GeoRDD manuscript

Maxime Rischard

December 6, 2017

Contents

Abstract

Regression discontinuity designs (RDDs) arise in observational studies when the treatment assignment is fully determined by a single "forcing" variable: all units on one side of a fixed threshold 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 adjacent 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 understood as a spatial analog to the usual approach to RDDs, and 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 then turn our attention to the definition and estimation of the average treatment effect along the border, and to hypothesis testing for GeoRDDs. We illustrate our methodology using publicly available house sale data from New York City, and show evidence of a discontinuity in price between neighboring school districts.

1 Introduction

- Problem we're trying to solve

  - treatment applied to one region and not a neighboring region
  - with no overlap
  - how to estimate the causal effect of the treatment?
  - if the outcome is not spatially varying, there's no problem
  - otherwise, the treatment is confounded with location

- In 1D, this is recognied as a regression discontinuity design

  - which is now a well-established methodology [citations]
  - and comes with a causal inference story [Imbens]

- In spatial settings, practicioners therefore attempt to use these tools

  - but they don't generalize easily to 2D
  - so often end up projecting onto distance from border

- We think this is a bad idea

  - ignores spatial structure / correlation
  - low power and could get the wrong answer

- Mention some of the more sophisticated approaches to this same problem [Keele]

  - maybe briefly mention why we don't like them

- Our approach: framework analogous to 1D RDD

  - 1D:
    1. fit the outcome **function** on both sides
    2. extrapolate to the **discontinuity point** $x*$

3. take difference to obtain $\tau(x*) \in \mathbb{R}$
   - 2D:
        1. fit the outcome **surface** on both sides
        2. extrapolate to **border curve** $\mathcal{B}$
        3. take difference to obtain $\tau(\mathcal{B}) : \mathbb{R} \to \mathbb{R}$ [help with notation? or is this too mathematical anyway?]

- Challenge 1: functional estimand is unusual
- Challenge 2: how to fit surface on both sides

   - in this framework there is not restriction on how surface fitting and extrapolation are performed
   - in 1D RDD, local linear regression has become standard
        * though other options have been explored (like splines?)
   - but this isn't suitable in 2D

- Challenge 3: summarizing functional estimand

   - how to take an average?
   - 1D manifold embedded in 2D space poses problems
   - pitfalls detailed in Section X

- Challenge 4: hypothesis testing on functional estimand
- We use Gaussian processes (kriging), which are widespread in spatial statistics

   - many advantages
        * flexible
        * known to perform well in spatial settings
        * analytic solutions

- we explore and address the 4 challenges using GPs
- by the way, GP's can also be used in 1D [cite Zach]

Consider an experiment performed on a population of units that are distributed geographically, in which a treatment gets applied to selected units, while the rest are used as a control group. For each unit, an outcome Y_i is measured. If the treatment assignment is randomized, standard estimators from randomized control trials apply, such as the difference in observed means within the treatment and control groups. If outcomes vary spatially, it would be wise to match units that are near each other, in order to prevent this variation from confounding the estimate [a citation here would be nice].

Now, we may find ourselves in a situation where the treatment does not get assigned at random. Instead, the land is arbitrarily divided into two contiguous regions, and the treatment is applied to units in one region, and withheld from those in the other. Ignoring the spatial covariates is particularly inadvisable in this scenario, as units in the two regions could be very different from each other. Matching is no longer directly applicable either, as there is no overlap in the spatial covariates. (Keele et al., 2015) develop a methodology to extend matching to GeoRDDs, by matching units across the border while minimizing the total sum of geographic distances between them. While sound, power is low because some units remain unmatched, and it requires a compromise between power and unbiasedness: matching units that are further away from the border reduces variance but introduces bias due to spatial variation.

Examples of GeoRDDs situation occasionally appear in the literature. 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.

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 regression discontinuity designs (RDD) enable the estimation of a causal effect despite the lack of overlap in the covariate. Our aim is to develop methodologies for our setting, which we recognize as geographic regression discontinuity designs (Keele and Titiunik, 2015).

Noticing this connection, practicioners often wish to use the well-established tools developed for one-dimensional RDDs for their spatial problem. Unfortunately, the most common methodologies, based on local linear regression on both sides of the border, do not extend naturally to two-dimensional settings. A temptation therefore arises to reduce a geographic RDD problem to a classical (one-dimensional) RDD by projecting locations onto the distance away from the border, an method that we will refer to as "projected RDD." This is the approach taken by, for example, (MacDonald et al., 2015) and (Chen et al., 2013). We can illustrate the inappropriateness of this approach with a simple example. Suppose we have units in a 2D square, with spatial coordinates $s_1 \in [-1, 1]$, and $s_2 \in [-1, 1]$, and with a border at $s_1 = 0$ separating a treatment region from a control region. Let us assume the null hypothesis, with outcomes driven only by s_2 (parallel to the border), given by $Y_i = \alpha s_{2i} + \epsilon_i$, 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 different in each quadrant of the square:

$$
\begin{aligned}
\rho(\mathbf{s}) &= 2\rho_0, \text{ where } \mathbf{s}_1 < 0, \ \mathbf{s}_2 > 0 && \text{(top left)} \\
\rho(\mathbf{s}) &= \rho_0, \text{ where } \mathbf{s}_1 > 0, \ \mathbf{s}_2 > 0 && \text{(top right)} \\
\rho(\mathbf{s}) &= 2\rho_0, \text{ where } \mathbf{s}_1 > 0, \ \mathbf{s}_2 < 0 && \text{(bottom right)} \\
\rho(\mathbf{s}) &= \rho_0, \text{ where } \mathbf{s}_1 < 0, \ \mathbf{s}_2 < 0 && \text{(bottom left)}
\end{aligned}
\tag{1}
$$

The projection RDD then considers a univariate RDD along s_1. The usual RDD estimand (5) can be obtained analytically, and equals $\tau = \frac{-\alpha}{3}$, despite assuming the null hypothesis. This is because s_2 acts as a hidden confounder, whose distribution changes discontinuously at the border, which leads to bias and inconsistency in the projected univariate 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.

We propose a framework for analysing GeoRDDs that is directly analogous to their univariate counterpart, and avoids the need to project positions onto a single dimension. Univariate RDD methodologies can be abstracted to three steps:

1. Fit a smooth **function** to the outcomes against the covariate on each side of the discontinuity;
2. Extrapolate the functions to the **discontinuity point** x^*; and
3. Subtract the two extrapolations to estimate the treatment effect at the threshold point $\tau(x^*) \in \mathbb{R}$.

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

1. Fit a smooth **surface** to the outcomes against the geographical covariates on each side of the discontinuity;
2. Extrapolate the surfaces to the **border curve** $\mathcal{B}$; and
3. Subtract the two extrapolations to estimate the treatment effect along the border $\tau(\mathbf{s}) : \mathcal{B} \to \mathbb{R}$.

While local linear regression is commonly used to fit and extrapolate the smooth response functions in RDDs, we propose to use kriging (Gaussian process regression) to fit the response surfaces in GeoRDDs. This is motivated by the well-established use of kriging for fitting smoothly varying spatial processes in the spatial statistics literature. 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 Gaussian process regression. The application of Gaussian process regression to univariate RDDs is presented in (Branson et al., 2017), with encouraging results compared to local linear regression.

Instead of a point estimate in the RDD case, the GeoRDD framework gives a functional estimate $\tau(\mathbf{s})$, defined on an irregular one-dimensional manifold (the border) that is embedded in two dimensional space (the surface of the Earth), and that is more challenging to interpret. Subsequently, analysts might often be interested in summarizing the information contained in the functional estimand. In section 4, we explore possible estimands for the average treatment effect (ATE), and elucidate their respective pitfalls and advantages. In section 5, we turn to hypothesis testing, and propose methods to test against the null hypothesis of no treatment effect. We also suggest Placebo tests [cite?] to examine the validity of the hypothesis tests.

4

2 Model Specification

2.1 Notation

We largely adopt the setup and notation for geographic regression discontinuity designs laid out in (Keele and Titiunik, 2015). n units are observed within an area $\mathcal{A}$ of 2-dimensional coordinate space. The n 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 $\mathcal{A}^T \cap \mathcal{A}^C = \emptyset$ and $\mathcal{A}^T \cup \mathcal{A}^C = \mathcal{A}$. 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 location $\mathbf{s}_i$, $Z_i = \mathbb{I}\left\{\mathbf{s}_i \in \mathcal{A}^T\right\}$. The observed outcome can be written as $Y_i = Z_i Y_{iT} + (1 - Z_i) Y_{iC}$.

The border between $\mathcal{A}^T$ and $\mathcal{A}^C$ is denoted $\mathcal{B}$. But for computational reasons, we will often represent the border as a set $\boldsymbol{b}$ of R "sentinel" points along the border, with each $\boldsymbol{b}_i \in \mathcal{B}$.

2.2 1GP solution [to be deleted]

Most straightforwardly, we model the observed outcome Y_i at location $\mathbf{s}_i$ as the sum of an intercept μ, linear trend with coefficients β, a spatial Gaussian process $f(\mathbf{s})$, a constant treatment effect τ in the treatment region, and iid normal noise ϵ.

$$
\begin{aligned}
Y_i &= \mu + \mathbf{s}_i^\mathsf{T} \beta + f(\mathbf{s}_i) + \tau Z_i + \epsilon_i \\
f &\sim \mathcal{GP}\left(0, k(\mathbf{s}, \mathbf{s}')\right) \\
k(\mathbf{s}, \mathbf{s}') &= \sigma_{GP}^2 \exp\left(-\frac{\left(\mathbf{s} - \mathbf{s}'\right)^\mathsf{T}\left(\mathbf{s} - \mathbf{s}'\right)}{2\ell^2}\right) \\
\epsilon_i &\overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\epsilon^2\right)
\end{aligned}
\tag{2}
$$

f is a smooth surface covering all of $\mathcal{A}$, specified as a Gaussian Process with squared exponential covariance kernel k with lengthscale ℓ and variance σ_{GP}^2. The squared exponential kernel is frequently used in spatial settings to model smoothly varying quantities. This model implies a constant treatment effect assumption $Y_{iT} = \tau + Y_{iC}$ for all units at all locations.

2.3 Gaussian process model

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 powerful method for fitting univariate 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, linear trend with coefficients β, 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.

$$Y_{iT} = \underbrace{m_T + \mathbf{s}_i^\mathsf{T} \beta_T + f_T(\mathbf{s}_i)}_{g_T(\mathbf{s}_i)} + \epsilon_i$$

$$Y_{iC} = \underbrace{m_C + \mathbf{s}_i^\mathsf{T} \beta_C + f_C(\mathbf{s}_i)}_{g_C(\mathbf{s}_i)} + \epsilon_i \qquad (3)$$

$$f_T, f_C \overset{\perp\!\!\!\perp}{\sim} \mathcal{GP}\left(0, k(\mathbf{s}, \mathbf{s}')\right)$$

$$k(\mathbf{s}, \mathbf{s}') = \sigma_{GP}^2 \exp\left(-\frac{(\mathbf{s} - \mathbf{s}')^\mathsf{T}(\mathbf{s} - \mathbf{s}')}{2\ell^2}\right)$$

Note that the treatment effect τ is not included explicitly in the model. Instead, the treatment effect at a location $\mathbf{s}$ is derived as the difference between the two (noise-free) surfaces g_T and g_C.

$$\tau(\mathbf{s}) = \left[m_T + \mathbf{s}^\mathsf{T}\beta_T + f_T(\mathbf{s})\right] - \left[m_C + \mathbf{s}^\mathsf{T}\beta_C + f_C(\mathbf{s})\right] \qquad (4)$$

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.

2.4 Inference

We proceed by placing normal priors on m_T, m_C and β. The model specification (3) can then be used to obtain covariances between the observations and these parameters. In fact, $(Y_C, Y_T, f_T, f_C, m_C, m_T, \beta_T, \beta_C) \mid \ell, \sigma_{GP}, \sigma_\epsilon$ is multi-variate normal, and so the distribution of any variable conditioned on the others can be obtained analytically and easily computed.

[I think the estimand should be made clear before the model is introduced (move up)]

Because the treatment and control regions do not overlap, inference on the treatment effect is only measurable near the border. In the classical one-dimensional regression discontinuity design, the estimand is therefore defined at the border $x = b$:

$$\tau = \lim_{x \downarrow b} \mathbb{E}\left[y \mid X = s\right] - \lim_{x \uparrow b} \mathbb{E}\left[y \mid X = x\right] = \mathbb{E}\left[Y_{iT} \mid X_i = b\right] - \mathbb{E}\left[Y_{iC} \mid X_i = b\right] \qquad (5)$$

Analogously, we focus on the treatment effect at the border $\mathcal{B}$ between the treatment and control regions, as defined in (Imbens and Zajonc, 2011) and (Keele et al., 2017).

$$\tau : \mathcal{B} \to \mathbb{R}, \ \tau(\boldsymbol{b}) = \mathbb{E}\left[Y_{iT} - Y_{iC} \mid \mathbf{s}_i = \boldsymbol{b}\right] \qquad (6)$$

Note that $\mathcal{B}$ is a one-dimensional manifold embedded in $\mathcal{A}$. 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 of R "sentinel" units distributed along the border $\boldsymbol{b} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_R\}$, $\boldsymbol{b}_i \in \mathcal{B}$. The extrapolation step then proceeds mechanically through multivariate-normal theory.

$$g_T(\boldsymbol{b}) \mid Y_T, S_T, \ell, \sigma_{GP}, \sigma_\epsilon \sim \mathcal{N}\left(\mu_{\boldsymbol{b}|T}, \Sigma_{\boldsymbol{b}|T}\right)$$

$$\mu_{\boldsymbol{b}|T} \equiv \text{cov}\left(g_T(\boldsymbol{b}), Y_T\right) \text{cov}\left(Y_T\right)^{-1} Y_T \qquad (7)$$

$$\Sigma_{\boldsymbol{b}|T} \equiv \text{cov}\left(g_T(\boldsymbol{b})\right) - \text{cov}\left(g_T(\boldsymbol{b}), Y_T\right) \text{cov}\left(Y_T\right)^{-1} \text{cov}\left(Y_T, g_T(\boldsymbol{b})\right)$$

with all the covariance matrices derived from the model specification. Analogously predictions for $g_C(\boldsymbol{b})$ are obtained using the data in the control region, and denote their posterior mean and covariance as $\mu_{\boldsymbol{b}|C}$ and $\Sigma_{\boldsymbol{b}|C}$. Since the two surfaces are modeled as independent, the treatment effect $\tau(\boldsymbol{b}) = g_T(\boldsymbol{b}) - g_C(\boldsymbol{b})$ along the border is also multivariate normal with posterior mean and covariance

$$\mu_{\boldsymbol{b}|Y} = \mathbb{E}\left(\tau(\boldsymbol{b}) \mid Y_T, Y_C\right) = \mu_{\boldsymbol{b}|T} - \mu_{\boldsymbol{b}|C}$$
$$\Sigma_{\boldsymbol{b}|Y} = \mathrm{cov}\left(\tau(\boldsymbol{b}) \mid Y_T, Y_C\right) = \Sigma_{\boldsymbol{b}|T} + \Sigma_{\boldsymbol{b}|C}. \tag{8}$$

2.5 1GP [to be deleted]

By modeling the spatial variation using Gaussian processes, we can leverage the properties of multivariate normals to obtain analytical forms for the estimate of the treatment effect.

We proceed by placing normal priors on μ, β and τ. The model specification can then be used to obtain covariances between the observations and these parameters. In fact, $(Y, \tau, \mu, \beta) \mid \ell, \sigma_{GP}, \sigma_\epsilon$ is multi-variate normal with variance-covariance given by

$$\tau \sim \mathcal{N}\left(0, \sigma_\tau^2\right)$$
$$\mu \sim \mathcal{N}\left(0, \sigma_\mu^2\right)$$
$$\beta \sim \mathcal{N}\left(0, \sigma_\beta^2\right)$$
$$\mathrm{cov}(Y_i, \tau) = \sigma_\tau^2 Z_i \tag{9}$$
$$\mathrm{cov}(Y_i, \mu) = \sigma_\mu^2$$
$$\mathrm{cov}(Y_i, \beta) = \sigma_\beta^2 \mathbf{s}^\mathsf{T}\mathbf{s}$$
$$\mathrm{cov}(Y_i, Y_j) = \sigma_\mu^2 + \sigma_\tau^2 Z_i Z_j + \sigma_\beta^2 \mathbf{s}_i^\mathsf{T}\mathbf{s}_j + k(\mathbf{s}_i, \mathbf{s}_j) + \delta_{ij}\sigma_\epsilon^2$$
$$\mathrm{cov}(Y_i, f(\mathbf{s}_j)) = \mathrm{cov}(f(\mathbf{s}_i), f(\mathbf{s}_j)) = k(\mathbf{s}_i, \mathbf{s}_j)$$

Multi-variate normal theory then allows us to condition any of these variables on the others. We are particularly interested in the posterior distribution $\tau \mid Y, \ell, \sigma_{GP}, \sigma_\epsilon$ which is given by

$$\tau \mid Y, \ell, \sigma_{GP} \sim \mathcal{N}\left(\mathrm{cov}\left(Y, \tau\right)^\mathsf{T}\mathrm{cov}\left(Y\right)^{-1}Y, \sigma_\tau^2 - \mathrm{cov}\left(Y, \tau\right)^\mathsf{T}\mathrm{cov}\left(Y\right)^{-1}\mathrm{cov}\left(Y, \tau\right)\right) \tag{10}$$

To proceed computationally, we define the treatment indicator vector $\mathbb{I}_T$ with ith entry equal to Z_i, the spatial covariate $n \times 2$ matrix S with ith row $\mathbf{s}_i$, and the $n \times n$ kernel covariance matrix $\mathbf{K}$ having entries $\mathbf{K}_{ij} = k(\mathbf{s}_i, \mathbf{s}_j)$. The posterior mean and variance are then easily computed.

$$\mathbb{E}\left(\tau \mid Y, \ell, \sigma_{GP}, \sigma_\epsilon\right) = \sigma_\tau^2 \mathbb{I}_T^\mathsf{T}\left\{\sigma_\mu^2 + \sigma_\tau^2 \mathbb{I}_T \mathbb{I}_T^\mathsf{T} + \sigma_\beta^2 SS^\mathsf{T} + \mathbf{K} + \sigma_\epsilon^2 \mathbf{I}\right\}^{-1} Y$$
$$\mathrm{var}\left(\tau \mid Y, \ell, \sigma_{GP}, \sigma_\epsilon\right) = \sigma_\tau^2 - \sigma_\tau^2 \mathbb{I}_T^\mathsf{T}\left\{\sigma_\mu^2 + \sigma_\tau^2 \mathbb{I}_T \mathbb{I}_T^\mathsf{T} + \sigma_\beta^2 SS^\mathsf{T} + \mathbf{K} + \sigma_\epsilon^2 \mathbf{I}\right\}^{-1} \mathbb{I}_T \tag{11}$$

What remains is the inference on the hyperparameters σ_ϵ, σ_{GP} and ℓ. The two approaches typically taken in modern spatial statistics are either to maximize the marginal likelihood of Y as a function of those three parameters, or to assign them a prior and take a Bayesian approach, requiring that the posterior of τ be integrated over those parameters. The compromise is clear: the Bayesian approach incorporates the uncertainty in the hyperparameters, thus giving more reliable inference on τ, but maximizing the marginal likelihood has a much lower computation cost. Therefore, we recommend taking the Bayesian approach whenever computationally possible, and maximizing the marginal likelihood when the data is larger.

3 Handling covariates

The Gaussian Process specification 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 $D\gamma$ on the $n \times p$ matrix of covariates D. In the spirit of ridge regression, we recommend placing

a normal prior $\mathcal{N}(0, \sigma_\gamma^2)$ on the regression coefficients. This preserves the multivariate normality of the model, with the simple addition of a term $\sigma_\gamma^2 D^\intercal D$ to the covariance of Y.

Our model becomes

$$
\begin{aligned}
Y_{iT} &= \underbrace{m_T + \mathbf{d}_i^\intercal \gamma + \mathbf{s}_i^\intercal \beta_T + f_T(\mathbf{s}_i)}_{g_T(\mathbf{s}_i)} + \epsilon_i \\
Y_{iC} &= \underbrace{m_C + \mathbf{d}_i^\intercal \gamma + \mathbf{s}_i^\intercal \beta_C + f_C(\mathbf{s}_i)}_{g_C(\mathbf{s}_i)} + \epsilon_i \\
f_T, f_C &\overset{\perp\!\!\!\perp}{\sim} \mathcal{GP}\left(0, k(\mathbf{s}, \mathbf{s}')\right) \\
k(\mathbf{s}, \mathbf{s}') &= \sigma_{GP}^2 \exp\left(-\frac{(\mathbf{s} - \mathbf{s}')^\intercal (\mathbf{s} - \mathbf{s}')}{2\ell^2}\right) \\
\gamma_j &\overset{\perp\!\!\!\perp}{\sim} \mathcal{N}\left(0, \sigma_\gamma^2\right) \text{ for } j = 1, 2, \dots, p
\end{aligned}
\tag{12}
$$

Unfortunately, the linear term induces a covariance between the treatment and control region. 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.

4 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 $\tau(\mathbf{s})$, with weight function $w_\mathcal{B}(\mathbf{s})$ defined everywhere on the border $\mathcal{B}$. The weighted mean integral can be approximated as a sum at the sentinels $\boldsymbol{b}_{1:R}$.

$$
\begin{aligned}
\tau^w &= \frac{\oint_\mathcal{B} w_\mathcal{B}(\mathbf{s}) \tau(\mathbf{s}) \, d\mathbf{s}}{\oint_\mathcal{B} w_\mathcal{B}(\mathbf{s}) \, d\mathbf{s}}, \\
&\approx \frac{\sum_{i=0}^R w_\mathcal{B}(\boldsymbol{b}_i) \tau(\boldsymbol{b}_i)}{\sum_{i=0}^R w_\mathcal{B}(\boldsymbol{b}_i)}.
\end{aligned}
\tag{13}
$$

We have shown the posterior distribution of $\tau(\mathbf{s})$ at the sentinels to be multivariate normal, with mean $\mu_{\boldsymbol{b}|Y}$ and covariance $\Sigma_{\boldsymbol{b}|Y}$ given in (8). Since τ^w is a linear transformation of $\tau(\boldsymbol{b})$, its posterior is also multivariate normal, with mean $\mu_{\tau^w|Y}$ and covariance $\Sigma_{\tau^w|Y}$ given by

$$
\begin{aligned}
\mu_{\tau^w|Y} &= \frac{w_\mathcal{B}(\boldsymbol{b})^\intercal \mu_{\boldsymbol{b}|Y}}{w_\mathcal{B}(\boldsymbol{b})^\intercal \mathbf{1}_R} \\
\Sigma_{\tau^w|Y} &= \frac{w_\mathcal{B}(\boldsymbol{b})^\intercal \Sigma_{\boldsymbol{b}|Y} w_\mathcal{B}(\boldsymbol{b})}{\left(w_\mathcal{B}(\boldsymbol{b})^\intercal \mathbf{1}_R\right)^2}
\end{aligned}
\tag{14}
$$

where $w_\mathcal{B}(\boldsymbol{b})$ is the vector of weights evaluated at the sentinels. For each estimator obtained in (14) as a weighted mean of $\mu_{\boldsymbol{b}|Y}$, we consider the "natural" estimand to be the same weighted mean applied to the truth $\tau(\mathbf{s})$, given by (13).

An alternative perspective on these estimators is given by the weights induced on the observations. Indeed, combining equations (7), (8), and (14), we obtain that the posterior mean of τ^w is a linear combination

$$\mathbb{E}\left(\tau^w \mid Y\right) = w_T^\mathsf{T} Y_T + w_C^\mathsf{T} Y_C \tag{15}$$

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

$$
\begin{aligned}
w_T &= \frac{1}{w_{\mathcal{B}}(\boldsymbol{b})^\mathsf{T} \mathbf{1}_R} K_{TT}^{-1} K_{Tb} w_{\mathcal{B}}(\boldsymbol{b}), \text{ and} \\
w_C &= -\frac{1}{w_{\mathcal{B}}(\boldsymbol{b})^\mathsf{T} \mathbf{1}_R} K_{CC}^{-1} K_{Cb} w_{\mathcal{B}}(\boldsymbol{b}),
\end{aligned}
\tag{16}
$$

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 $w_{\mathcal{B}}(\mathbf{s})$, and explore interpretation, advantages, and disadvantages. The six border weight functions and unit weights for a simulated example are shown in Figure 3(a-f). Interestingly, the projected RDD estimator can also be written in the form (15), and so we also illustrate the resulting unit weights in Figure 3(g).

4.1 Uniform ATE

The simplest choice is uniform weights $w_{\mathcal{B}}(\mathbf{s}) = 1$, a seemingly reasonable and unopinionated decision. We estimate τ^{UNIF}, the uniformly weighted mean of $\tau(\mathbf{s})$ along the border, by averaging the entries of the mean posterior at the sentinels. Following (13) and (14):

$$
\begin{aligned}
\tau^{\text{UNIF}} &\equiv \frac{\oint_{\mathcal{B}} \tau(x)\, d\mathbf{s}}{\oint_{\mathcal{B}} dx} \\
\tau^{\text{UNIF}} \mid Y_T, Y_C, \sigma_{\text{GP}}, \sigma_\epsilon, \ell &\sim \mathcal{N}\left(\mu_{\tau^{\text{UNIF}}\mid Y}, \Sigma_{\tau^{\text{UNIF}}\mid Y}\right) \\
\mu_{\tau^{\text{UNIF}}\mid Y} &= \left(\mathbf{1}^\mathsf{T} \mu_{b\mid Y}\right)/R \\
\Sigma_{\tau^{\text{UNIF}}\mid Y} &= \left(\mathbf{1}^\mathsf{T} \Sigma_{b\mid Y} \mathbf{1}\right)/R^2
\end{aligned}
\tag{17}
$$

Note that if the sentinels are not evenly spaced, then each entry needs to be re-weighted by the length of the border that the sentinel occupies. The uniformly weighted estimand takes on a geometric interpretation: equal-length segments of the border are given equal weight. The procedure is mathematically sound, but the choice of this estimator suffers from issues that we describe and address in the next two sections.

4.2 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, $\tau(\mathbf{s})$ 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: τ^{UNIF} will drown them in noise.

We can address this issue by weighting the treatment effect at each sentinel location by the local density. That is we choose $w_{\mathcal{B}}(\mathbf{s}) = \rho(\mathbf{s})$, 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.
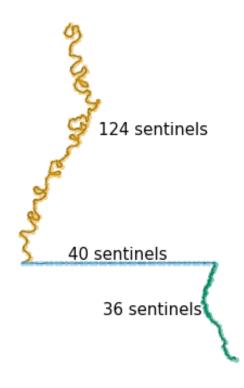
9

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

4.3 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 $\Sigma_{b|Y}$: 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 use of the inverse-variance weighted mean τ^{INV}

$$\tau^{\text{INV}} \mid Y_{\text{T}}, Y_{\text{C}}, \sigma_{\text{GP}}, \sigma_\epsilon, \ell \sim \mathcal{N}\left(\mu_{\tau^{\text{INV}}|Y}, \Sigma_{\tau^{\text{INV}}|Y}\right),$$

$$\mu_{\tau^{\text{INV}}|Y} = \left(\mathbf{1}^{\mathsf{T}} \Sigma_{b|Y}^{-1} \mu_{b|Y}\right) / \left(\mathbf{1}^{\mathsf{T}} \Sigma_{b|Y}^{-1} \mathbf{1}\right), \tag{18}$$

$$\Sigma_{\tau^{\text{INV}}|Y} = 1 / \left(\mathbf{1}^{\mathsf{T}} \Sigma_{b|Y}^{-1} \cdot \mathbf{1}\right)$$

This estimator efficiently extracts the information from the posterior treatment effect, and is guaranteed to yield the lowest posterior variance amongst weighted averages of the form (13). 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. The approach is in keeping with the philosophy of regression discontinuity designs: we let information be our guide when weighting the mean treatment effect, just like it guided the regression discontinuity design to only focus on the treatment effect at the border. This estimand isn't chosen by the scientist, but rather it is dictated by the limitations of the observed data.

The estimand is still a weighted mean, with weights for the sentinels given by $w_{\mathcal{B}}(\boldsymbol{b}) = \Sigma_{b|Y}^{-1} \mathbf{1}$. This can put negative weights on some sentinels, as seen for example in Figure 3(c), and generally this estimand doesn't lend itself to an intuitive scientific interpretation. 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 unorthodox in the causal inference literature, analogous "estimands of convenience" have been proposed in other settings, for example matching methods that exclude some unmatched units from the analysis, or (Li et al., 2016) in the context of balancing treatment and control populations with little overlap in their covariate distributions.

4.4 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. In the spirit of a finite-population estimate, each unit figuratively gets one vote, though instead of a direct vote, it has a representative at the border. 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 (17), replacing the cliff-face mean vector $\mu_{b|Y}$ and covariance matrix $\Sigma_{b|Y}$ with equivalent quantities obtained at the projected sentinels, to obtain the posterior mean and covariance of τ^{PROJ}:

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

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

$$\Sigma_{\tau^{\text{PROJ}}|Y} = \frac{1}{(n_{\text{C}} + n_{\text{T}})^2} \sum_{i=1}^{n_{\text{T}}+n_{\text{C}}} \sum_{j=1}^{n_{\text{T}}+n_{\text{C}}} \text{cov}\left[\tau\left(\text{proj}_{\mathcal{B}}(\mathbf{s}_i)\right), \tau\left(\text{proj}_{\mathcal{B}}(\mathbf{s}_j)\right) \mid Y_{\text{T}}, Y_{\text{C}}, \sigma_{\text{GP}}, \sigma_\epsilon, \ell\right]. \tag{19}$$

The posterior expectations and covariances in (19) can be obtained as in (8), but using the projected sentinels. Note that τ^{PROJ} is still within the class of weighted mean estimands (13), with weight function $w_{\mathcal{B}}(\mathbf{s}) = \sum_{i=1}^{n_{\text{T}}+n_{\text{C}}} \delta\left(\mathbf{s} - \text{proj}_{\mathcal{B}}(\mathbf{s}_i)\right)$, 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 much easier to understand and interpret, and may feel more familiar to practicioners used to finite-population inference.

If there are units very far away from the border, they 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) should be projected onto the border to become sentinels.

4.5 Projected land ATE

In certain applications, estimands that depend on the position of measurements are undesirable, and geography-weighted estimands are more natural. **[LukeB: Citation? Better sentence?]** The "geometry-based" unweighted estimand τ^{UNIF} has the property that each segment of the border has equal weight. If instead, we desire each patch of land to have equal weight, we can borrow the projection method from τ^{PROJ} and apply it to an infinite population of uniform density on both sides of the border. This yields the geography-based projected land ATE τ^{GEO}. To estimate τ^{GEO}, a tight grid of evenly spaced points is first generated within $\mathcal{A}^{\text{T}}$ and $\mathcal{A}^{\text{C}}$. Each point on this grid is then projected onto the border and becomes a sentinel. The 2GP procedure applied to these unevenly spaced sentinels then yields a mean vector and covariance matrix for the treatment effect at these positions. 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} estimator 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 $w_{\mathcal{B}}(\mathbf{s})$ in (13) proportional to the area of $\mathcal{A}^{\text{T}}$ and $\mathcal{A}^{\text{C}}$ that $\mathbf{s}$ is nearest to:

$$w_{\mathcal{B}}(\mathbf{s}) = \int_{\mathcal{A}} \mathbb{I}\left\{\mathbf{s} = \text{proj}_{\mathcal{B}}\left(\mathbf{s}'\right)\right\} d\mathbf{s}' \tag{20}$$

[**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.

4.6 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.

4.7 Wiggly Border Simulation

We illustrate the differences between the four average treatment effect 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. The outcomes Y are simulated from a Gaussian process with squared exponential kernel ($\ell = 0.4$, $\sigma = 0.5$), to which we add a treatment effect $\tau(S_1, S_2) = S_1$.
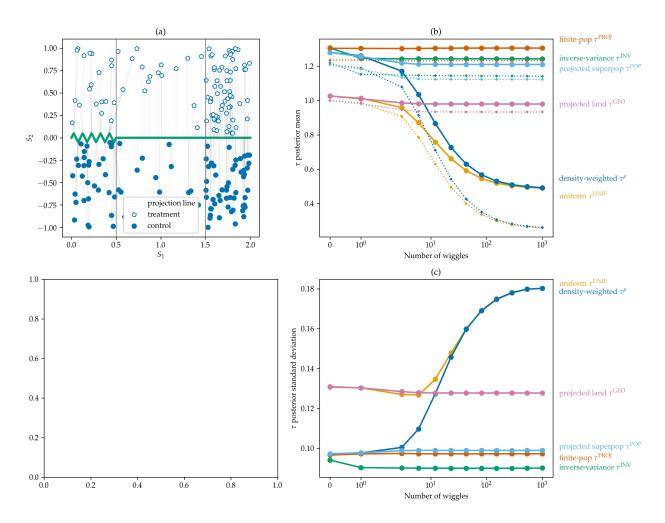
Figure 2: (a) Spatial positions of units and border; (b) Estimator (posterior mean) and estimand (dotted lines of same color) behavior as left border gets wigglier; (c) Posterior standard deviation for each estimator as left border gets wigglier. **(no plot in bottom left yet, final layout TBD)**

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 $\rho = 2.0$
τ	weak	medium	strong

We fit Gaussian processes on each side of the border, using the known hyperparameters and constant mean equal to the empirical mean, and estimate the average treatment effect using the six methods proposed above. The simulation setting is illustrated in Figure 2(a). For each estimator, we show in Figure 2(b) the estimand and posterior mean estimate evolving as the number of wiggles increases. The behavior of the posterior standard deviation is shown in Figure 2(c).

As the border is a straight line and $\mathcal{A}^{\mathrm{T}}$ and $\mathcal{A}^{\mathrm{C}}$ 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, the finite-population and the infinite-population projected estimands are similar, though the difference can be more pronounced with smaller sample popu-
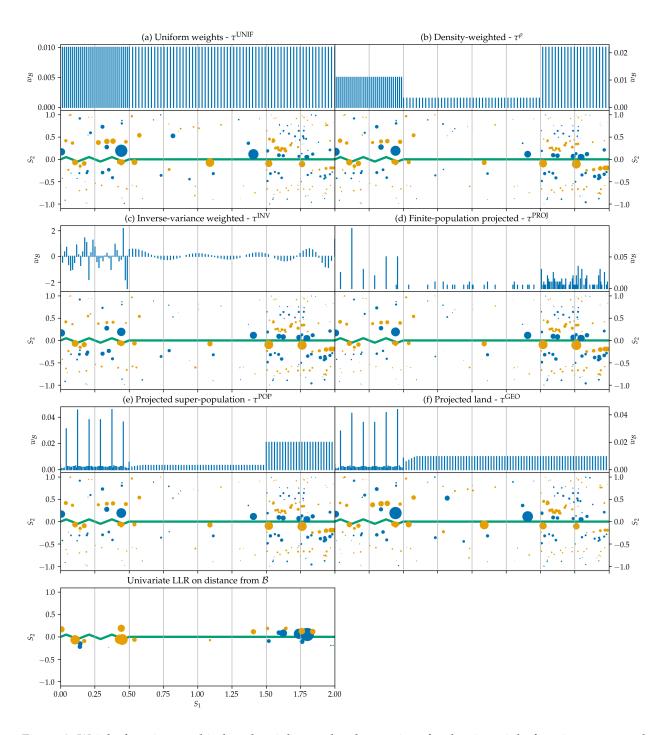
13

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.

lations.

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. However, τ^{INV} is guaranteed to have 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 automatically 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 ($w_{\mathcal{B}}(\mathbf{s})$) and unit weights ($w_{\mathrm{T}}$ 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. 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 will 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.

4.8 Summary of Estimator Properties

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

15

Symbol	Description	Border Topology	Sentinels	Principle	Variance
τ^{GEO}	Projected land ATE	Robust	Projected Grid	Geography-based	High
τ^{POP}	Projected superpopulation ATE	Robust	Projected Grid	Population-based	Low

5 Testing for non-zero effect

Following the 2GP procedure, we might naturally wonder whether we can claim to have detected a significant treatment effect anywhere along 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.

5.1 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 (18). While the posterior is a Bayesian object, we can use it heuristically to derive a pseudo-p-value

$$
\begin{aligned}
Z_0 &\sim \mathcal{N}\left(0, \Sigma_{\tau^{\text{INV}}|Y}\right) \\
p^{\text{INV}} &= \mathbb{P}\left(|Z_0| > \left|\mu_{\tau^{\text{INV}}|Y}\right|\right) \\
&= 2\Phi\left(-\frac{\left|\mu_{\tau^{\text{INV}}|Y}\right|}{\sqrt{\Sigma_{\tau^{\text{INV}}|Y}}}\right)
\end{aligned}
\tag{21}
$$

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.

5.2 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

$$
t = \log \mathbb{P}\left(Y_T, Y_C \mid \mathcal{M}_1\right) - \log \mathbb{P}\left(Y_T, Y_C \mid \mathcal{M}_0\right)
\tag{22}
$$

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

$$
t^* = \log \mathbb{P}\left(Y_T^*, Y_C^* \mid \mathcal{M}_1\right) - \log \mathbb{P}\left(Y_T^*, Y_C^* \mid \mathcal{M}_0\right)
\tag{23}
$$

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.

$$p^{lik} = \mathbb{P}\left(t^* > t_{obs} \mid \mathcal{M}_0\right) \tag{24}$$

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.

5.3 "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 posterior treatment effect along the border, approximated in (8) by its mean vector $\mu_{b|Y}$ and covariance matrix $\Sigma_{b|Y}$ at the sentinel positions b. We will use μ and Σ as shorthand throughout this section. If a k-vector y has multivariate distribution $\mathcal{N}(\mu, \Sigma)$, with mean vector μ known and covariance Σ known, then under the null hypothesis that $\mu = 0$, the test statistic $y^\mathsf{T}\Sigma^{-1}y$ has distribution χ_k^2. See for example (**?**) section 5.2.2 for a classical derivation of this test. This suggests that we could use $S^2 = \mu^\mathsf{T}\Sigma^{-1}\mu$ as a test statistic, and obtain a p-value from a χ_R^2 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

$$
\begin{aligned}
p &= \frac{1}{B}\sum_{t=1}^{T}\mathbb{I}\left\{S_{(b)}^2 < S^2\right\}, \\
S_{(b)}^2 &= \left(\mu_{(b)}\right)^\mathsf{T}\Sigma^{-1}\mu_{(b)}
\end{aligned}
\tag{25}
$$

where $\mu_{(b)}$ is the result of applying (8) to $Y_T^{(b)}$ and $Y_C^{(b)}$, themselves drawn from $\mathcal{M}_0$ at the same locations 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.

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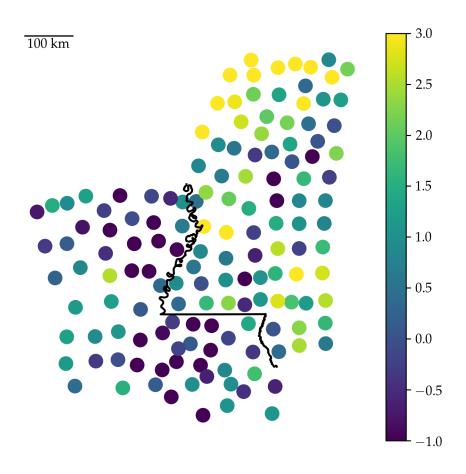
Figure 4: Set-up of the imaginary experiment in Louisiana and Mississippi. Each unit is at the centroid of a county. The colors indicated the observed outcomes in one draw of the simulation under $\tau = 1.5$. In this particular run, the p-values were 0.0016, 0.0018, and 0.0013 for the mLL, χ^2, and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

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

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 C.

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). Therefore, in studies where the main interest is an overall (average) increase or decrease in outcomes, we recommend using the inverse-variance test to maximize power.

5.5 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.

6 Spatial advantage [candidate for removal]

Classical regression discontinuity designs often suffer from low power, requiring many units near the border for inference to be possible. In the spatial RDD setting, we might worry that the situation is worse, as geographical datasets with many units packed along the border are uncommon. In geographical settings, each unit (e.g. household or counties) normally takes up space, so there is a limit to how densely packed units can be near the border. And boundaries often include sparsely populated segments, e.g. running through parks, industrial areas, or farmland. The intuition that spatial RDDs will therefore suffer from low power is correct, inasmuch as at any given point along the border, the posterior variance of $\tau(\mathcal{B})$ will typically be high. But once we pool the information into an average treatment effect, or perform a sharp test, spatial RDDs can be more powerful than classical RDDs, with the same number of units at the same distance from the border.

We illustrate this statement once more with the Louisiana-Mississippi example. The variance of the inverse-variance weighted treatment effect τ^{INV} is thence only a function of the positions of the units, available analytically by plugging the posterior variance (8) into the inverse-variance estimator (18). Following this procedure, we obtain a posterior standard deviation of the average treatment effect of 0.31. We then

19

create a one-dimensional regression discontinuity design for the same setting, by using each unit's distance from the border as the covariate x, the distribution of which is shown in Figure 4(b). Following the exact same 2GP procedure with the same hyperparameters as in the spatial setting, and with a discontinuity at $x = 0$, we again compute the posterior standard deviation of the treatment effect at the border (now a single number rather than a continuous function) , this time obtaining 0.58. This higher figure indicates that, perhaps counter-intuitively, the spatial experiment actually has more power than its one-dimensional analog.

To gain intuition about the higher power of the spatial RDD, we turn to the interpretation of regression discontinuity designs as natural experiments [need reference]. 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. This in turn allows us to claim that a natural randomized experiment took place near the border, with treatment and control units coming from the same population. We can extend this interpretation to the spatial setting, by conceiving of multiple correlated experiments taking place all along the border. The average treatment effect estimator then pools the information supplied by all of these experiments. The question then becomes: do we get more powerful inference by grouping all the units into a single experiment, or by spreading them along a multitude of weaker experiments? There are two sources of uncertainty in our model: the observation noise ϵ_i, and the underlying processes g_T and g_C. Adding more units to a single experiment allows us to cancel out more of the observation noise, but if the new units aren't added closer to the discontinuity, uncertainty always remains in g_T and g_C. In the spatial setting, however, we observe multiple realizations of the Gaussian process, and therefore do not suffer from the same diminishing returns.

7 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 twelve 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?

7.1 Preprocessing

In order to model the property sale prices with a stationary Gaussian process, we need to obtain their location on a Euclidean grid. 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 as follows, by removing 1. sales of properties without a reported sale price 1. sales of properties outside of the residential building class categories ("one family dwellings", "two family dwellings", "three family dwellings", "tax class 1 condos", "coops - walkup apartments", "coops - elevator apartments", "condos - walkup apartments", "condos - elevator apartments", "condos - 2-10 unit residential", "condo coops"), 2. any sale with missing data in the sale price, square footage, property covariates, geographical coordinates (due to failed geocoding), 3. sales outside of any NYC school district, 4. properties smaller than 100 sq ft, and 5. outliers in the price per square foot.
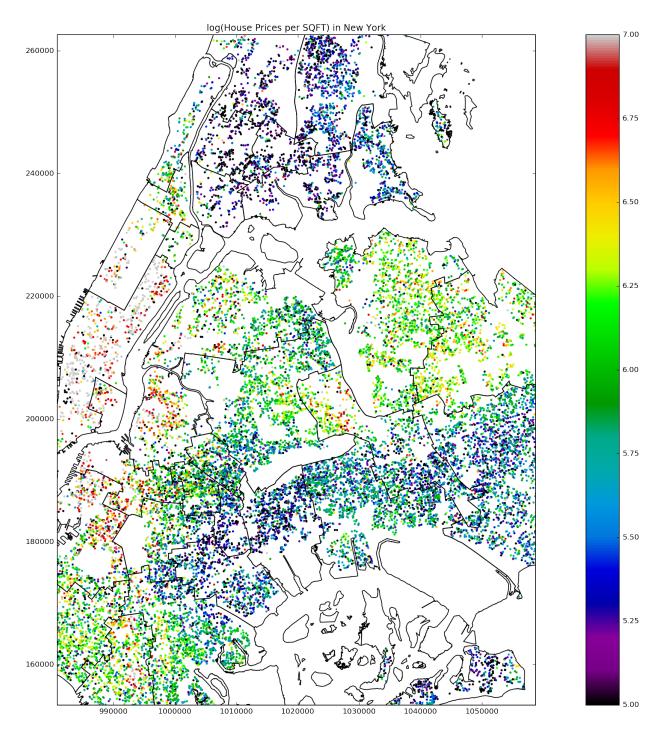
Figure 5: sales map

7.2 Exploratory analysis

7.3 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]:

$$
\begin{aligned}
Y_i = \log\left(\frac{\texttt{SalePrice}_i}{\texttt{SQFT}_i}\right) = {} & \mu_{\texttt{District[i]}} + \beta_{1,\texttt{TaxClass[i]}} + \beta_{2,\texttt{BuildingClass[i]}} \\
& + f_{\texttt{District[i]}}(\mathbf{s}_i) + \epsilon_i \\
\epsilon_i \sim {} & \mathcal{N}\left(0, \sigma_y^2\right) \\
\mu_j \sim {} & \mathcal{N}\left(\bar{Y}_j, \sigma_\mu^2\right) \\
\beta_{1j}, \beta_{2j} \sim {} & \mathcal{N}\left(0, \sigma_\beta^2\right) \\
f_j(\mathbf{s}_i) \sim {} & \mathcal{GP}\left(0, k(\mathbf{s}, \mathbf{s}')\right) \\
k(\mathbf{s}, \mathbf{s}') = {} & \sigma_{GP}^2 \exp\left\{-\frac{(\mathbf{s} - \mathbf{s}')^\mathsf{T}(\mathbf{s} - \mathbf{s}')}{2\ell^2}\right\}
\end{aligned}
\tag{26}
$$

A visual inspection of the house sales map above 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 causal effect solely to the difference in 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 an interpretation that is not directly supported by the data. A histogram 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.

7.4 parameter optimization

We initially fit the hyperparameters σ_β, σ_{GP}, ℓ and σ_ϵ by optimizing the marginal log-likelihood of the data within a single district. We choose district 27 as it contains more sales. We hold σ_μ fixed to 10 to give the district means μ_j a fairly uninformative prior. The fitted hyperparameters were $\sigma_\epsilon = 0.4179$, $\sigma_{GP} = 0.2426$, $\sigma_\beta = 0.1306$, and $\ell = 3378.5800$ ft.

7.5 cliff face

We seek the treatment effect function $\tau(\mathcal{B})$ between the two districts. We could proceed by computing the joint predictive distributions $g_T(\boldsymbol{b}), g_C(\boldsymbol{b}) \mid Y_T, Y_C, \sigma_\beta, \sigma_{GP}, \ell, \sigma_\epsilon$, 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 $g_T(\boldsymbol{b})$ and $g_C(\boldsymbol{b})$ so they become independent multivariate normal distributions $g_T(\boldsymbol{b}) \mid Y_T, \hat{\beta}, \sigma_{GP}, \ell, \sigma_\epsilon$ and $g_C(\boldsymbol{b}) \mid Y_C, \hat{\beta}, \sigma_{GP}, \ell, \sigma_\epsilon$. 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 $g_C(\boldsymbol{b})$ and $g_T(\boldsymbol{b})$, which are uncorrelated conditional on $\beta = \hat{\beta}$, we can now take their difference according to the procedure outline in section 2.3, to obtain the posterior distribution of the cliff-face $\tau(\boldsymbol{b})$ obtained at the sentinel locations. The cliff-face is shown in Figure 7, and shows that the estimated $\tau(\boldsymbol{b})$ 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, and therefore it isn't clear that this effect isn't due to random variation.
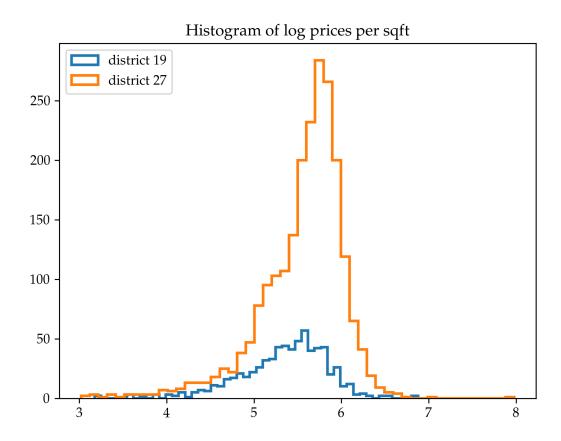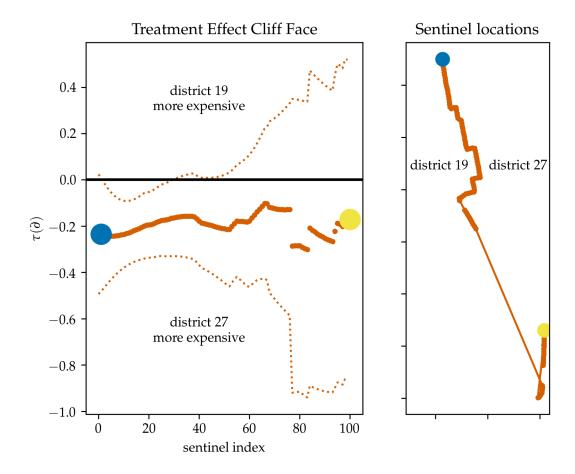
Figure 6:

Figure 7: NYC cliff face

The treatment effect can also be visualized directly in Figure 8 as the difference between the two log-price mean surfaces $g(\mathbf{s})$. This pictures also gives a better sense of the important spatial variation in prices captured by the model, which explains the wide credible envelope in the cliff face, despite the large number of sales in both districts.

7.6 Average Log-Price Increase

The cliff-face plot shows a negative treatment effect everywhere along the border, which can be averaged by the estimators we developed in section 4. The most obvious approach is to take an unweighted mean at each equispaced sentinel, which has posterior distribution

$$\tau^{\text{UNIF}} \mid Y_{\text{T}}, Y_{\text{C}}, \sigma_{\text{GP}}, \sigma_\epsilon, \hat{\beta}, \ell \sim \mathcal{N}\left(-0.20, 0.09^2\right) \text{ and tail probability}$$

$$\mathbb{P}\left(\tau^{\text{UNIF}} > 0 \mid Y_{\text{T}}, Y_{\text{C}}, \sigma_{\text{GP}}, \sigma_\epsilon, \hat{\beta}, \ell\right) = 1.48\%.$$
(27)

The inverse-variance weighted mean estimator is robust to changes in the border topology, and gives higher weight to sections of the border where the difference in house prices is easier to measure. It is guaranteed to minimize the posterior variance amongst weighted mean estimators, which is reflected here by the narrower posterior distribution
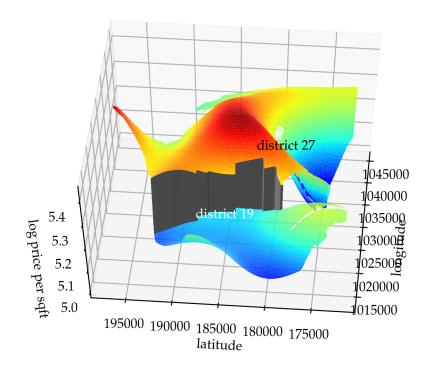
Figure 8: NYC surface plot

$$\tau^{\text{INV}} \mid Y_{\text{T}}, Y_{\text{C}}, \sigma_{\text{GP}}, \sigma_{\epsilon}, \hat{\beta}, \ell \sim \mathcal{N}\left(-0.20, 0.05^2\right) \text{ and reduced tail probability}$$

$$\mathbb{P}\left(\tau^{\text{INV}} > 0 \mid Y_{\text{T}}, Y_{\text{C}}, \sigma_{\text{GP}}, \sigma_{\epsilon}, \hat{\beta}, \ell\right) = 0.01\%. \tag{28}$$

The posterior mean estimate corresponds to a 21% increase in price per square foot from district 19 to district 27.

All estimators [except τ^ρ for now] are shown in Table 4. For each estimand, we show the mean and standard deviation of its posterior distribution, and the tail probability $\mathbb{P}\left(\tau > 0 \mid Y_{\text{T}}, Y_{\text{C}}, \sigma_{\text{GP}}, \sigma_{\epsilon}, \hat{\beta}, \ell\right)$ of the average treatment being greater than zero.

Table 4: Average treatment effect estimators for New York school district house prices example.

Estimand	Posterior Mean	Posterior Standard Deviation	Posterior Tail Prob
τ^{UNIF}	-0.20	0.09	1.48%
τ^{INV}	-0.20	0.05	0.01%
τ^{PROJ}	-0.21	0.08	0.46%
τ^{GEO}	-0.19	0.10	2.54%
τ^{POP}	-0.19	0.06	0.04%

7.7 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 5. In applied settings, running multiple tests invalidates their results, 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	0.003
mLL bootstrap	0.001
τ^{INV} uncalibrated	0.0003
τ^{INV} calibrated	0.0004

[no longer true] The three tests tell very different stories. The χ^2 test fails to reject the null hypothesis even at the $\alpha = 0.1$ level. This strongly contradicts the inverse-variance test and its low p-value of 0.0005, backed by the likelihood-ratio test with $p = 0.0015$. The possibility of such a contradiction was anticipated in section XX, where we saw that the χ^2 has the lowest power of the three tests, and therefore could easily fail to reject an effect that is easily detected by the inverse-variance test. Because the inverse-variance test has the highest power in detecting constant treatment effects, we would recommend its use in applications — such as this one — where a very heterogenous treatment effect is not expected.

7.7.1 placebo tests

To assess the validity of the three tests, we apply the placebo tests that we devised in Section X. 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
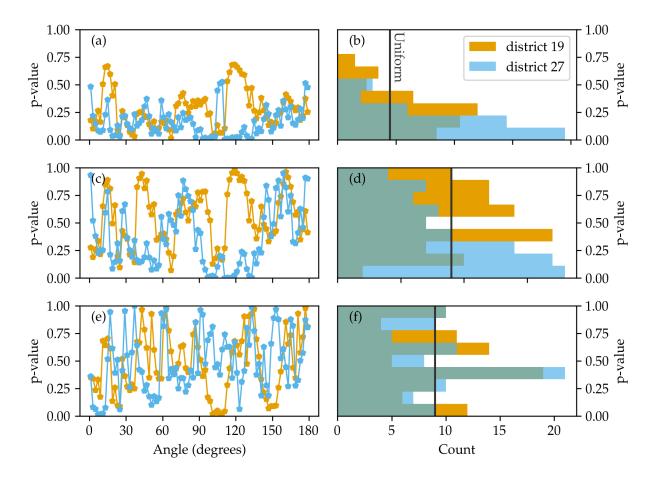
Figure 9: Placebo tests for significance tests applied to NYC school district house price example, applied within districts 19 and 27. The three rows respectively show results for the marginal log-likelihood bootstrap test, chi-squared bootrap test, and calibrated inverse-variance test. The first column 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.

see a uniform distribution of placebo p-values. However, these tests will be highly correlated, and so the low effective sample size could lead to some apparent departures from uniformity. There is in fact visible autocorrelation in the graphs of placebo p-values as a function of angle.

The mLL placebo p-values show a pronounced bias towards low values. This seems to confirm our concern that the marginal log-likelihood may be sensitive to features of the data other than the discontinuity at the border. In particular, model misspecification, which is a big 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 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 argument for its use in most applications.

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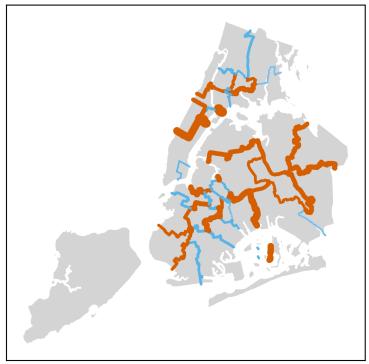
Figure 10: Pairwise effect size between adjacent districts. The thickness of each border is proprtional to the posterior effect size defined as $\mathbb{E}(\tau^{\mathrm{INV}} \mid Y)/(\mathrm{var}(\tau^{\mathrm{INV}} \mid Y))^{1/2}$, indicating the strength of the evidence towards a discontinuity in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

7.8 pairwise treatment effect (all districts)

8 Conclusion

- GeoRDDs are popping up in a lot of places
- Few methods so far to analyze them
- We proposed a general framework
- And an method that fits within this framework
- The primary output of the method is an estimate of the treatment effect everywhere along the border

    - It's a functional estimand!

- But that raises further question: what should we make of this estimate?
- We looked at two possibilities: averaging and hypothesis-testing
- Averaging:

    - many possible ways to average
    - general framework: estimator and estimand are weighted means along the border
    - we recommend either the inverse-variance weighted ATE for its high power/efficiency
    - or the projected finite-population ATE for its ease of interpretability
    - step back: different philosophies

- Hypothesis testing:

  - 3 approaches
  - sharp null vs. weak null
  - power
  - placebo tests to defend against misspecification

- NYC example

A particularly intriguing aspect of GeoRDDs is the functional estimand defined on a one-dimensional manifold embedded in 2D space. Our approach yields a multivariate normal posterior distribution of the functional treatment effect for a collection of "sentinel" points along the border: the "cliff face" estimate. But this primary output is somewhat unsatisfactory, as it doesn't answer two immediate questions.

Firstly, can we estimate the average treatment effect (ATE)? Defining an "average treatment effect" estimand turns out to have surprising pitfalls. The geometry of GeoRDDs makes the very reasonable idea of integrating the treatment effect linearly along the border quite unsuitable. We showed that this ATE estimand is undesirably sensitive to the topology of the border: straight segments are given less weight than wiggly segments of the border, which is unlikely to be the analyst's intention. A more suitable approach is to weight sentinels by the inverse of the posterior covariance, which yields an ATE estimate that is robust to the topology of the border, and is also the minimum variance ATE estimate, thus fully utilizing the information available in the GeoRDD. But it lacks an easily interpretable estimand, motivated as it is by the properties of its posterior estimator. Another approach is to project observed units onto the border, instead of evenly spaced sentinels, an take the unweighted mean of the treatment effect estimated at those locations. This estimate is also robust to border topology, and the estimand allows for a finite population interpretation that may feel more natural to practicioners.

Secondly, when can we say that we have detected a statistically significant treatment effect? We proposed three tests, one based on the inverse-variance weighted ATE, on a marginal likelihood comparison, and a "chi-squared" test. A tail probability can be derived heuristically from the inverse-variance weighted ATE's posterior distribution, but a calibration step, either through a parametric bootstrap or analytically, is necessary to turn this Bayesian pseudo-p-value into a valid frequentist p-value. Note that the same procedure to obtain a valid test could be followed using any of the average treatment effect estimates defined in section 4. The second test compares the marginal likelihood under the GeoRDD (3) versus a single Gaussian process with the same covariance parameters spanning both regions, the null model. We argue that the marginal likelihood test is sensitive to model misspecification and doesn't explicitly target the detection of a treatment effect at the border, and therefore do not recommend its use. The "chi-squared test" addresses this issue, as it uses a statistic that is calculated directly from the border treatment effect's posterior mean vector and covariance matrix. The test statistic is inspired by a classical chi-squared test for testing whether the mean of a multivariate normal distribution is zero, but in this setting it does not have a χ^2 distribution in this setting. We therefore again use a parametric bootstrap to obtain its null distribution. The "chi-squared" test targets the sharp null hypothesis of no treatment effect anywhere along the border, while the inverse-variance test targets the weak null hypothesis of zero ATE. While the sharp null is more general, we find that the "chi-squared" test has lower power than the inverse-variance test. Therefore, we recommend the use of the inverse-variance test in contexts where the treatment effect, if non-zero, is expected to be mostly positive everywhere, or negative everywhere along the border, or when the average treatment effect is of primary scientific interest.

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 effect is zero everywhere along the border? We have not yet explored this avenue of research.

Datasets that allow for a GeoRDD approach are numerous, but typically tools from classical univariate RDDs (local linear regression), or from causal inference (matching) have been borrowed for this purpose. In particular, the direct use of methods developed for univariate RDDs, treating distance from the border as the forcing variable, is asymptotically biased in simple examples. Our aim was to recognize the importance of the geographical aspect of the problem, and therefore draw from the spatial statistics literature, which brings a rich set of tools designed specifically to model and exploit spatial correlations to obtain more pow-

erful inference. We therefore used Gaussian process regression, known as kriging in the spatial statistics literature, as the foundation of our GeoRDD method.

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.

Because of the need to extrapolate the fitted processes to the border, GeoRDDs are particularly vulnerable to the limitations of Gaussian processes when extrapolating. The distinction between interpolation and extrapolation of spatial models is explored in some depth in (**?**). 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. Note that the estimand is no longer defined as a discontinuity in the expectation function of the potential outcomes, as in (6), but as the difference between two latent processes that are hypothesized by the model.

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   Covariances in 2GP model

[needs a bit of love]

$$m_T, m_C \sim \mathcal{N}\left(0, \sigma_\mu^2\right)$$

$$\beta \sim \mathcal{N}\left(0, \sigma_\beta^2\right)$$

$$\text{cov}(Y_{iT}, m_T) = \sigma_\mu^2$$

$$\text{cov}(Y_{iC}, m_C) = \sigma_\mu^2$$

$$\text{cov}(Y_{iT}, \beta) = \sigma_\beta^2 \mathbf{s}_i^\mathsf{T} \mathbf{s}_i$$

$$\text{cov}(Y_{iC}, \beta) = \sigma_\beta^2 \mathbf{s}_i^\mathsf{T} \mathbf{s}_i \tag{29}$$

$$\text{cov}\left(Y_{iT}, f_T(\mathbf{s}')\right) = k(\mathbf{s}_i, \mathbf{s}')$$

$$\text{cov}\left(Y_{iC}, f_C(\mathbf{s}')\right) = k(\mathbf{s}_i, \mathbf{s}')$$

$$\text{cov}\left(Y_{iT}, f_C(\mathbf{s}')\right) = 0$$

$$\text{cov}\left(Y_{iC}, f_T(\mathbf{s}')\right) = 0$$

$$\text{cov}(Y_{iT}, Y_{jT}) = \sigma_\mu^2 + \sigma_\beta^2 \mathbf{s}_i^\mathsf{T} \mathbf{s}_j + k(\mathbf{s}_i, \mathbf{s}_j) + \delta_{ij} \sigma_\epsilon^2$$

$$\text{cov}(Y_{iT}, Y_{jC}) = \sigma_\beta^2 \mathbf{s}_i^\mathsf{T} \mathbf{s}_j$$

B   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?]

$$
\begin{aligned}
\Sigma_{Y|\beta} &\equiv \text{cov}\left(Y \mid \beta\right) & \text{conditional variance of } Y \\
\text{cov}\left(Y_i, Y_j \mid \beta\right) &= \sigma_\epsilon^2 \delta_{ij} + k\left(\mathbf{s}_i, \mathbf{s}_j\right) \delta_{\texttt{District[i]},\texttt{District[j]}} & \text{(block diagonal)} \\
\Sigma_\beta &\equiv \text{cov}\left(\beta\right) = \sigma_\beta^2 I_p & \text{prior variance of } \beta \\
\Sigma_Y &\equiv \text{cov}\left(Y\right) = \Sigma_{Y|\beta} + D^\mathsf{T} \Sigma_\beta D & \text{unconditional variance of } Y \\
T_\beta &= D^\mathsf{T} \Sigma_{Y|\beta}^{-1} D + \Sigma_\beta^{-1} & \text{precision matrix of } \beta \\
\hat{\beta} &= \left(T_\beta^{-1} D\right)\left(\Sigma_{Y|\beta}^{-1}\left(Y - \mu\right)\right) & \text{posterior mean of } \beta
\end{aligned}
\tag{30}
$$

C   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.

$$
\begin{aligned}
\tau^{\text{INV}} \mid Y_T, Y_C, \sigma_{\text{GP}}, \sigma_\epsilon, \ell &\sim \mathcal{N}\left(\mu_{\tau^{\text{INV}}|Y}, \Sigma_{\tau^{\text{INV}}|Y}\right) \\
\mu_{\tau^{\text{INV}}|Y} &\approx \left(\mathbf{1}^\mathsf{T} \Sigma_{b|Y}^{-1} \mu_{b|Y}\right) / \left(\mathbf{1}^\mathsf{T} \Sigma_{b|Y}^{-1} \mathbf{1}\right) \\
\mu_{b|T} &\equiv \text{cov}\left(g_T(b), Y_T\right) \text{cov}\left(Y_T\right)^{-1} Y_T \\
\mu_{b|C} &\equiv \text{cov}\left(g_T(b), Y_C\right) \text{cov}\left(Y_C\right)^{-1} Y_C \\
\mu_{\mathcal{B}|Y} &= \mu_{b|T} - \mu_{b|C} \\
\mu_{\tau^{\text{INV}}|Y} &= \left(\mathbf{1}^\mathsf{T} \Sigma_{b|Y}^{-1} \mu_{b|Y}\right) / \left(\mathbf{1}^\mathsf{T} \Sigma_{b|Y}^{-1} \mathbf{1}\right)
\end{aligned}
\tag{31}
$$

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

$$\text{cov}\left(\begin{pmatrix} Y_T \\ Y_C \end{pmatrix} \mid H_0\right) = \begin{bmatrix} \Sigma_{TT} & \Sigma_{TC} \\ \Sigma_{TC}^T & \Sigma_{CC} \end{bmatrix} \text{ where}$$

$$\Sigma_{TT} \equiv K_{TT} + \sigma_\epsilon^2 I_{n_T}$$
$$\Sigma_{CC} \equiv K_{CC} + \sigma_\epsilon^2 I_{n_C} \qquad (32)$$
$$\Sigma_{TC} \equiv K_{TC}$$

where the entries of K_{TT}, K_{CC} and K_{TC} are obtained simply be 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 $\mu_{b|T}$ and $\mu_{b|T}$ are obtained by left-multiplying Y_T and Y_C by matrices that are deterministic functions of the unit locations and the hyperparameters

$$A_T \equiv \text{cov}\left(g_T(\boldsymbol{b}), Y_T\right) \text{cov}\left(Y_T\right)^{-1} = K_{bT}\Sigma_{TT}^{-1}, \text{ and}$$
$$A_C \equiv \text{cov}\left(g_C(\boldsymbol{b}), Y_C\right) \text{cov}\left(Y_C\right)^{-1} = K_{bC}\Sigma_{CC}^{-1}. \qquad (33)$$

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

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

$$\text{cov}\left(\begin{pmatrix} A_T Y_T \\ A_C Y_C \end{pmatrix} \mid H_0\right) = \begin{bmatrix} A_T \Sigma_{TT} A_T^T & A_T \Sigma_{TC} A_C^T \\ (A_T \Sigma_{TC} A_C^T)^T & A_C \Sigma_{CC} A_C^T \end{bmatrix} \qquad (34)$$

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

$$\text{cov}\left(\mu_{\mathcal{B}|Y} \mid H_0\right) = \text{cov } A_T Y_T - A_C Y_C$$
$$= A_T \Sigma_{TT} A_T^T + A_C \Sigma_{CC} A_C^T - A_T \Sigma_{TC} A_C^T - (A_T \Sigma_{TC} A_C^T)^T \qquad (35)$$

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 $\boldsymbol{v}$, its variance is given by

$$\text{var}\left(\bar{\tau}^v \mid H_0\right) = \text{cov}\left(\frac{\boldsymbol{v}^T \mu_{\mathcal{B}|Y}}{\boldsymbol{1}_R^T \boldsymbol{v}}\right)$$
$$= \frac{\boldsymbol{v}^T \text{cov}\left(\mu_{\mathcal{B}|Y}\right) \boldsymbol{v}}{(\boldsymbol{1}_R^T \boldsymbol{v})^2}. \qquad (36)$$

From this null distribution the p-value follows:

$$\mathbb{P}\left(|\bar{\tau}^v| > |\bar{\tau}^v_{obs}| \mid H_0\right) = 2\Phi\left(-\frac{|\bar{\tau}^v_{obs}|}{\sqrt{\text{var}\left(\bar{\tau}^v \mid H_0\right)}}\right). \qquad (37)$$

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

D   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}
1	0	0.85 (0.08)	1.00	1.16 (0.05)	1.20	1.30 (0.05)	1.34	1.28 (0.05)	1.33	0.85 (0.08)	1.00	1.30 (0.05)	1.34
2	1	0.84 (0.08)	0.99	1.06 (0.04)	1.13	1.28 (0.05)	1.32	1.28 (0.05)	1.32	0.84 (0.08)	0.98	1.27 (0.05)	1.31
3	1	0.84 (0.08)	0.99	1.06 (0.04)	1.13	1.28 (0.05)	1.32	1.28 (0.05)	1.32	0.84 (0.08)	0.98	1.27 (0.05)	1.31
4	2	0.81 (0.07)	0.95	1.05 (0.04)	1.12	1.23 (0.05)	1.27	1.29 (0.05)	1.33	0.82 (0.07)	0.96	1.24 (0.05)	1.28
5	3	0.78 (0.07)	0.91	1.05 (0.04)	1.12	1.17 (0.05)	1.20	1.29 (0.05)	1.33	0.81 (0.07)	0.95	1.23 (0.05)	1.26
6	6	0.69 (0.07)	0.79	1.05 (0.04)	1.12	1.01 (0.05)	1.03	1.29 (0.05)	1.33	0.81 (0.07)	0.94	1.22 (0.05)	1.25
7	10	0.61 (0.06)	0.67	1.05 (0.04)	1.12	0.86 (0.05)	0.86	1.29 (0.05)	1.33	0.81 (0.07)	0.93	1.22 (0.05)	1.25
8	15	0.54 (0.07)	0.58	1.05 (0.04)	1.12	0.74 (0.06)	0.73	1.29 (0.05)	1.33	0.81 (0.07)	0.93	1.22 (0.05)	1.25
9	25	0.47 (0.07)	0.48	1.05 (0.04)	1.12	0.60 (0.06)	0.58	1.29 (0.05)	1.33	0.81 (0.07)	0.93	1.22 (0.05)	1.25
10	39	0.42 (0.07)	0.41	1.05 (0.04)	1.12	0.51 (0.07)	0.49	1.29 (0.05)	1.33	0.81 (0.07)	0.93	1.22 (0.05)	1.25
11	63	0.38 (0.07)	0.36	1.05 (0.04)	1.12	0.44 (0.07)	0.41	1.29 (0.05)	1.33	0.80 (0.07)	0.93	1.22 (0.05)	1.25
12	100	0.35 (0.08)	0.32	1.05 (0.04)	1.12	0.39 (0.08)	0.35	1.29 (0.05)	1.33	0.80 (0.07)	0.93	1.22 (0.05)	1.25
13	158	0.33 (0.08)	0.30	1.05 (0.04)	1.12	0.36 (0.08)	0.32	1.29 (0.05)	1.33	0.80 (0.07)	0.93	1.22 (0.05)	1.25
14	251	0.32 (0.08)	0.28	1.05 (0.04)	1.12	0.34 (0.08)	0.29	1.29 (0.05)	1.33	0.80 (0.07)	0.93	1.22 (0.05)	1.25
15	398	0.31 (0.08)	0.27	1.05 (0.04)	1.12	0.32 (0.08)	0.28	1.29 (0.05)	1.33	0.80 (0.07)	0.93	1.22 (0.05)	1.25
16	630	0.31 (0.08)	0.26	1.05 (0.04)	1.12	0.31 (0.08)	0.27	1.29 (0.05)	1.33	0.80 (0.07)	0.93	1.22 (0.05)	1.25
17	1000	0.30 (0.08)	0.26	1.05 (0.04)	1.12	0.31 (0.08)	0.26	1.29 (0.05)	1.33	0.80 (0.07)	0.93	1.22 (0.05)	1.25

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.

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]*.

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)*.

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

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.