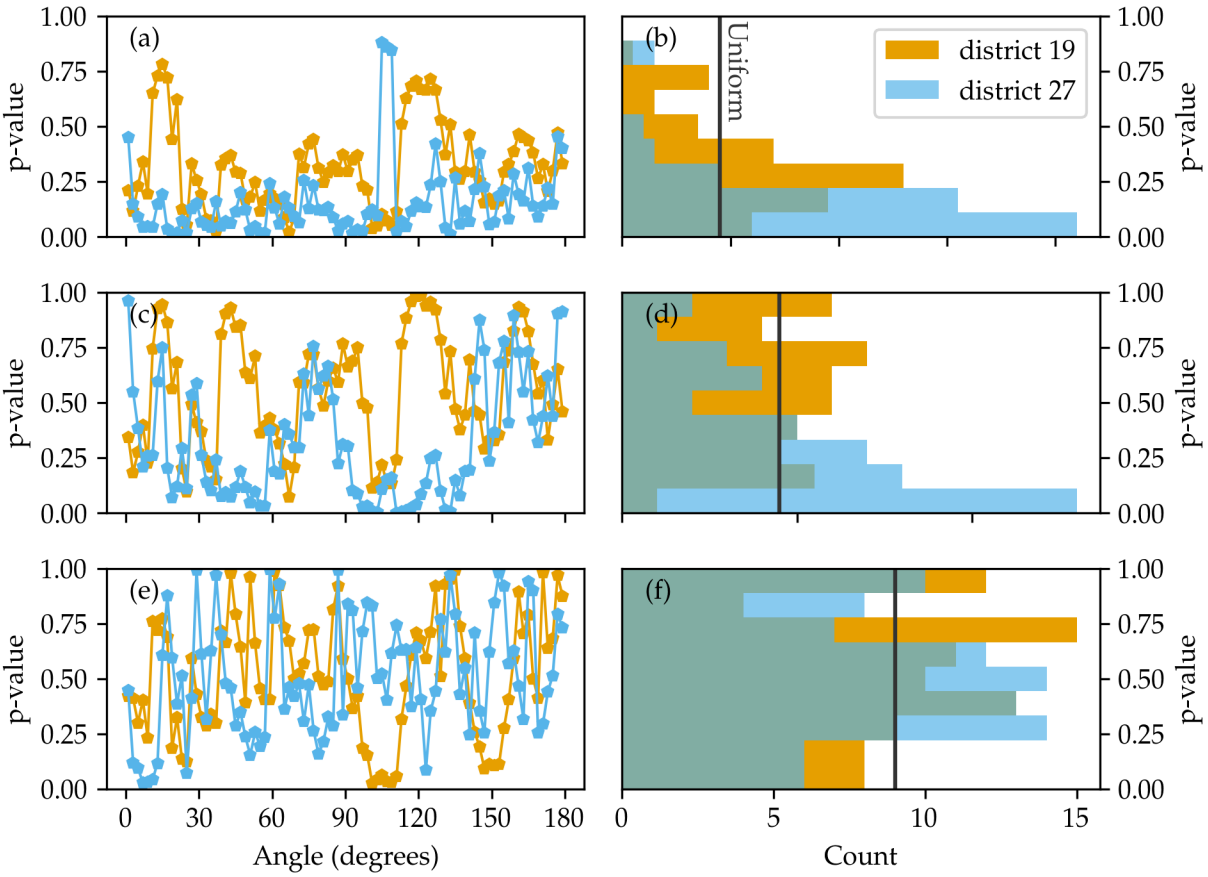

Figure 9: Placebo tests for significance tests applied to NYC school district house price example, applied within districts 19 and 27. The three rows respectively show results for the marginal log-likelihood bootstrap test, chi-squared bootrap test, and calibrated inverse-variance test. The first column shows the placebo p-value as a function of the border angle, in order to visualize the high auto-correlation of the placebo tests. The second column shows histograms of the placebo p-values, with the black vertical line indicating the uniform distribution for comparison.

nuity at the border. In particular, model misspecification, which is a concern in spatial models, makes the interpretation of the mLL test unreliable. Based on this vulnerability, and its manifestation in this example, we do not recommend relying on the likelihood-ratio test.

The χ^2 test shows more robustness, with Figure 9(d) showing some negative bias in district 27, and some positive bias in district 19, which could simply be due to the low effective sample size. We therefore believe that the χ^2 test will continue to be reliable under misspecification. It is only due to its low power that we hesitate to recommend its use in applications where the treatment effect is expected to be fairly homogenous.

Lastly, the calibrated inverse-variance placebo p-values display no obvious bias, with Figure 9(f) close to uniformly distributed, and Figure 9(e) showing a lower auto-correlation than the mLL and χ^2 tests. The high power and robustness of the inverse-variance test make a strong case for its use in most applications.

5.6 pairwise treatment effect (all districts)

6 Conclusion

Geographic regression discontinuity designs (GeoRDDs) arise when a treatment is assigned to one region, but not to another adjacent region. For outcomes that vary spatially, a direct comparison of mean outcomes between Y_T and Y_C, such as a t-test, is an invalid estimator of the treatment effect, as it is confounded by the spatial covariates. However, under smoothness assumptions, units adjacent to the border are comparable, and form a natural experiment. The same idea underpins causal interpretations of one-dimensional regression discontinuity designs (1D RDDs), where a single "forcing" variable controls the treatment assignment instead of a border separating two geographical regions. We use this similarity to motivate a general framework for the analysis of GeoRDDs. One-dimensional methods can be abstracted to three steps: (1) fit a smooth function on either side of the threshold; (2) extrapolate the functions to the threshold; and (3) take the difference of the two extrapolations to estimate the treatment effect at the threshold. For GeoRDDs we propose to (1) fit a smooth surface on either side of the border; (2) extrapolate the surfaces to the border; and (3) take the difference of the two extrapolations to estimate the treatment effect along the border.

Previous research has focused on extending methods developed for 1D RDDs to GeoRDDs. In applied settings, some have used the signed distance from the border as the forcing variable in a 1D RDD, but the resulting estimator is still spatially confounded. In this paper, our aim was to recognize the importance of the geographical aspect of the problem, and therefore draw from the spatial statistics literature, which brings

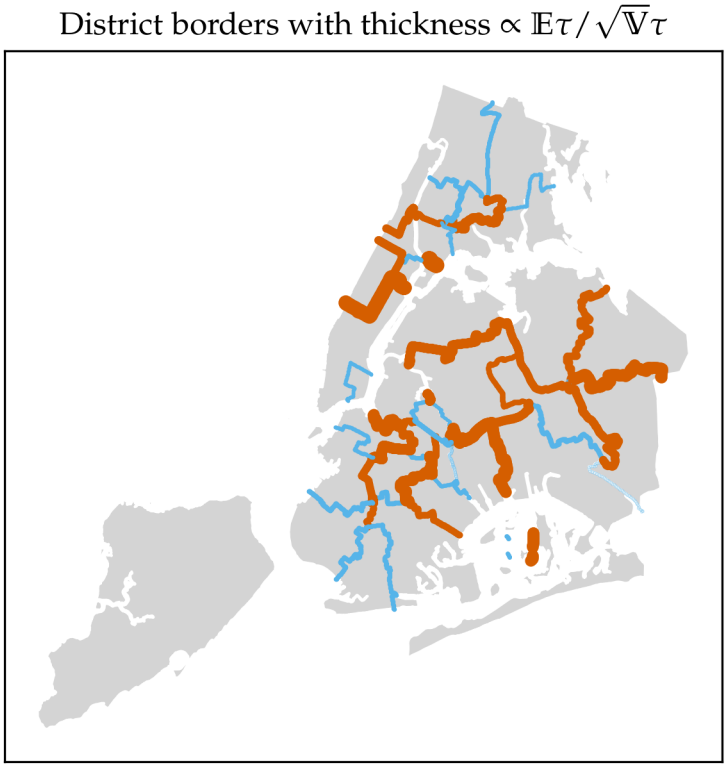

Figure 10: Pairwise effect size between adjacent districts. The thickness of each border is proprtional to the posterior effect size defined as $\mathbb{E}(\tau^{\text{INV}} \mid Y)/(\text{var}(\tau^{\text{INV}} \mid Y))^{1/2}$, indicating the strength of the evidence towards a discontinuity in house prices at that border. Effect sizes greater than 2 are shown in orange.

a rich set of tools designed specifically to model and exploit spatial correlations to obtain more powerful inference. We therefore used Gaussian process regression, known as kriging in the spatial statistics literature, to fit the smooth surfaces to the outcomes in step (1) of our general framework. Our approach yields a multivariate normal posterior distribution of the treatment effect for a collection of "sentinel" locations along the border.

Defining an "average treatment effect" estimand turns out to have surprising pitfalls. Simply integrating the treatment effect uniformly along the border yields an estimand that is inefficient and undesirably sensitive to the topology of the border. More sophisticated estimands, summarized in Table 2, are robust to this effect, and use the information available in the data more efficiently.

There are multiple valid approaches to hypothesis testing against the null hypothesis of zero treatment effect along the border. We recommend the use of the calibrated inverse-variance test, derived from the posterior distribution of the inverse-variance ATE estimator. It has generally high statistical power and behaved well in placebo tests in the NYC empirical example.

We applied our method to a publicly available dataset of New York City property sales, and detected a significant difference in house prices between school districts 19 and 27. We estimated a roughly 20% average increase in real estate prices when crossing the border from district 19 to district 27. However, the border between districts is also the border between boroughs, so we cannot interpret this difference as the causal effect of the school districts.

The main limitation of our approach to GeoRDDs is the reliance on modeling assumptions. We modeled the response surfaces as two independent Gaussian processes, with iid normal noise for each observation. Furthermore, we assumed an isotropic covariance function, with parameters equal in the treatment and control regions, and in particular we chose the squared exponential covariance. None of these assumptions are justified a priori, so their reasonableness should be evaluated with each application. The squared exponential covariance function makes smoothness assumptions that are often considered unrealistic in spatial settings, and so the Matérn covariance family is often recommended as a more robust alternative.

We use Gaussian processes as non-parametric smoothing devices used to capture spatial correlations, but do not think of them as truthful approximations to the data generating mechanism. Some care must therefore be taken not to lean heavily on this modeling assumption. In particular, we recommend that hypothesis tests always be accompanied by placebo tests. By applying the same procedure on data where no treatment was applied, we can verify that the test behaves correctly under the null hypothesis despite any potential model misspecification.

Because of the need to extrapolate the fitted processes to the border, our GeoRDD method is particularly vulnerable to the limitations of Gaussian processes when extrapolating. The distinction between interpola-

tion and extrapolation of spatial models is explored in some depth in (Stein, 2012). We expect that methodological advances that improve the extrapolating behavior of Gaussian processes would also improve the robustness of our method. For example, (Wilson and Adams, 2013) develop spectral mixture (SM) covariance kernels with good extrapolating behavior, which could be applied beneficially to GeoRDDs. However, SM kernels are aimed at time-series signals with some periodic or oscillatory behavior, which is more unusual in spatial applications, and may therefore not be as well-suited for use with GeoRDDs.

The use of Gaussian process regression to analyse GeoRDDs gives flexibility and extensibility to the method. This presents many opportunities for future research, inspired by the past and future development of methods in spatial statistics and machine learning that are based on Gaussian processes. In spatial statistics, kriging has been used as the foundation for a plethora of spatial models, which may be adapted for the purposes of analyzing GeoRDDs. (Banerjee et al., 2014) provides a good introduction to the richness of the spatial statistics field. For example, if the outcomes are binary, proportions, or counts, then binomial or Poisson likelihoods could be substituted for the iid normal likelihood used in this paper. Besides ATE and hypothesis testing, another potential question of interest is whether heterogeneous treatment effects are detectable along the border. In other words, can we reject the null hypothesis that the first derivative of the treatment

The framework and techniques of this paper could also be extended to spatio-temporal settings. If the treatment is only applied to the treatment region after a time $t*$, one could envision a three-dimensional RDD consisting of the geographical border in the spatial dimensions, and a straight line through $t*$ in the temporal dimension. The only necessary modification to our approach is that the Gaussian process model would need to be augmented with a temporal component, for example with an anisotropic squared exponential covariance function. Longitudinal studies could also be handled by such an approach, with the addition of random intercepts for each unit. We leave spatio-temporal RDDs using Gaussian process models to future research.

A Spatial confounding of 1D RDD applied to GeoRDD

Analysing GeoRDDs by using the signed distance from the border as a forcing variable in a 1D RDD can lead to spatial confounding. We demonstrate this with a simple artificial example, depicted in Figure 11.

Suppose we have units in a 2D square, with spatial coordinates $\mathbf{s}_1 \in [-1, 1]$, and $\mathbf{s}_2 \in [-1, 1]$, and with a horizontal border at $s_2 = 0$ separating a treatment region from a control region. Let us assume the null hypothesis, with outcomes driven only by $\mathbf{s}_1$ (parallel to the border), given by $Y_i = \alpha \mathbf{s}_{1i} + \epsilon_i$, where ϵ_i is an iid noise term $\epsilon_i \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$. Lastly, let us consider the situation where the density $\rho(\mathbf{s})$ of units is

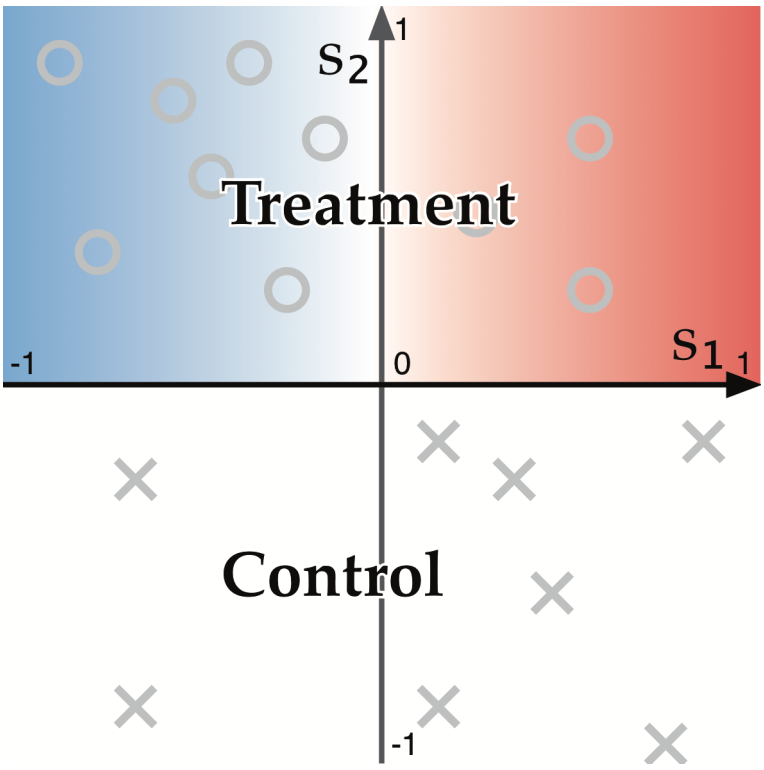

Figure 11: Confounding.

different in each quadrant of the square:

$$\begin{aligned}
\rho(\mathbf{s}) &= 2\rho_0, \text{ where } \mathbf{s}_1 < 0,\ \mathbf{s}_2 > 0 \quad \text{(top left)} \\
\rho(\mathbf{s}) &= \rho_0, \text{ where } \mathbf{s}_1 > 0,\ \mathbf{s}_2 > 0 \quad \text{(top right)} \\
\rho(\mathbf{s}) &= 2\rho_0, \text{ where } \mathbf{s}_1 > 0,\ \mathbf{s}_2 < 0 \quad \text{(bottom right)} \\
\rho(\mathbf{s}) &= \rho_0, \text{ where } \mathbf{s}_1 < 0,\ \mathbf{s}_2 < 0 \quad \text{(bottom left)}
\end{aligned} \tag{22}$$

The projection RDD then considers a 1D RDD along s_2. The usual RDD estimand (1) can be obtained analytically, and equals $\tau = \frac{-\alpha}{3}$, despite assuming the null hypothesis. This is because s_1 acts as a hidden confounder, whose distribution changes discontinuously at the border, which leads to bias and inconsistency in the projected RDD estimate. In geographical settings, a discontinuous change in the density of units at the border is not unusual: for example a border could run alongside a park, or a small body of water, therefore with zero population density on one side of the border. A visual inspection of Figure 5 showing the locations of units in a New York City property sales dataset reveals many examples of this.

B Covariances for Gaussian process model

$$
\begin{aligned}
m_T, m_C &\sim \mathcal{N}\left(0, \sigma_\mu^2\right) \\
\operatorname{cov}(Y_{iT}, m_T) &= \sigma_\mu^2 \\
\operatorname{cov}(Y_{iC}, m_C) &= \sigma_\mu^2 \\
\operatorname{cov}(Y_{iT}, m_C) &= \operatorname{cov}(Y_{iC}, m_T) = 0 \\
\operatorname{cov}\left(Y_{iT}, f_T(\mathbf{s}')\right) &= k(\mathbf{s}_i, \mathbf{s}') \\
\operatorname{cov}\left(Y_{iC}, f_C(\mathbf{s}')\right) &= k(\mathbf{s}_i, \mathbf{s}') \\
\operatorname{cov}\left(Y_{iT}, f_C(\mathbf{s}')\right) &= \operatorname{cov}\left(Y_{iC}, f_T(\mathbf{s}')\right) = 0 \\
\operatorname{cov}(Y_{iT}, Y_{jT}) &= \operatorname{cov}(Y_{iC}, Y_{jC}) = \sigma_\mu^2 + k(\mathbf{s}_i, \mathbf{s}_j) + \delta_{ij}\sigma_\epsilon^2 \\
\operatorname{cov}(Y_{iT}, Y_{jC}) &= 0
\end{aligned}
\tag{23}
$$

C Posterior mean of $\hat{\beta}$

[Derivation of $\hat{\beta}$ below: should it be in the covariates section? should it be in an appendix? is it too elementary to be in this paper?]

$$
\begin{aligned}
\Sigma_{Y|\beta} &\equiv \operatorname{cov}\left(Y \mid \beta\right) && \text{conditional variance of } Y \\
\operatorname{cov}\left(Y_i, Y_j \mid \beta\right) &= \sigma_\epsilon^2 \delta_{ij} + k\left(\mathbf{s}_i, \mathbf{s}_j\right) \delta_{\texttt{District[i],District[j]}} && \text{(block diagonal)} \\
\Sigma_\beta &\equiv \operatorname{cov}\left(\beta\right) = \sigma_\beta^2 I_p && \text{prior variance of } \beta \\
\Sigma_Y &\equiv \operatorname{cov}\left(Y\right) = \Sigma_{Y|\beta} + D^\intercal \Sigma_\beta D && \text{unconditional variance of } Y \\
T_\beta &= D^\intercal \Sigma_{Y|\beta}^{-1} D + \Sigma_\beta^{-1} && \text{precision matrix of } \beta \\
\hat{\beta} &= \left(T_\beta^{-1} D\right)\left(\Sigma_{Y|\beta}^{-1}\left(Y - \mu\right)\right) && \text{posterior mean of } \beta
\end{aligned}
\tag{24}
$$

D Calibration of inverse-variance test

First, let's remind ourselves how the inverse-variance posterior mean estimate was obtained. We will then derive its distribution under the null hypothesis.

$$
\begin{aligned}
\tau^{\text{INV}} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell &\sim \mathcal{N}\left(\mu_{\tau^{\text{INV}}|Y}, \Sigma_{\tau^{\text{INV}}|Y}\right) \\
\mu_{\tau^{\text{INV}}|Y} &\approx \left(\mathbf{1}_R^\intercal \, \Sigma_{\boldsymbol{b}_{1:R}|Y}^{-1} \mu_{\boldsymbol{b}_{1:R}|Y}\right) / \left(\mathbf{1}_R^\intercal \, \Sigma_{\boldsymbol{b}_{1:R}|Y}^{-1} \mathbf{1}_R\right) \\
\mu_{\boldsymbol{b}_{1:R}|T} &\equiv \operatorname{cov}\left(g_T(\boldsymbol{b}), Y_T\right) \operatorname{cov}\left(Y_T\right)^{-1} Y_T \\
\mu_{\boldsymbol{b}_{1:R}|C} &\equiv \operatorname{cov}\left(g_T(\boldsymbol{b}), Y_C\right) \operatorname{cov}\left(Y_C\right)^{-1} Y_C \\
\mu_{\mathcal{B}|Y} &= \mu_{\boldsymbol{b}_{1:R}|T} - \mu_{\boldsymbol{b}_{1:R}|C} \\
\mu_{\tau^{\text{INV}}|Y} &= \left(\mathbf{1}_R^\intercal \, \Sigma_{\boldsymbol{b}_{1:R}|Y}^{-1} \mu_{\boldsymbol{b}_{1:R}|Y}\right) / \left(\mathbf{1}_R^\intercal \, \Sigma_{\boldsymbol{b}_{1:R}|Y}^{-1} \mathbf{1}_R\right)
\end{aligned}
\tag{25}
$$

Under our parametric null hypothesis H_0, Y_T and Y_C are drawn from a single smooth Gaussian process, with no discontinuity at the border. Their joint covariance is

$$
\begin{aligned}
\operatorname{cov}\left(\begin{pmatrix} Y_T \\ Y_C \end{pmatrix} \mid H_0\right) &= \begin{bmatrix} \Sigma_{TT} & \Sigma_{TC} \\ \Sigma_{TC}^\intercal & \Sigma_{CC} \end{bmatrix} \text{ where} \\
\Sigma_{TT} &\equiv K_{TT} + \sigma_\epsilon^2 I_{n_T} \\
\Sigma_{CC} &\equiv K_{CC} + \sigma_\epsilon^2 I_{n_C} \\
\Sigma_{TC} &\equiv K_{TC}
\end{aligned}
\tag{26}
$$

where the entries of K_{TT}, K_{CC} and K_{TC} are obtained by evaluating the Gaussian process kernel for each pair of points within and between the treatment and control regions. The predicted mean outcomes at the sentinels $\mu_{\boldsymbol{b}_{1:R}|T}$ and $\mu_{\boldsymbol{b}_{1:R}|T}$ are obtained by left-multiplying Y_T and Y_C by matrices that are deterministic functions of the unit locations and the hyperparameters

$$
\begin{aligned}
A_T &\equiv \operatorname{cov}\left(g_T(\boldsymbol{b}_{1:R}), Y_T\right) \operatorname{cov}\left(Y_T\right)^{-1} = K_{\boldsymbol{b}_{1:R}T} \Sigma_{TT}^{-1}, \text{ and} \\
A_C &\equiv \operatorname{cov}\left(g_C(\boldsymbol{b}_{1:R}), Y_C\right) \operatorname{cov}\left(Y_C\right)^{-1} = K_{\boldsymbol{b}_{1:R}C} \Sigma_{CC}^{-1}.
\end{aligned}
\tag{27}
$$

where we dropped the explicit conditioning on the null hypothesis for readability.

The joint distribution of $\mu_{\boldsymbol{b}_{1:R}|T}$ and $\mu_{\boldsymbol{b}_{1:R}|T}$ is consequently also multivariate normal with mean zero and covariance

$$
\operatorname{cov}\left(\begin{pmatrix} A_T Y_T \\ A_C Y_C \end{pmatrix} \mid H_0\right) = \begin{bmatrix} A_T \Sigma_{TT} A_T^\intercal & A_T \Sigma_{TC} A_C^\intercal \\ (A_T \Sigma_{TC} A_C^\intercal)^\intercal & A_C \Sigma_{CC} A_C^\intercal \end{bmatrix}
\tag{28}
$$

Continuing in this fashion, $\mu_{\mathcal{B}|Y}$ is yet another zero-mean multivariate normal with covariance

$$\begin{aligned}\operatorname{cov}\left(\mu_{\mathcal{B}|Y} \mid H_0\right) &= \operatorname{cov} A_T Y_T - A_C Y_C \\ &= A_T \Sigma_{TT} A_T{}^\intercal + A_C \Sigma_{CC} A_C{}^\intercal - A_T \Sigma_{TC} A_C{}^\intercal - (A_T \Sigma_{TC} A_C{}^\intercal)^\intercal\end{aligned} \tag{29}$$

Weighted mean estimators are linear transformation of $\mu_{\mathcal{B}|Y}$, and so under H_0, they are normally distributed with mean zero. For a weight vector $\boldsymbol{\nu}$, its variance is given by

$$\begin{aligned}\operatorname{var}\left(\bar{\tau}^{\boldsymbol{\nu}} \mid H_0\right) &= \operatorname{cov}\left(\frac{\boldsymbol{\nu}^\intercal \mu_{\mathcal{B}|Y}}{\mathbf{1}_R^\intercal \boldsymbol{\nu}}\right) \\ &= \frac{\boldsymbol{\nu}^\intercal \operatorname{cov}\left(\mu_{\mathcal{B}|Y}\right) \boldsymbol{\nu}}{\left(\mathbf{1}_R^\intercal \boldsymbol{\nu}\right)^2}.\end{aligned} \tag{30}$$

From this null distribution the p-value follows:

$$\mathbb{P}\left(\left|\bar{\tau}^{\boldsymbol{\nu}}\right| > \left|\bar{\tau}^{\boldsymbol{\nu}}_{obs}\right| \mid H_0\right) = 2\Phi\left(-\frac{\left|\bar{\tau}^{\boldsymbol{\nu}}_{obs}\right|}{\sqrt{\operatorname{var}\left(\bar{\tau}^{\boldsymbol{\nu}} \mid H_0\right)}}\right). \tag{31}$$

Our calibrated inverse-variance test is the special case of this final step where the weights are chosen to be $\boldsymbol{\nu} = \Sigma_{\mathcal{B}|Y}^{-1} \mathbf{1}_R$.

E Wiggly border simulation results

n_{wiggles}	$\widehat{\tau^{\text{UNIF}}}$	τ^{UNIF}	$\widehat{\tau^{\text{INV}}}$	τ^{INV}	$\widehat{\tau^{\rho}}$	τ^{ρ}	$\widehat{\tau^{\text{PROJ}}}$	τ^{PROJ}	$\widehat{\tau^{\text{GEO}}}$	τ^{GEO}	$\widehat{\tau^{\text{POP}}}$	τ^{POP}
0	1.02 (0.14)	1.00	1.24 (0.10)	1.23	1.21 (0.10)	1.21	1.23 (0.10)	1.24	1.02 (0.14)	1.00	1.21 (0.10)	1.21
1	1.01 (0.13)	0.99	1.15 (0.09)	1.16	1.19 (0.10)	1.19	1.24 (0.10)	1.24	1.00 (0.13)	0.97	1.18 (0.10)	1.17
2	0.98 (0.13)	0.95	1.14 (0.09)	1.16	1.14 (0.10)	1.14	1.23 (0.10)	1.24	0.97 (0.13)	0.94	1.14 (0.10)	1.14
3	0.94 (0.13)	0.91	1.14 (0.09)	1.16	1.09 (0.10)	1.08	1.23 (0.10)	1.23	0.96 (0.13)	0.93	1.13 (0.10)	1.12
5	0.86 (0.13)	0.82	1.14 (0.09)	1.15	0.99 (0.11)	0.96	1.23 (0.10)	1.23	0.95 (0.13)	0.92	1.12 (0.10)	1.11
10	0.72 (0.14)	0.67	1.14 (0.09)	1.15	0.80 (0.13)	0.76	1.23 (0.10)	1.23	0.95 (0.13)	0.92	1.12 (0.10)	1.11
20	0.58 (0.15)	0.52	1.14 (0.09)	1.15	0.63 (0.15)	0.58	1.23 (0.10)	1.23	0.95 (0.13)	0.92	1.12 (0.10)	1.11
40	0.48 (0.16)	0.41	1.14 (0.09)	1.15	0.50 (0.16)	0.44	1.23 (0.10)	1.23	0.95 (0.13)	0.92	1.12 (0.10)	1.11
80	0.41 (0.17)	0.34	1.14 (0.09)	1.15	0.42 (0.17)	0.35	1.23 (0.10)	1.23	0.95 (0.13)	0.92	1.12 (0.10)	1.11
160	0.38 (0.18)	0.30	1.14 (0.09)	1.15	0.39 (0.18)	0.30	1.24 (0.10)	1.23	0.95 (0.13)	0.92	1.12 (0.10)	1.11
320	0.35 (0.18)	0.27	1.14 (0.09)	1.15	0.36 (0.18)	0.28	1.23 (0.10)	1.23	0.95 (0.13)	0.92	1.12 (0.10)	1.11
640	0.35 (0.18)	0.26	1.14 (0.09)	1.15	0.35 (0.18)	0.26	1.24 (0.10)	1.23	0.95 (0.13)	0.92	1.12 (0.10)	1.11
1000	0.34 (0.18)	0.26	1.14 (0.09)	1.15	0.34 (0.18)	0.26	1.24 (0.10)	1.23	0.95 (0.13)	0.92	1.12 (0.10)	1.11

Table 6: Table of posterior mean, posterior standard deviation and true value for each average treatment effect estimand as the wiggliness of the border is increased.

References

Banerjee, S., B. P. Carlin, and A. E. Gelfand, 2014: *Hierarchical modeling and analysis for spatial data*. Crc Press.

Branson, Z., M. Rischard, L. Bornn, and L. Miratrix, 2017: A nonparametric bayesian methodology for regression discontinuity designs. URL https://arxiv.org/abs/1704.04858, 1704.04858.

Chen, Y., A. Ebenstein, M. Greenstone, and H. Li, 2013: Evidence on the impact of sustained exposure to air pollution on life expectancy from china's huai river policy. *Proceedings of the National Academy of Sciences*, **110 (32)**, 12 936–12 941.

Cook, T. D., 2008: "waiting for life to arrive": a history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, **142 (2)**, 636–654.

Ding, P., 2014: A paradox from randomization-based causal inference. URL https://arxiv.org/abs/1402.0142, 1402.0142.

Imbens, G., and K. Kalyanaraman, 2012: Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies*, **79 (3)**, 933–959.

Imbens, G., and T. Zajonc, 2011: Regression discontinuity design with multiple forcing variables. *Report, Harvard University.[972]*.

Imbens, G. W., and T. Lemieux, 2008: Regression discontinuity designs: A guide to practice. *Journal of econometrics*, **142 (2)**, 615–635.

Keele, L., S. Lorch, M. Passarella, D. Small, and R. Titiunik, 2017: *An Overview of Geographically Discontinuous Treatment Assignments with an Application to Children's Health Insurance*, chap. 4, 147–194. Emerald Publishing Limited, doi:10.1108/S0731-905320170000038007, URL http://www.emeraldinsight.com/doi/abs/10.1108/S0731-905320170000038007, http://www.emeraldinsight.com/doi/pdf/10.1108/S0731-905320170000038007.

Keele, L., R. Titiunik, and J. R. Zubizarreta, 2015: Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **178 (1)**, 223–239.

Keele, L. J., and R. Titiunik, 2015: Geographic boundaries as regression discontinuities. *Political Analysis*, **23 (1)**, 127–155, doi:10.1093/pan/mpu014.

Li, F., K. L. Morgan, and A. M. Zaslavsky, 2016: Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, **(just-accepted)**.

MacDonald, J. M., J. Klick, and B. Grunwald, 2015: The effect of private police on crime: evidence from a geographic regression discontinuity design. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Papay, J. P., J. B. Willett, and R. J. Murnane, 2011: Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, **161 (2)**, 203–207.

Rasmussen, C. E., and C. K. Williams, 2006: *Gaussian processes for machine learning*, Vol. 1. MIT press Cambridge.

Rencher, A. C., 2003: *Methods of multivariate analysis*, Vol. 492. John Wiley & Sons.

Stein, M. L., 2012: *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Thistlethwaite, D. L., and D. T. Campbell, 1960: Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, **51 (6)**, 309.

Wilson, A., and R. Adams, 2013: Gaussian process kernels for pattern discovery and extrapolation. *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 1067–1075.