# GeoRDD manuscript

Maxime Rischard

October 26, 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 boundary
- 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 **boundary 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 boundary, 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 boundary. 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 $\mathbf{s}_1 \in [-1, 1]$, and $\mathbf{s}_2 \in [-1, 1]$, 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 $\mathbf{s}_2$ (parallel to the border), given by $Y_i = \alpha \mathbf{s}_{2i} + \epsilon_i$, where $\epsilon_i$ 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 method 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 boundary, 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, or even a busy road, 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 geographical regression discontinuity designs 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 **boundary curve** $\mathcal{B}$; and
3. Subtract the two extrapolations to estimate the treatment effect along the boundary $\tau(\mathbf{s}) : \mathcal{B} \to \mathbb{R}$.

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.

# 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 $n_{\boldsymbol{b}}$ "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 $\mu$, linear trend with coefficients $\beta$, a spatial Gaussian process $f(\mathbf{s})$, a constant treatment effect $\tau$ in the treatment region, and iid normal noise $\epsilon$.

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

$f$ is a smooth surface covering all of $\mathcal{A}$, specificed as a Gaussian Process with squared exponential covariance kernel $k$ with lengthscale $\ell$ and variance $\sigma^2_{GP}$. 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 $\beta$, a spatial Gaussian process $f(\mathbf{s})$, and iid normal noise $\epsilon$. The Gaussian process has zero mean, and its covariance function is a modeling choice. There is a rich literature of possible covariance functions ("kernels" in the machine learning world) **[LukeB: do you have a good reference summarizing popular options?]**, but in this paper, we will use the squared exponential kernel, for its ease of understanding and its prevalence in applied spatial statistics.

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

Note that the treatment effect $\tau$ 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 $g_C$.

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

In this specification, the parameters $\ell$, $\sigma_{GP}$, and $\sigma_\epsilon$ 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 $\beta$. 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.

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

$$
\tau = \lim_{x \downarrow b} \mathbb{E}\left[y \mid X = s\right] - \lim_{x \uparrow b} \mathbb{E}\left[y \mid X = x\right] = \mathbb{E}\left[Y_T \mid X = b\right] - \mathbb{E}\left[Y_C \mid X = b\right] \tag{5}
$$

Analogously, we focus on the treatment effect at the boundary $\mathcal{B}$ between the treatment and control regions. Note that $\mathcal{B}$ is a one-dimensional manifold embedded in $\mathcal{A}$. We proceed by extrapolating both Gaussian processes to the boundary, and then taking the difference of the predictions to obtain the posterior treatment effect $\tau(\mathcal{B})$ along the boundary. Computationally, we need to represent this boundary as a set of $n_{\boldsymbol{b}}$ "sentinel" units distributed along the boundary $\boldsymbol{b} = \left\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{n_{\boldsymbol{b}}}\right\}$, $\boldsymbol{b}_i \in \mathcal{B}$. The extrapolation step then proceeds mechanically through multivariate-normal theory.

$$
\begin{aligned}
g_T(\boldsymbol{b}) \mid Y_T, S_T, \ell, \sigma_{GP}, \sigma_\epsilon &\sim \mathcal{N}\left(\mu_{\boldsymbol{b}|T}, \Sigma_{\boldsymbol{b}|T}\right) \\
\mu_{\boldsymbol{b}|T} &\equiv \operatorname{cov}\left(g_T(\boldsymbol{b}), Y_T\right) \operatorname{cov}\left(Y_T\right)^{-1} Y_T \\
\Sigma_{\boldsymbol{b}|T} &\equiv \operatorname{cov}\left(g_T(\boldsymbol{b})\right) - \operatorname{cov}\left(g_T(\boldsymbol{b}), Y_T\right) \operatorname{cov}\left(Y_T\right)^{-1} \operatorname{cov}\left(Y_T, g_T(\boldsymbol{b})\right)
\end{aligned}
\tag{6}
$$

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 boundary is also multivariate normal with posterior mean and covariance

$$
\begin{aligned}
\mu_{\boldsymbol{b}|Y} &= \mathbb{E}\left(\tau(\boldsymbol{b}) \mid Y_T, Y_C\right) = \mu_{\boldsymbol{b}|T} - \mu_{\boldsymbol{b}|C} \\
\Sigma_{\boldsymbol{b}|Y} &= \operatorname{cov}\left(\tau(\boldsymbol{b}) \mid Y_T, Y_C\right) = \Sigma_{\boldsymbol{b}|T} + \Sigma_{\boldsymbol{b}|C}\,.
\end{aligned}
\tag{7}
$$

### 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 $\mu$, $\beta$ and $\tau$. 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

$$\begin{aligned}
\tau &\sim \mathcal{N}\left(0, \sigma_\tau^2\right) \\
\mu &\sim \mathcal{N}\left(0, \sigma_\mu^2\right) \\
\beta &\sim \mathcal{N}\left(0, \sigma_\beta^2\right) \\
\operatorname{cov}(Y_i, \tau) &= \sigma_\tau^2 Z_i \\
\operatorname{cov}(Y_i, \mu) &= \sigma_\mu^2 \\
\operatorname{cov}(Y_i, \beta) &= \sigma_\beta^2 \mathbf{s}^\intercal \mathbf{s} \\
\operatorname{cov}(Y_i, Y_j) &= \sigma_\mu^2 + \sigma_\tau^2 Z_i Z_j + \sigma_\beta^2 \mathbf{s}_i^\intercal \mathbf{s}_j + k(\mathbf{s}_i, \mathbf{s}_j) + \delta_{ij}\sigma_\epsilon^2 \\
\operatorname{cov}(Y_i, f(\mathbf{s}_j)) &= \operatorname{cov}(f(\mathbf{s}_i), f(\mathbf{s}_j)) = k(\mathbf{s}_i, \mathbf{s}_j)
\end{aligned} \tag{8}$$

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

$$\tau \mid Y, \ell, \sigma_{GP} \sim \mathcal{N}\left(\operatorname{cov}(Y, \tau)^\intercal \operatorname{cov}(Y)^{-1} Y, \sigma_\tau^2 - \operatorname{cov}(Y, \tau)^\intercal \operatorname{cov}(Y)^{-1} \operatorname{cov}(Y, \tau)\right) \tag{9}$$

To proceed computationally, we define the treatment indicator vector $\mathbb{I}_T$ with $i$th entry equal to $Z_i$, the spatial covariate $n \times 2$ matrix $S$ with $i$th 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.

$$\begin{aligned}
\mathbb{E}\left(\tau \mid Y, \ell, \sigma_{GP}, \sigma_\epsilon\right) &= \sigma_\tau^2 \mathbb{I}_T^\intercal \left\{\sigma_\mu^2 + \sigma_\tau^2 \mathbb{I}_T \mathbb{I}_T^\intercal + \sigma_\beta^2 SS^\intercal + \mathbf{K} + \sigma_\epsilon^2 \mathbf{I}\right\}^{-1} Y \\
\operatorname{var}\left(\tau \mid Y, \ell, \sigma_{GP}, \sigma_\epsilon\right) &= \sigma_\tau^2 - \sigma_\tau^2 \mathbb{I}_T^\intercal \left\{\sigma_\mu^2 + \sigma_\tau^2 \mathbb{I}_T \mathbb{I}_T^\intercal + \sigma_\beta^2 SS^\intercal + \mathbf{K} + \sigma_\epsilon^2 \mathbf{I}\right\}^{-1} \mathbb{I}_T
\end{aligned} \tag{10}$$

What remains is the inference on the hyperparameters $\sigma_\epsilon, \sigma_{GP}$ and $\ell$. 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 $\tau$ be integrated over those parameters. The compromise is clear: the Bayesian approach incorporates the uncertainty in the hyperparameters, thus giving more reliable inference on $\tau$, 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

$$
\begin{aligned}
Y_{iT} &= \underbrace{m_T + \mathbf{d}_i^\intercal \gamma + \mathbf{s}_i^\intercal \beta_T + f_T(\mathbf{s}_i)}_{g_T(\mathbf{s}_i)} + \epsilon_i \\
Y_{iC} &= \underbrace{m_C + \mathbf{d}_i^\intercal \gamma + \mathbf{s}_i^\intercal \beta_C + f_C(\mathbf{s}_i)}_{g_C(\mathbf{s}_i)} + \epsilon_i \\
f_T, f_C &\overset{\perp\!\!\!\perp}{\sim} \mathcal{GP}\left(0, k(\mathbf{s}, \mathbf{s}')\right) \\
k(\mathbf{s}, \mathbf{s}') &= \sigma_{GP}^2 \exp\left(-\frac{(\mathbf{s}-\mathbf{s}')^\intercal (\mathbf{s}-\mathbf{s}')}{2\ell^2}\right) \\
\gamma_j &\overset{\perp\!\!\!\perp}{\sim} \mathcal{N}\left(0, \sigma_\gamma^2\right) \text{ for } j = 1, 2, \ldots, p
\end{aligned}
\tag{11}
$$

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 estimand class of weighted means of the functional treatment effect $\tau(\mathbf{s})$, with weight function $w(\mathbf{s})$ defined everywhere on the border $\mathcal{B}$. The weighted mean integral can be approximated as a sum at the sentinels $\boldsymbol{b}_{1:K}$.

$$
\begin{aligned}
\tau^w &= \frac{\oint_{\mathcal{B}} w(\mathbf{s})\tau(\mathbf{s})\, d\mathbf{s}}{\oint_{\mathcal{B}} w(\mathbf{s})\, d\mathbf{s}}, \\
&\approx \frac{\sum_{i=0}^{K} w(\boldsymbol{b}_i)\tau(\boldsymbol{b}_i)}{\sum_{i=0}^{K} w(\boldsymbol{b}_i)}.
\end{aligned}
\tag{12}
$$

We have shown the posterior distribution of $\tau(\mathbf{s})$ at the sentinels to be multivariate normal, with mean $\mu_{b|Y}$ and covariance $\Sigma_{b|Y}$ given in (7). Since $\tau^w$ is approximated as 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

$$
\begin{aligned}
\mu_{\tau^w|Y} &= \frac{w(\boldsymbol{b})^\intercal \mu_{b|Y}}{w(\boldsymbol{b})^\intercal \mathbf{1}_K} \\
\Sigma_{\tau^w|Y} &= \frac{w(\boldsymbol{b})^\intercal \Sigma_{b|Y} w(\boldsymbol{b})}{\left(w(\boldsymbol{b})^\intercal \mathbf{1}_K\right)^2}
\end{aligned}
\tag{13}
$$

where $w(\boldsymbol{b})$ is the vector of weights evaluated at the sentinels.

## 4.1 Uniform ATE

The question remains: what is the most sensible choice of weights? The simplest choice is uniform weights $w(\mathbf{s}) = 1$, a seemingly reasonable and unopinionated decision. We estimate $\tau^{UNIF}$, the uniformly weighted mean of $\tau(\mathbf{s})$ along the boundary, by averaging the entries of the mean posterior at the sentinels. Following (12) and (13):

$$
\begin{aligned}
\tau^{\text{UNIF}} &\equiv \frac{\oint_{\mathcal{B}} \tau(x)\, ds}{\oint_{\mathcal{B}} dx} \\
\tau^{\text{UNIF}} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell &\sim \mathcal{N}\left(\mu_{\tau^{\text{UNIF}}|Y}, \Sigma_{\tau^{\text{UNIF}}|Y}\right) \\
\mu_{\tau^{\text{UNIF}}|Y} &= \left(\mathbf{1}^\intercal\, \mu_{b|Y}\right)/n_b \\
\Sigma_{\tau^{\text{UNIF}}|Y} &= \left(\mathbf{1}^\intercal\, \Sigma_{b|Y}\, \mathbf{1}\right)/n_b^2
\end{aligned}
\tag{14}
$$

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 weighting, 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 $\tau^{\text{UNIF}}$, potentially jeopardizing the successful detection of otherwise strong treatment effects: $\tau^{\text{UNIF}}$ 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 $w(\mathbf{s}) = \rho(\mathbf{s})$, where $\rho$ is the local population density. The resulting estimand $\tau^\rho$ 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)

## 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 $\tau^{\text{UNIF}}$ 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 $\tau^{\text{INV}}$

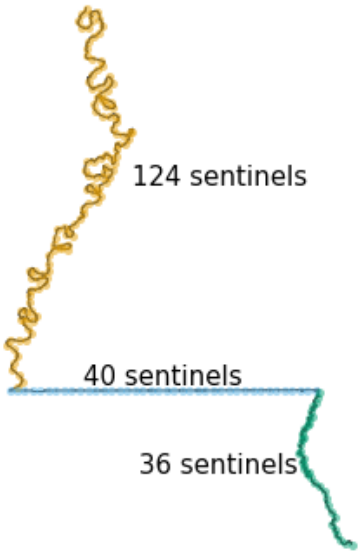


Figure 1: Evenly spaced sentinels along the border between Mississippi and Louisiana.

$$\begin{aligned}
\tau^{\text{INV}} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell &\sim \mathcal{N}\left(\mu_{\tau^{\text{INV}}|Y}, \Sigma_{\tau^{\text{INV}}|Y}\right), \\
\mu_{\tau^{\text{INV}}|Y} &= \left(\mathbf{1}^\intercal \Sigma_{b|Y}^{-1} \mu_{b|Y}\right) / \left(\mathbf{1}^\intercal \Sigma_{b|Y}^{-1} \mathbf{1}\right), \\
\Sigma_{\tau^{\text{INV}}|Y} &= 1 / \left(\mathbf{1}^\intercal \Sigma_{b|Y}^{-1} \cdot \mathbf{1}\right)
\end{aligned} \tag{15}$$

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 (12). 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 $w(\boldsymbol{b}) = \Sigma_{b|Y}^{-1}\mathbf{1}$. This can put negative weights on some sentinels, 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. It is not unprecented hower (Li et al., 2016).

## 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 $\tau^{\text{PROJ}}$ is then the uniformly weighted mean applied with the projected sentinels instead of the evenly spaced sentinels. We can therefore modify (14), 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 $\tau^{\text{PROJ}}$:

$$\begin{aligned}
\tau^{\text{PROJ}} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell &\sim \mathcal{N}\left(\mu_{\tau^{\text{PROJ}}|Y}, \Sigma_{\tau^{\text{PROJ}}|Y}\right), \\
\mu_{\tau^{\text{PROJ}}|Y} &= \frac{1}{n_C + n_T} \sum_{i=1}^{n_T+n_C} \mathbb{E}\left[\tau\left(\text{proj}_{\mathcal{B}}(\mathbf{s}_i)\right) \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell\right], \\
\Sigma_{\tau^{\text{PROJ}}|Y} &= \frac{1}{(n_C + n_T)^2} \sum_{i=1}^{n_T+n_C} \sum_{j=1}^{n_T+n_C} \text{cov}\left[\tau\left(\text{proj}_{\mathcal{B}}(\mathbf{s}_i)\right), \tau\left(\text{proj}_{\mathcal{B}}(\mathbf{s}_j)\right) \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell\right].
\end{aligned} \tag{16}$$

The posterior expectations and covariances in (16) can be obtained as in (7), but using the projected sentinels. Note that $\tau^{\text{PROJ}}$ is still within the class of weighted mean estimands (12), with weight function $w(\mathbf{s}) = \sum_{i=1}^{n_T+n_C} \delta\left(\mathbf{s} - \text{proj}_{\mathcal{B}}(\mathbf{s}_i)\right)$, where $\delta$ 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 $\tau^{\text{UNIF}}$ 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 $\tau^{\text{PROJ}}$ and apply it to an infinite population of uniform density on both sides of the border. This yields the geography-based projected land ATE $\tau^{\text{GEO}}$. To estimate $\tau^{\text{GEO}}$, a tight grid of evenly spaced points is first generated within $\mathcal{A}^T$ and $\mathcal{A}^C$. Each point on this grid is then projected onto the border 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 $\tau^{\text{GEO}}$. In other words, $\tau^{\text{GEO}}$ is estimated by applying the $\tau^{\text{UNIF}}$ estimator with sentinels obtained by projecting the grid points, instead of equispaced sentinels. $\tau^{\text{GEO}}$ remains in the category of weighted-mean estimands, with the weight function $w(\mathbf{s})$ in (12) proportional to the area of $\mathcal{A}^T$ and $\mathcal{A}^C$ that $\mathbf{s}$ is nearest to:

$$w(\mathbf{s}) = \int_{\mathcal{A}} \mathbb{I}\left\{\mathbf{s} = \text{proj}_{\mathcal{B}}(\mathbf{s}')\right\} d\mathbf{s}' \tag{17}$$

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

### 4.7 Wiggly Border Simulation

We illustrate the differences between the four average treatment effect estimators with a simulation study. 1000 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 the table below. 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 $\tau(S_1, S_2) = S_1$.

| | 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 $\rho = 0.1$ | high $\rho = 2.8$ |
| $\boldsymbol{\tau}$ | weak | medium | strong |

We fit Gaussian processes on each side of the boundary, 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), (b), and (c), and the resulting estimand and estimator behavior as the number of wiggles increases is shown in Figure 2(d).

When the border is a straight line and $\mathcal{A}^{\text{T}}$ and $\mathcal{A}^{\text{C}}$ are rectangles, and because the treatment effect does not depend on the vertical axis $S_2$, the density-weighted estimand $\tau^{\rho}$ equals the projected superpopulation estimand $\tau^{\text{POP}}$, and they are in fact both equal to the infinite-population average treatment effect $\mathbb{E}\left[Y(1) - Y(0)\right]$. Correspondingly, the posteriors of $\tau^{\rho}$ and $\tau^{\text{POP}}$ are identical. With 1000 units, the finite-population and the infinite-population projected estimands are similar, though the difference can be more pronounced with smaller sample populations.

The geometry- and geography-based ATE $\tau^{\text{UNIF}}$ and $\tau^{\text{GEO}}$ 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 $\tau^{\rho}$ and $\tau^{\text{POP}}$.

Lastly, the information-based inverse-variance estimand $\tau^{\text{INV}}$ does not coincide with any others. The estimand here is between the population-based and the geography-based estimators, but that need not be true in the general case. However, $\tau^{\text{INV}}$ is guaranteed to have the lowest posterior variance within the class of ATEs under consideration, and here we indeed see that it has the narrowest band.

As we introduce wiggles into the leftmost band, $\tau^{\rho}$ and $\tau^{\text{UNIF}}$ 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, $\tau^{\text{INV}}$ 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 $\tau^{\text{PROJ}}$, $\tau^{\text{GEO}}$, and $\tau^{\text{POP}}$ 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

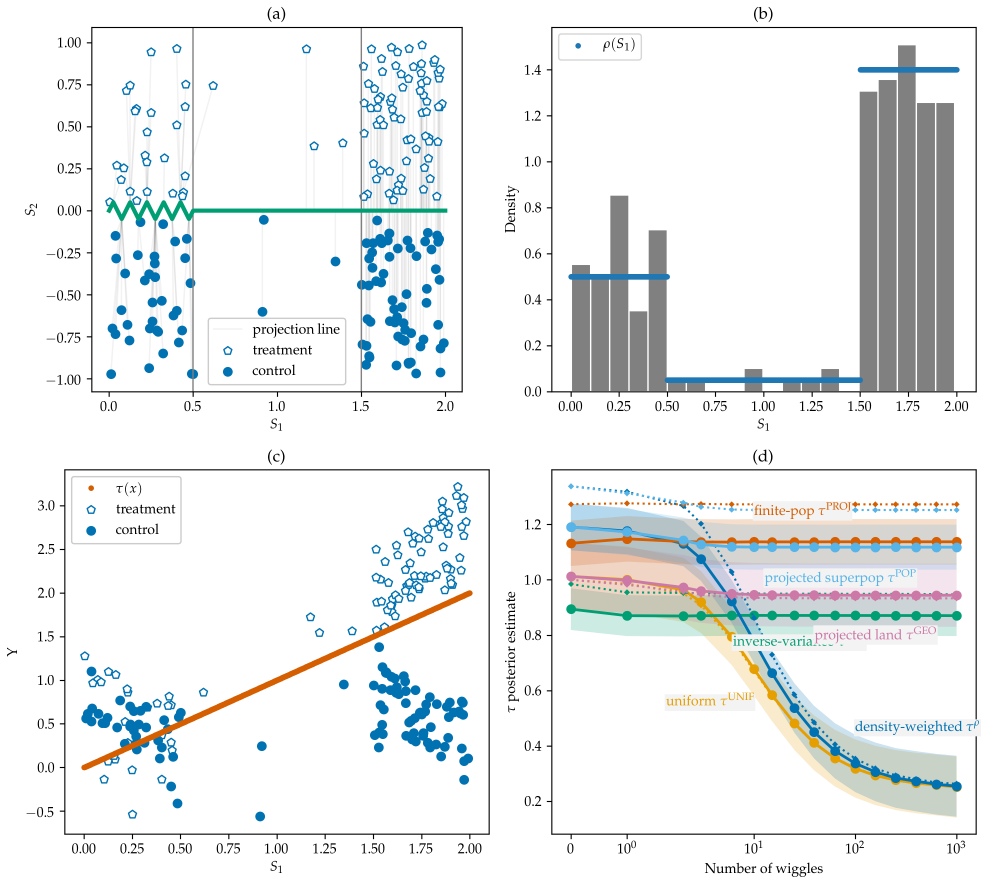


Figure 2: (a) Spatial positions of units and border; (b) Density of units along $S_1$; (c) Simulated outcomes and true treatment effect; (d) Estimator (posterior mean $\pm 1\sigma$) and estimand (dotted lines) behavior as left border gets wigglier.

observed units.

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 $\tau^{POP}$ 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

| Symbol | Description | Border Topology | Sentinels | Principle | Variance |
|---|---|---|---|---|---|
| $\tau^{UNIF}$ | Uniform ATE | Sensitive | Equispaced | Geometry-based | High |
| $\tau^{\rho}$ | Density-weighted ATE | Sensitive | Equispaced | Population-based | Low |
| $\tau^{INV}$ | Inverse-variance weighted ATE | Robust | Equispaced | Information-based | Lowest |
| $\tau^{PROJ}$ | Projected finite population ATE | Robust | Projected Units | Finite-population | Low |
| $\tau^{GEO}$ | Projected land ATE | Robust | Projected Grid | Geography-based | High |
| $\tau^{POP}$ | 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 boundary. 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 boundary: $\tau(\mathcal{B}) = 0$, 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 $\tau^{INV}$ has posterior given by (15). While the posterior is a Bayesian object, we can use it heuristically to derive a pseudo-p-value

$$\begin{aligned} Z_0 &\sim \mathcal{N}\left(0, \Sigma_{\tau^{INV}|Y}\right) \\ p^{INV} &= \mathbb{P}\left(|Z_0| > \left|\mu_{\tau^{INV}|Y}\right|\right) \\ &= 2\Phi\left(-\frac{\left|\mu_{\tau^{INV}|Y}\right|}{\sqrt{\Sigma_{\tau^{INV}|Y}}}\right) \end{aligned} \tag{18}$$

While we didn't derive this pseudo-p-value through a rigorous procedure, our simulations show that it actually has good frequentist properties.

## 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 boundary, 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

$$t = \log \mathbb{P}\left(Y_T, Y_C \mid \mathcal{M}_1\right) - \log \mathbb{P}\left(Y_T, Y_C \mid \mathcal{M}_0\right) \tag{19}$$

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

$$t^* = \log \mathbb{P}\left({Y_T}^*, {Y_C}^* \mid \mathcal{M}_1\right) - \log \mathbb{P}\left({Y_T}^*, {Y_C}^* \mid \mathcal{M}_0\right) \tag{20}$$

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.

$$p^{lik} = \mathbb{P}\left(t^* > t_{obs} \mid \mathcal{M}_0\right) \tag{21}$$

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 $\chi^2$ 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 boundary, without any explicit reference to the boundary $\mathcal{B}$. The test statistic and p-values can be computed without access to the sentinel positions, using only the treatment and control indicators.

To address this oddity, we can derive a test statistic directly from the posterior treatment effect along the boundary, approximated in (7) by its mean vector $\boldsymbol{\mu}_{b|Y}$ and covariance matrix $\Sigma_{b|Y}$ at the sentinel positions $\boldsymbol{b}$. We will use $\boldsymbol{\mu}$ and $\Sigma$ as shorthand throughout this section. If a $k$-vector $\mathbf{y}$ has multivariate distribution $\mathcal{N}0, \Sigma$, then $\mathbf{y}^\intercal \Sigma^{-1} \mathbf{y}$ has distribution $\chi^2_k$. This suggests that we could use $S = \boldsymbol{\mu}^\intercal \Sigma^{-1} \boldsymbol{\mu}$ as a test statistic, and obtain a p-value by comparing it to a $\chi^2_k$ distribution, where $k$ is the number of sentinels. However, we face two problems. Firstly, this test derived heuristically from a Bayesian posterior is invalid from a frequentist perspective. Secondly, while $\Sigma$ is mathematically full-rank, it is typically numerically rank-deficient. Therefore, $k$ overestimates the true degrees of freedom of $\Sigma$, which invalidates the test.

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

We therefore again take the parametric bootstrap approach, this time using $S$ as the test statistic instead of the likelihood ratio. Because $S$ involves inverting a matrix $\Sigma$ that is mathematically of full rank, but numerically of low rank, we may worry about the numerical stability of computing $S$. We rely on julia's matrix division polyalgorithm to ensure numerical stability, and check in simulated examples that adding a small constant to the diagonal of $\Sigma$ does not greatly affect the computed $S$. Furthermore, even if numerical stability was an issue, the parametric bootstrap ensures the frequentist validity of the test, though its power could be lowered.

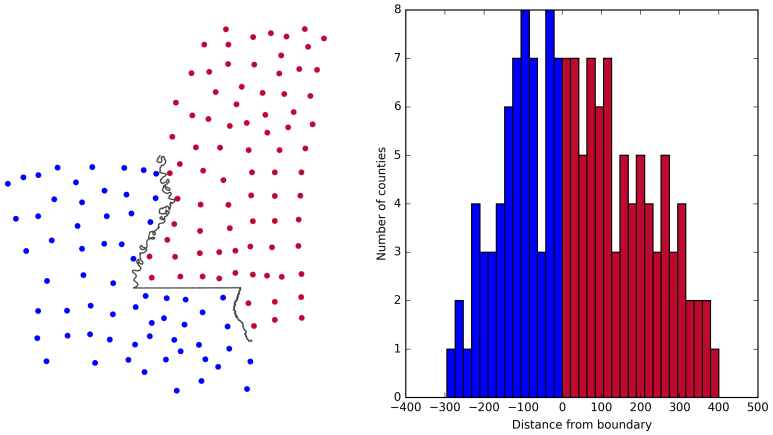


Figure 3: Position of units in an imaginary experiment in Louisiana and Mississippi

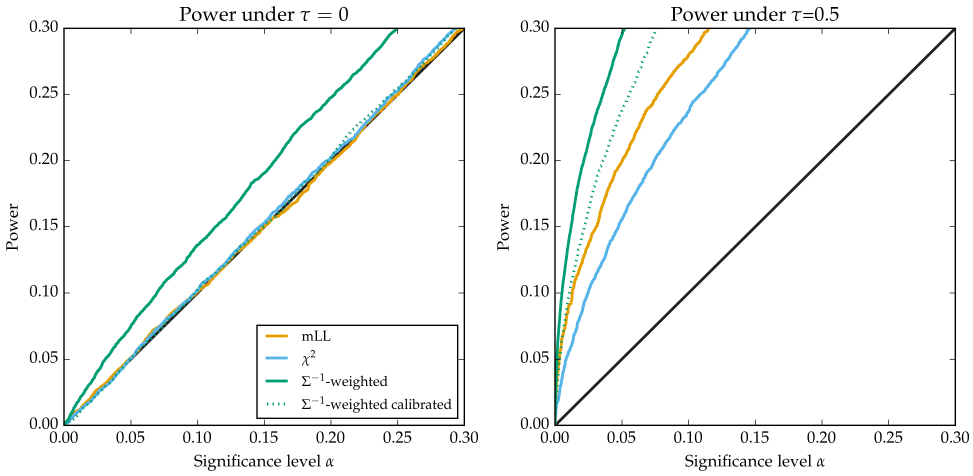


Figure 4: Power of hypothesis tests

## 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 boundary between Louisiana and Mississippi, we imagine an experiment where the unit of analysis is the county, located at its centroid, as shown in Figure 3(a). 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 $\ell = 50\,\text{km}$. We then add a constant treatment effect $\tau$ to all the outcomes in Louisiana. The results for $\tau = 0$ (null hypothesis) and $\tau = 0.5$ are shown in Figure 4.

We see that under the null, the p-values of the $\chi^2$ and likelihood ratio tests are uniformly distributed. 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 for example if we set $\alpha = 0.05$, we will falsely reject the null 7.5% 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. It can be calibrated using the same parametric bootstrap approach that was used for the likelihood and $\chi^2$ tests. The calibration can also be achieved analytically, since $\mu_{\tau^{\text{INV}}|Y}$ is normally distributed

under the null hypothesis.

Even after the calibration, the hypothesis test based on the inverse-variance mean has the highest power to detect the constant treatment effect. This can lead to a paradox: we may reject the weak null hypothesis, but fail to reject the sharp null hypothesis (using the $\chi^2$ 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 scientific contexts 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 boundary, creating new treatment and control groups. Because the boundary was chosen arbitrarily by us, we should not expect there to be a discontinuous jump in outcomes at this boundary. 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 $1^\circ, 2^\circ, 3^\circ, \ldots, 180^\circ$. 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

Classical regression discontinuity designs often suffer from low power, requiring many units near the boundary 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 boundary 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 boundary. 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 boundary, 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 boundary.

We illustrate this statement once more with the Louisiana-Mississippi example. The variance of the inverse-variance weighted treatment effect $\tau^{\text{INV}}$ is thence only a function of the positions of the units, available analytically by plugging the posterior variance (7) into the inverse-variance estimator (15). Following this procedure, we obtain a posterior standard deviation of the average treatment effect of 0.31. We then create a one-dimensional regression discontinuity design for the same setting, by using each unit's distance from the boundary as the covariate $x$, the distribution of which is shown in Figure 3(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 boundary (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 boundary, 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 boundary. 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 $\epsilon_i$, 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.

## 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]:

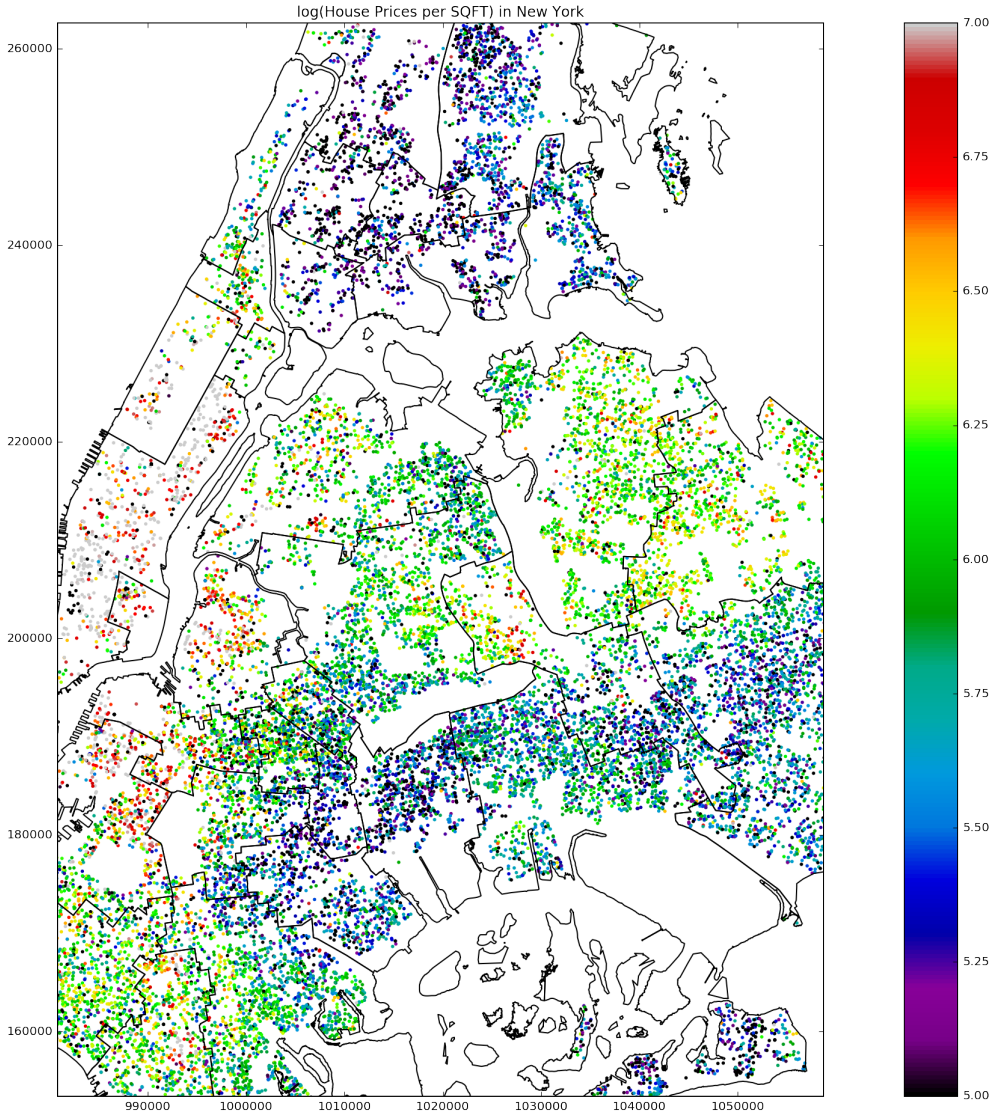


Figure 5: sales map

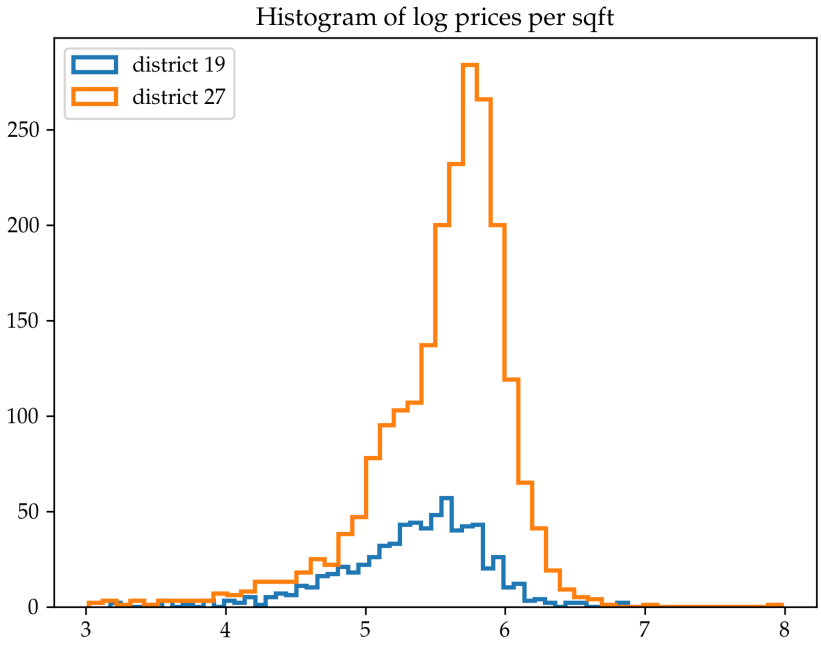


Figure 6:

$$
\begin{aligned}
Y_i = \log\left(\frac{\texttt{SalePrice}_i}{\texttt{SQFT}_i}\right) = {} & \mu_{\texttt{District[i]}} + \beta_{1,\texttt{TaxClass[i]}} + \beta_{2,\texttt{BuildingClass[i]}} \\
& + f_{\texttt{District[i]}}(\mathbf{s}_i) + \epsilon_i \\
\epsilon_i \sim \mathcal{N}\left(0, \sigma_y^2\right) & \\
\mu_j \sim \mathcal{N}\left(\bar{Y}_j, \sigma_\mu^2\right) & \\
\beta_{1j}, \beta_{2j} \sim \mathcal{N}\left(0, \sigma_\beta^2\right) & \\
f_j(\mathbf{s}_i) \sim \mathcal{GP}\left(0, k(\mathbf{s}, \mathbf{s}')\right) & \\
k(\mathbf{s}, \mathbf{s}') = \sigma_{GP}^2 \exp\left\{-\frac{(\mathbf{s}-\mathbf{s}')^\intercal(\mathbf{s}-\mathbf{s}')}{2\ell^2}\right\} &
\end{aligned}
\tag{22}
$$

A visual inspection of the house sales map above suggests examining the boundary between districts 19 and 27. Importantly, the boundary between the two districts is also part of the boundary 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 boundary, and not merely an underlying trend that spans both districts.

## 7.4 parameter optimization

We initially fit the hyperparameters $\sigma_\beta$, $\sigma_{GP}$, $\ell$ and $\sigma_\epsilon$ by optimizing the marginal log-likelihood of the data within a single district. We choose district 27 as it contains more sales. We hold $\sigma_\mu$ fixed to 10 to give the

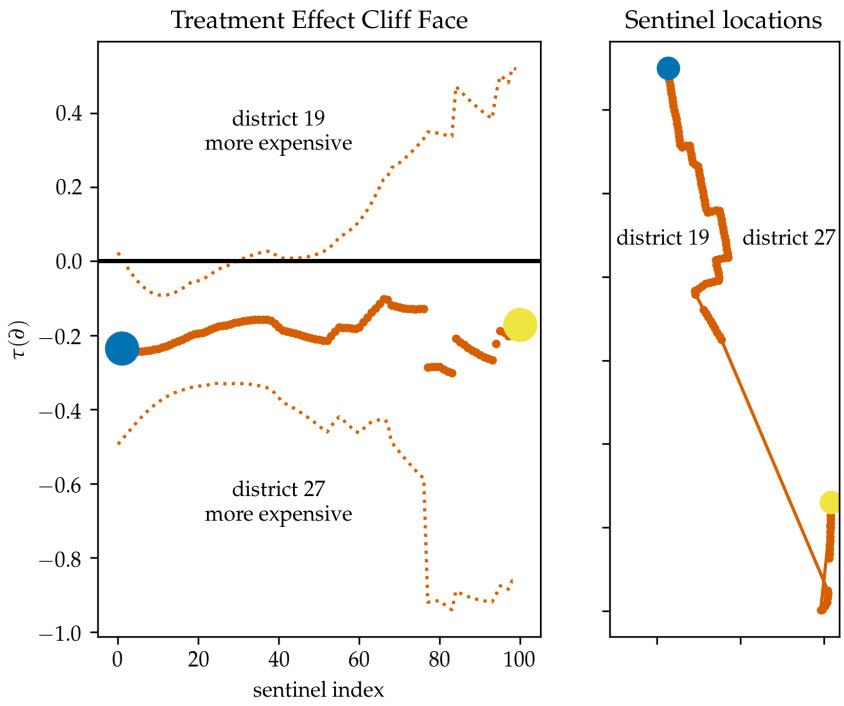


Figure 7: NYC cliff face

district means $\mu_j$ 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 $2n_{\boldsymbol{b}}$-dimensional multivariate normal distribution. Instead, we obtain the posterior means of the $\beta_{1j}$ and $\beta_{2j}$ 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 $\beta$ is low, and therefore the two approaches yield very similar results, but conditioning on the estimate of $\beta$ 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 XX, 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.

The treatment effect can also be visualized directly in Figure XX 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

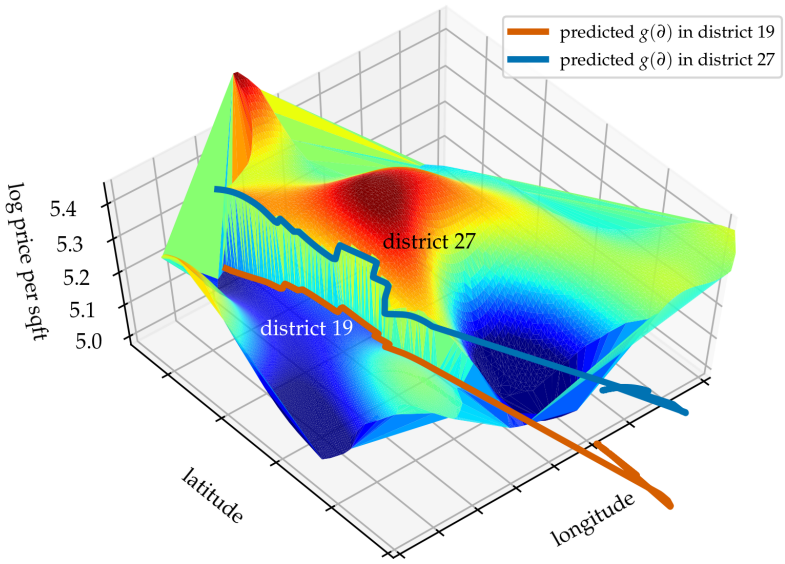


Figure 8: NYC surface plot

each equispaced sentinel, which has posterior distribution

$$\begin{aligned} \tau^{\text{UNIF}} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \hat{\beta}, \ell \sim \mathcal{N}\left(-0.20, 0.09^2\right) \text{ and tail probability} \\ \mathbb{P}\left(\tau^{\text{UNIF}} > 0 \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \hat{\beta}, \ell\right) = 1.48\%\,. \end{aligned} \tag{23}$$

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

$$\begin{aligned} \tau^{\text{INV}} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \hat{\beta}, \ell \sim \mathcal{N}\left(-0.20, 0.05^2\right) \text{ and reduced tail probability} \\ \mathbb{P}\left(\tau^{\text{INV}} > 0 \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \hat{\beta}, \ell\right) = 0.01\%\,. \end{aligned} \tag{24}$$

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

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

| Estimand | Posterior Mean | Posterior Standard Deviation | Posterior Tail Prob |
|---|---|---|---|
| $\tau^{\text{UNIF}}$ | -0.20 | 0.09 | 1.48% |
| $\tau^{\text{INV}}$ | -0.20 | 0.05 | 0.01% |
| $\tau^{\text{PROJ}}$ | -0.21 | 0.08 | 0.46% |

| Estimand | Posterior Mean | Posterior Standard Deviation | Posterior Tail Prob |
|---|---|---|---|
| $\tau^{GEO}$ | -0.19 | 0.10 | 2.54% |
| $\tau^{POP}$ | -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 XX. 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 XX.

| Test | p-value |
|---|---|
| $\chi^2$ bootstrap | 0.145 |
| mLL bootstrap | 0.0015 |
| $\tau^{INV}$ uncalibrated | 0.0003 |
| $\tau^{INV}$ calibrated | 0.0005 |

The three tests tell very different stories. The $\chi^2$ 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 $\chi^2$ 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 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 confirms our earlier concern that the marginal log-likelihood may be sensitive to features of the data other than the discontinuity at the boundary. 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 $\chi^2$ test shows more robustness, with 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 $\chi^2$ 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. Its high power and robustness to misspecification make a strong argument for its use in most applications.

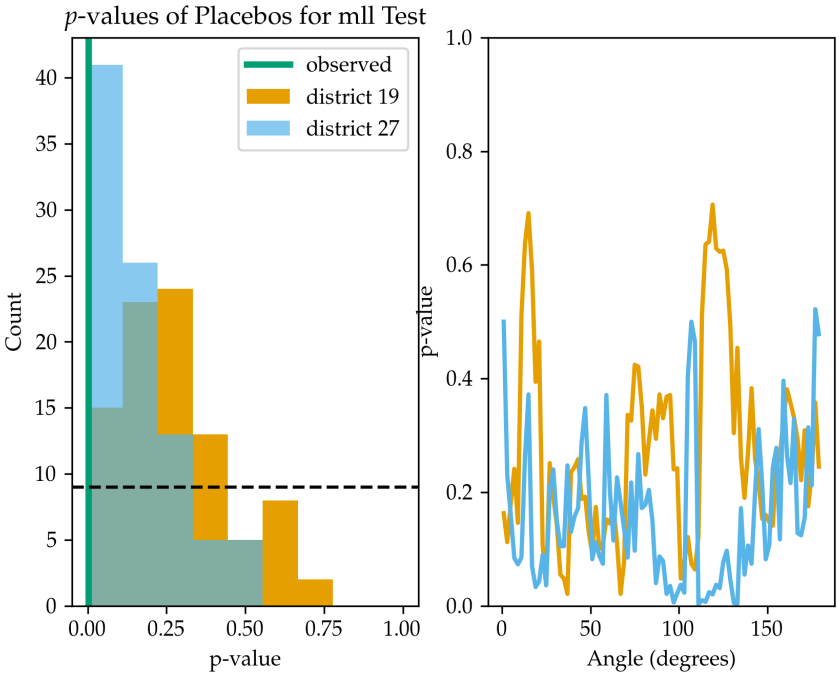


Figure 9: Placebo test for mLL test

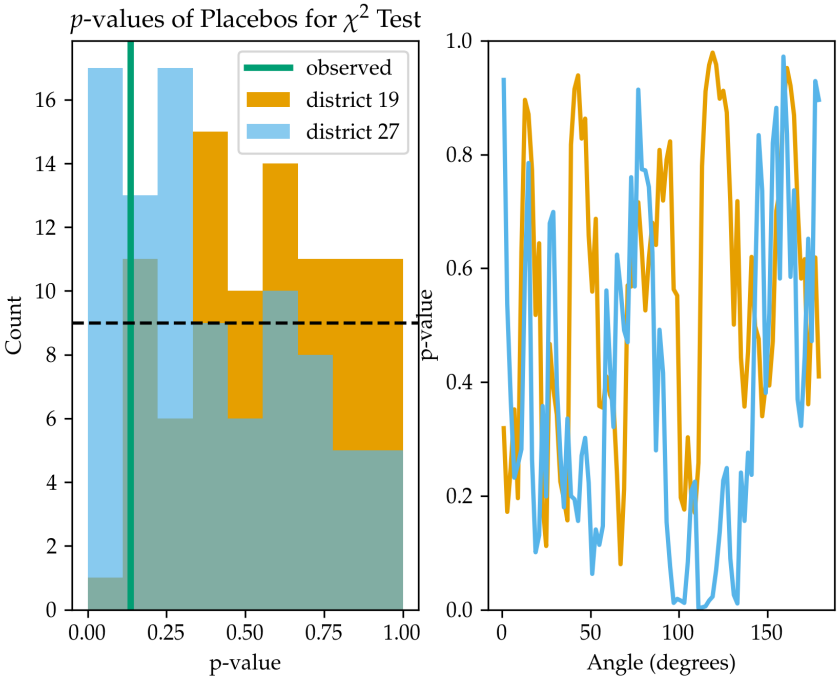


Figure 10: Placebo test for $\chi^2$ test

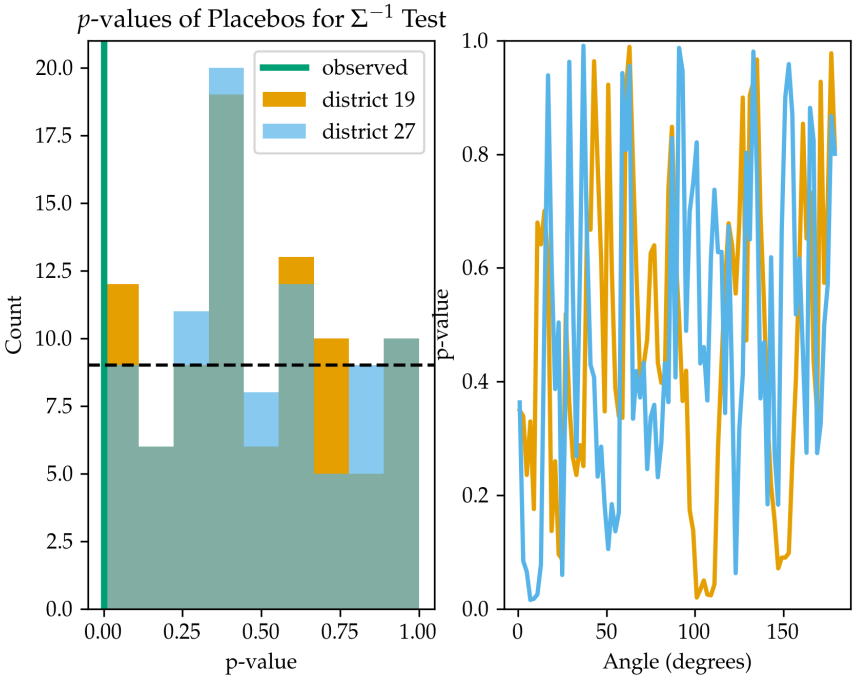


Figure 11: Placebo test for $\Sigma^{-1}$ test

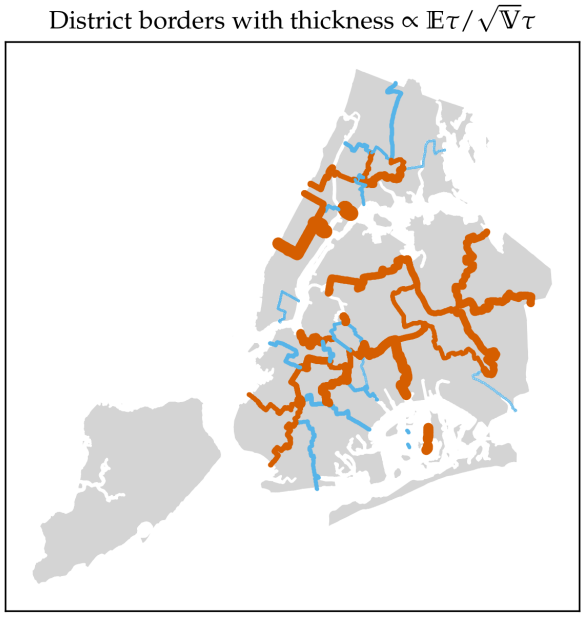


Figure 12: Pairwise effect size between adjacent districts.

### 7.8 pairwise treatment effect (all districts)

# 8 Conclusion

# A Covariances in 2GP model

**[needs a bit of love]**

$$
\begin{aligned}
m_T, m_C &\sim \mathcal{N}\left(0, \sigma_\mu^2\right) \\
\beta &\sim \mathcal{N}\left(0, \sigma_\beta^2\right) \\
\operatorname{cov}(Y_{iT}, m_T) &= \sigma_\mu^2 \\
\operatorname{cov}(Y_{iC}, m_C) &= \sigma_\mu^2 \\
\operatorname{cov}(Y_{iT}, \beta) &= \sigma_\beta^2 \mathbf{s}_i^\intercal \mathbf{s}_i \\
\operatorname{cov}(Y_{iC}, \beta) &= \sigma_\beta^2 \mathbf{s}_i^\intercal \mathbf{s}_i \\
\operatorname{cov}\left(Y_{iT}, f_T(\mathbf{s}')\right) &= k(\mathbf{s}_i, \mathbf{s}') \\
\operatorname{cov}\left(Y_{iC}, f_C(\mathbf{s}')\right) &= k(\mathbf{s}_i, \mathbf{s}') \\
\operatorname{cov}\left(Y_{iT}, f_C(\mathbf{s}')\right) &= 0 \\
\operatorname{cov}\left(Y_{iC}, f_T(\mathbf{s}')\right) &= 0 \\
\operatorname{cov}(Y_{iT}, Y_{jT}) &= \sigma_\mu^2 + \sigma_\beta^2 \mathbf{s}_i^\intercal \mathbf{s}_j + k(\mathbf{s}_i, \mathbf{s}_j) + \delta_{ij}\sigma_\epsilon^2 \\
\operatorname{cov}(Y_{iT}, Y_{jC}) &= \sigma_\beta^2 \mathbf{s}_i^\intercal \mathbf{s}_j
\end{aligned}
\tag{25}
$$

# 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?]**

$$
\begin{aligned}
\Sigma_{Y|\beta} &\equiv \operatorname{cov}\left(Y \mid \beta\right) && \text{conditional variance of } Y \\
\operatorname{cov}\left(Y_i, Y_j \mid \beta\right) &= \sigma_\epsilon^2 \delta_{ij} + k\left(\mathbf{s}_i, \mathbf{s}_j\right) \delta_{\texttt{District[i],District[j]}} && \text{(block diagonal)} \\
\Sigma_\beta &\equiv \operatorname{cov}\left(\beta\right) = \sigma_\beta^2 I_p && \text{prior variance of } \beta \\
\Sigma_Y &\equiv \operatorname{cov}\left(Y\right) = \Sigma_{Y|\beta} + D^\intercal \Sigma_\beta D && \text{unconditional variance of } Y \\
T_\beta &= D^\intercal \Sigma_{Y|\beta}^{-1} D + \Sigma_\beta^{-1} && \text{precision matrix of } \beta \\
\hat{\beta} &= \left(T_\beta^{-1} D\right)\left(\Sigma_{Y|\beta}^{-1}\left(Y - \mu\right)\right) && \text{posterior mean of } \beta
\end{aligned}
\tag{26}
$$

# 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.

$$
\begin{aligned}
\tau^{\text{INV}} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell &\sim \mathcal{N}\left(\mu_{\tau^{\text{INV}}|Y}, \Sigma_{\tau^{\text{INV}}|Y}\right) \\
\mu_{\tau^{\text{INV}}|Y} &\approx \left(\mathbf{1}^\intercal \Sigma_{b|Y}^{-1} \mu_{b|Y}\right) / \left(\mathbf{1}^\intercal \Sigma_{b|Y}^{-1} \mathbf{1}\right) \\
\mu_{b|T} &\equiv \operatorname{cov}\left(g_T(b), Y_T\right) \operatorname{cov}\left(Y_T\right)^{-1} Y_T \\
\mu_{b|C} &\equiv \operatorname{cov}\left(g_T(b), Y_C\right) \operatorname{cov}\left(Y_C\right)^{-1} Y_C \\
\mu_{\mathcal{B}|Y} &= \mu_{b|T} - \mu_{b|C} \\
\mu_{\tau^{\text{INV}}|Y} &= \left(\mathbf{1}^\intercal \Sigma_{b|Y}^{-1} \mu_{b|Y}\right) / \left(\mathbf{1}^\intercal \Sigma_{b|Y}^{-1} \mathbf{1}\right)
\end{aligned}
\tag{27}
$$

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

$$
\begin{aligned}
\operatorname{cov}\left(\begin{pmatrix} Y_T \\ Y_C \end{pmatrix} \mid H_0\right) &= \begin{bmatrix} \Sigma_{TT} & \Sigma_{TC} \\ \Sigma_{TC}^\intercal & \Sigma_{CC} \end{bmatrix} \text{ where} \\
\Sigma_{TT} &\equiv K_{TT} + \sigma_\epsilon^2 I_{n_T} \\
\Sigma_{CC} &\equiv K_{CC} + \sigma_\epsilon^2 I_{n_C} \\
\Sigma_{TC} &\equiv K_{TC}
\end{aligned}
\tag{28}
$$

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

$$
\begin{aligned}
A_T &\equiv \operatorname{cov}\left(g_T(b), Y_T\right) \operatorname{cov}\left(Y_T\right)^{-1} = K_{bT} \Sigma_{TT}^{-1}, \text{ and} \\
A_C &\equiv \operatorname{cov}\left(g_C(b), Y_C\right) \operatorname{cov}\left(Y_C\right)^{-1} = K_{bC} \Sigma_{CC}^{-1}.
\end{aligned}
\tag{29}
$$

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

$$
\operatorname{cov}\left(\begin{pmatrix} A_T Y_T \\ A_C Y_C \end{pmatrix} \mid H_0\right) = \begin{bmatrix} A_T \Sigma_{TT} A_T^\intercal & A_T \Sigma_{TC} A_C^\intercal \\ (A_T \Sigma_{TC} A_C^\intercal)^\intercal & A_C \Sigma_{CC} A_C^\intercal \end{bmatrix}
\tag{30}
$$

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

$$
\begin{aligned}
\operatorname{cov}\left(\mu_{\mathcal{B}|Y} \mid H_0\right) &= \operatorname{cov} A_T Y_T - A_C Y_C \\
&= A_T \Sigma_{TT} A_T^\intercal + A_C \Sigma_{CC} A_C^\intercal - A_T \Sigma_{TC} A_C^\intercal - (A_T \Sigma_{TC} A_C^\intercal)^\intercal
\end{aligned}
\tag{31}
$$

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

$$
\begin{aligned}
\operatorname{var}\left(\bar{\tau}^{\boldsymbol{\nu}} \mid H_0\right) &= \operatorname{cov}\left(\frac{\boldsymbol{\nu}^\intercal \mu_{\mathcal{B}|Y}}{\mathbf{1}_{n_b}^\intercal \boldsymbol{\nu}}\right) \\
&= \frac{\boldsymbol{\nu}^\intercal \operatorname{cov}\left(\mu_{\mathcal{B}|Y}\right) \boldsymbol{\nu}}{\left(\mathbf{1}_{n_b}^\intercal \boldsymbol{\nu}\right)^2}.
\end{aligned}
\tag{32}
$$

From this null distribution the p-value follows:

$$\mathbb{P}\left(\left|\bar{\tau}^{\boldsymbol{\nu}}\right| > \left|\bar{\tau}^{\boldsymbol{\nu}}_{\mathrm{obs}}\right| \mid \mathrm{H}_0\right) = 2\Phi\left(-\frac{\left|\bar{\tau}^{\boldsymbol{\nu}}_{\mathrm{obs}}\right|}{\sqrt{\mathrm{var}\left(\bar{\tau}^{\boldsymbol{\nu}} \mid \mathrm{H}_0\right)}}\right). \tag{33}$$

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

# D Wiggly boundary simulation results

| | $n_{\text{wiggles}}$ | $\widehat{\tau^{\text{UNIF}}}$ | $\tau^{\text{UNIF}}$ | $\widehat{\tau^{\text{INV}}}$ | $\tau^{\text{INV}}$ | $\widehat{\tau^{\rho}}$ | $\tau^{\rho}$ | $\widehat{\tau^{\text{PROJ}}}$ | $\tau^{\text{PROJ}}$ | $\widehat{\tau^{\text{GEO}}}$ | $\tau^{\text{GEO}}$ | $\widehat{\tau^{\text{POP}}}$ | $\tau^{\text{POP}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 5: Table of posterior mean, posterior standard deviation and true value for each average treatment effect estimand as the wiggliness of the boundary 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 T. Zajonc, 2011: Regression discontinuity design with multiple forcing variables. *Report, Harvard University.[972]*.

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.