# A Bayesian Nonparametric Approach to Geographic Regression Discontinuity Designs: Do School Districts Affect NYC House Prices?

Maxime Rischard * [a], Zach Branson[a], Luke Miratrix[b], and Luke Bornn[c]

[a]Department of Statistics, Harvard University
[b]Graduate School of Education, Harvard University
[c]Simon Fraser University

July 10, 2018

**Abstract**

Most research on regression discontinuity designs (RDDs) has focused on univariate cases, where only those units with a "forcing" variable on one side of a threshold value receive a treatment. Geographical regression discontinuity designs (GeoRDDs) extend the RDD to multivariate settings with spatial forcing variables. We propose a framework for analysing GeoRDDs, which we implement using Gaussian process regression. This yields a Bayesian posterior distribution of the treatment effect at every point along the border. We address nuances of having a functional estimand defined on a border with potentially intricate topology, particularly when defining and estimating causal estimands of the local average treatment effect (LATE). The Bayesian estimate of the LATE can also be used as a test statistic in a hypothesis test with good frequentist properties, which we validate using simulations and placebo tests. We demonstrate our methodology with a dataset of property sales in New York City, to assess whether there is a discontinuity in housing prices at the border between two school district. We find a statistically significant difference in price across the border between the districts with $p$=0.002, and estimate a 20% higher price on average for a house on the more desirable side.

*Keywords*: Gaussian processes; kriging; bayesian testing; causal inference; regression discontinuity; treatment effect; housing market

*This research was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1144152, by the National Science Foundation under Grant No. 1461435, by DARPA under Grant No. FA8750-14-2-0117, by ARO under Grant No. W911NF- 15-1-0172, and by NSERC. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, DARPA, ARO, or NSERC.

# 1 Introduction

Regression discontinuity designs (RDDs) are natural experiments characterized by the treatment assignment being fully determined by some covariates, which are termed "forcing" variables. A typical RDD scenario arises when a treatment is given to all units with a forcing variable that falls below (or above) an arbitrary threshold value, and is withheld from units on the other side of the threshold. If, as is often the case, the forcing variable is also predictive of the outcome of interest, then treatment assignment and outcomes are confounded, but by focusing on units near the threshold, a causal treatment effect can nonetheless be estimated. The theory and methods for RDDs date from the 1960s, starting with Thistlethwaite and Campbell (1960). Cook (2008) trace the history of how interest in RDDs subsequently waned, until the late 1990s when the design saw renewed attention, theoretical progress, and applications in the social sciences. More recently, beginning with Papay et al. (2011), methods have been developed to analyse RDDs with multiple forcing variables. Imbens and Zajonc (2011) extend the local linear regression methods (see Imbens and Lemieux, 2008) that are popular for analysing RDDs with a single forcing variable (1D RDDs) to settings with multiple forcing variables.

Geographical regression discontinuity designs (GeoRDDs) are such RDDs where the forcing variables are spatial, meaning that units within a certain region are assigned to treatment, while units in a neighboring region are assigned to control. For example, in MacDonald et al. (2015), a private police force patrols a neighborhood, but stays out of surrounding areas, and a causal effect on crime rates is sought. In Chen et al. (2013), a policy applies south of the Huai River in China but not in the north, and pollution levels and life expectancies are measured to infer environmental and health impacts of the policy. In our application, we seek to estimate the effect of school districts on house prices in New York City.

Practitioners often wish to use the well-established methods and software developed for 1D RDDs with their spatial data. It is therefore tempting to reduce a GeoRDD problem to a 1D RDD by using the signed distance from the boundary (positive for treatment and negative for control) as the forcing variable, a method that we refer to as "projected 1D RDD," and which is used by both examples cited above. However, this method can fail to capture the spatial variation in the outcomes, resulting in spatial confoundedness of the estimator. We demonstrate the resulting bias in a simple example in Section S-1 of the Supplementary Materials. See also Section 4.2 of Keele and Titiunik (2015) for a discussion of this issue.

A more principled treatment of the GeoRDD is offered by Keele and Titiunik (2015), who build theoretical foundations for the analysis of GeoRDDs. They extend the identification assumptions that were formalized by Hahn et al. (2001) for 1D RDDs, and Imbens and Zajonc (2011) for multivariate RDDs, to GeoRDDs. The main requirement for identification of the treatment effect is continuity of the conditional regression functions near the border. Notably, this is violated if units can sort around the border, crossing the border to seek or avoid the treatment, a particular concern in GeoRDDs. Keele and Titiunik (2015) discuss further pitfalls of GeoRDDs, such as the issue of compound treatments—when a geographical border determines the assignment of the treatment of interest, but also of other differences.

Several methods for estimating the treatment effect in GeoRDDs have been proposed. Keele and Titiunik (2015) and Keele et al. (2017) estimate the treatment effect using a modification of the projected 1D RDD method by applying it locally around each point along

the border, thus alleviating the problem of spatial confounding. This is slightly different from the method of Imbens and Zajonc (2011) developed for multivariate regression discontinuities that are not necessarily geographical: they use a multivariate local linear regression to estimate the treatment effect at each boundary point, thus avoiding the need to specify a distance metric in the space of covariates. Keele et al. (2015) propose an alternative approach: they deploy the matching methods of Zubizarreta (2012) to match units on opposite sides of the border that are near each other geographically and in other covariates, and then analyze the matched outcomes as if they came from a randomized experiment. Their method requires the analyst to choose a "buffer" distance from the border, so that within this distance spatial variation can plausibly be assumed to be negligible, resulting in estimates that are robust to spatial confounding.

In this paper, we propose a framework for analysing GeoRDDs that is a spatial analogue of 1D RDD methods. Broadly, 1D RDD methodologies are composed of three steps: (1) fit a smooth *function* to the outcomes against the forcing variable on each side of the threshold, (2) extrapolate the functions to the *threshold point*, and (3) take the difference between the two extrapolations to estimate the treatment effect at the threshold point. Reusing the same methodological skeleton and applying it to geographical RDDs, our framework proceeds analogously: (1) fit a smooth *surface* to the outcomes against the geographical covariates in each region, (2) extrapolate the surfaces to the *border curve*, and (3) take the *pointwise* difference between the two extrapolations to estimate the treatment effect along the border.

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

In Section 2, we use GPR to estimate the treatment effect along the border by extending the model of Branson et al. (2017) to geographical settings. A peculiarity of GeoRDDs is that the estimand is a function defined everywhere along the border, which is a one-dimensional manifold embedded in two-dimensional space. Furthermore, geographical borders, whether they be political or natural, are rarely simple straight lines. The topology of borders complicates the definition and interpretation of estimands for the local average treatment effect (LATE), which we address in Section 3. We obtain Bayesian estimators for multiple possible LATE estimands and discuss their properties. In Section 4 we turn to hypothesis testing, and propose a method to test against the null hypothesis of no treatment effect along the border.

In Section 5, we apply our methodology to a publicly available dataset of property sales in NYC to determine whether school districts affect property prices. Focusing on a single border between two school districs, we estimate the treatment effect everywhere along the border, obtain estimates of the LATE, and perform and validate a hypothesis test. We find a statistically significant difference in price across the border with a $p$-value of 0.002, and estimate that the same house located near the border will on average fetch an almost 20% higher price in district 27 than in district 19. However, this effect can not be attributed solely to the reputation of the school district, as this border also separates the boroughs of Brooklyn and Queens, thus confounding the causal effect of the districts.

## 2 GeoRDD Modeling with Gaussian processes

We largely adopt the setup and notation for GeoRDDs laid out in Keele and Titiunik (2015). The outcomes $Y_i$ of $n$ units with spatial coordinates $\boldsymbol{s}_i$ are observed within an area $\mathcal{A}$ of 2-dimensional coordinate space. The units are separated 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 GeoRDDs is that the two areas are adjacent but non-overlapping, intersecting only at the border $\mathcal{B}$ between them. Throughout this paper, points on the border are denoted by $\boldsymbol{b}$. Under the potential outcomes framework for causal inference, each unit $i$ has potential outcomes $Y_{iT}$ and $Y_{iC}$ under treatment and control respectively. Let $Z_i$ denote the treatment indicator, which is equal to one if unit $i$ is in the treatment area, and zero if it is in the control area. Unlike traditional randomized experiments, treatment assignment is a deterministic function of a unit's geographical coordinates $\boldsymbol{s}_i$: $Z_i = \mathbb{I}\{\boldsymbol{s}_i \in \mathcal{A}_T\}$. The observed outcome for unit $i$ is $Y_i = Z_i Y_{iT} + (1 - Z_i) Y_{iC}$. We denote the vector of observed outcomes of the treatment units and control units respectively by $\boldsymbol{Y}_T$ and $\boldsymbol{Y}_C$, and $\boldsymbol{Y}$ the vector formed by concatening $\boldsymbol{Y}_T$ and $\boldsymbol{Y}_C$.

For 1D RDDs, because the treatment and control regions do not overlap, the treatment effect is typically only inferred at the threshold $X = b$. As was already recognized by Thistlethwaite and Campbell (1960), this choice requires the least extrapolation of the fitted regression functions, which makes the estimated treatment more credible. The estimand at the threshold can be obtained as the difference of the two limits of the expectation of the conditional regression functions

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

where the second equality requires the assumption that the conditional regression functions $\mathbb{E}[Y_{iT} \mid X_i = x]$ and $\mathbb{E}[Y_{iC} \mid X_i = x]$ are continuous in $x$ (see Assumption 2.1 in Imbens and Lemieux (2008) and the discussion that follows). Analogously, we focus on the treatment effect at the border $\mathcal{B}$ between the treatment and control regions:

$$\tau \colon \mathcal{B} \to \mathbb{R} \quad \text{defined as} \quad \tau(\boldsymbol{b}) = \mathbb{E}[Y_{iT} - Y_{iC} \mid \boldsymbol{s}_i = \boldsymbol{b}]\,. \tag{2}$$

This is the functional estimand defined in Imbens and Zajonc (2011) and Keele and Titiunik (2015). For any $\boldsymbol{b} \in \mathcal{B}$, $\tau(\boldsymbol{b})$ can be obtained as the difference of the two limits of the expected outcomes, approaching $\boldsymbol{b}$ from the treatment or the control side of the border, given the assumption that the conditional regression functions $\mathbb{E}[Y_{iT} \mid \boldsymbol{s}_i = \boldsymbol{s}]$ and $\mathbb{E}[Y_{iC} \mid \boldsymbol{s}_i = \boldsymbol{s}]$ are continuous in $\boldsymbol{s}$ within $\mathcal{A}$. This result is formalized under Assumption 2.2.2 by Imbens and Zajonc (2011) and Assumption 1 in Keele and Titiunik (2015).

For computational reasons, we often represent the border as a set $\boldsymbol{b}_{1:R} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_R\}$, $\boldsymbol{b}_r \in \mathcal{B}$ of $R$ "sentinel points" along the border. We denote by $\tau(\boldsymbol{b}_{1:R})$ the $R$-vector with $r^{\text{th}}$ entry $\tau(\boldsymbol{b}_r)$ of the treatment effect evaluated at $\boldsymbol{b}_r$.

## 2.1 Model Specification

Our GeoRDD framework allows any method to be used to fit the outcomes on either side of the border. In this paper we use Gaussian process regression (GPR) for this purpose. GPR, known as kriging in the spatial statistics literature, is a Bayesian non-parametric method for fitting smooth functions. Recently, Branson et al. (2017) showed GPR to be a promising approach for the analysis 1D RDDs. Further inspired by the popularity of GPR in spatial statistics, we extend the model of Branson et al. (2017) to geographical RDDs.

On each side of the border, we model the observed outcomes $Y_i$ at location $\boldsymbol{s}_i$ as the sum of an intercept $m$, a spatial Gaussian process $f(\boldsymbol{s})$, 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, known as "kernels" in machine learning; see Banerjee et al. (2014) and Rasmussen and Williams (2006) for examples. In this paper we use the squared exponential kernel for its ease of understanding and its prevalence in applied spatial statistics. This yields the outcomes model:

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

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

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

This can be visualized as the height of a cliff along the border $\mathcal{B}$ separating the two smooth plains of the treatment and control regions.

In this specification, the hyperparameters $\ell$, $\sigma_{\text{GP}}$, and $\sigma_\epsilon$ are the same in the treatment and control regions, so we assume that the spatial smoothness of the responses is not affected by the treatment. We expect that this assumption will be reasonable in many applications, but it can be easily relaxed, as discussed in Branson et al. (2017).

## 2.2 Inference of the Treatment Effect

If $m_T$ and $m_C$ are given normal priors with variance $\sigma^2_m$, then the model specification (3) can be used to obtain covariances between the observations, the Gaussian processes, and the mean parameters. Given hyperparameters $\boldsymbol{\theta} = (\ell, \sigma_{\text{GP}}, \sigma_\epsilon, \sigma_m)$, any vector with entries consisting of observations, points on the potential outcomes surface $f_T$ and $f_C$, and the mean parameters $m_C, m_T$ is jointly multivariate normal. Therefore the distribution of any such vector conditioned on another is also multivariate normal, with mean and covariances analytically tractable, and easily computed.

In accordance with the framework laid out in Section 1, we proceed by extrapolating both Gaussian processes to the border, and then taking the difference of the predictions to obtain the posterior treatment effect along the border. Computationally, we need to represent this

border as a set $\boldsymbol{b}_{1:R} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_R\}$ of $R$ "sentinel" units dotted along $\mathcal{B}$. The extrapolation step then follows mechanically through multivariate normal theory. On the treatment side, for example:

$$\begin{aligned} &g_T(\boldsymbol{b}_{1:R}) \mid \boldsymbol{Y}_T, \boldsymbol{\theta} \sim \mathcal{N}\big(\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|T}, \Sigma_{\boldsymbol{b}_{1:R}|T}\big)\,, \text{ with} \\ &\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|T} = \boldsymbol{K}_{\mathcal{B}T}\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{Y}_T \quad \text{and} \quad \Sigma_{\boldsymbol{b}_{1:R}|T} = \boldsymbol{K}_{\mathcal{B}\mathcal{B}} - \boldsymbol{K}_{\mathcal{B}T}\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{K}_{\mathcal{B}T}^{\intercal}\,. \end{aligned} \tag{5}$$

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

$$\begin{aligned} &\tau(\boldsymbol{b}_{1:R}) \mid \boldsymbol{Y}, \boldsymbol{\theta} \sim \mathcal{N}\big(\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}, \Sigma_{\boldsymbol{b}_{1:R}|Y}\big)\,, \text{ with} \\ &\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y} = \boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|T} - \boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|C} \quad \text{and} \quad \Sigma_{\boldsymbol{b}_{1:R}|Y} = \Sigma_{\boldsymbol{b}_{1:R}|T} + \Sigma_{\boldsymbol{b}_{1:R}|C}\,. \end{aligned} \tag{6}$$

The posterior mean and covariance of the $\tau(\boldsymbol{b}_{1:R})$ are the primary output of our GeoRDD analysis, and we refer to (6) as the "cliff height" estimator.

This leaves the choice of the hyperparameters: $\boldsymbol{\theta} = \ell$, $\sigma_{\text{GP}}$, $\sigma_\epsilon$, and $\sigma_m$. For $\sigma_m$, we arbitrarily pick a large number, so that the prior on the mean parameters is weak. The rest are optimized by maximizing the marginal likelihood of the observations $\mathbb{P}(\boldsymbol{Y} \mid \ell, \sigma_{\text{GP}}, \sigma_\epsilon)$, which is available analytically and easily computed for GPR. This empirical Bayes approach is common in spatial and machine learning applications of Gaussian processes. An alternative would be to also specify a prior on the hyperparameters, which would be preferable in order to fully account for the uncertainty in the model, but fully Bayesian inference of large Gaussian process models tends to be computationally expensive.

## 2.3 Handling Nonspatial Covariates

The Gaussian process specification also makes it easy, mathematically and computationally, to incorporate a linear model on non-spatial covariates. The models are modified by the addition of the linear regression term $\boldsymbol{D}\boldsymbol{\beta}$ on the $n \times p$ matrix of covariates $\boldsymbol{D}$, where $p$ is the number of non-spatial covariates. We recommend placing a normal prior $\mathcal{N}(0, \sigma_\beta^2)$ on the regression coefficients, as this preserves the multivariate normality of the model, with the simple addition of a term $\sigma_\beta^2 \boldsymbol{D}\boldsymbol{D}^{\intercal}$ to the covariance $\boldsymbol{\Sigma}_Y$ of $\boldsymbol{Y}$. Let $\boldsymbol{d}_i$ be the $p$-vector of non-spatial covariates of unit $i$. Our model becomes:

$$\begin{aligned} &Y_{iT} = \underbrace{m_T + f_T(\boldsymbol{s}_i)}_{g_T(\boldsymbol{s}_i)} + \boldsymbol{d}_i^{\intercal}\boldsymbol{\beta} + \epsilon_i \quad \text{and} \quad Y_{iC} = \underbrace{m_C + f_C(\boldsymbol{s}_i)}_{g_C(\boldsymbol{s}_i)} + \boldsymbol{d}_i^{\intercal}\boldsymbol{\beta} + \epsilon_i\,, \text{ with} \\ &\beta_j \overset{\perp\!\!\!\perp}{\sim} \mathcal{N}\big(0, \sigma_\beta^2\big)\,, \text{ for } j = 1, 2, \ldots, p\,, \end{aligned} \tag{7}$$

and $f_T$ and $f_C$ as in (3).

Unfortunately, the linear term induces a covariance between the treatment and control region; $\boldsymbol{\Sigma}_Y$ is no longer black diagonal, which roughly quadruples the computational cost of the analysis, as it requires the inversion of an $(n_T + n_C) \times (n_T + n_C)$ covariance matrix

instead of (in the absence of correlations between the treatment and control units) the separate inversions of the $n_T \times n_T$ covariance matrix $\boldsymbol{\Sigma}_{TT}$, and $n_C \times n_C$ covariance matrix $\boldsymbol{\Sigma}_{CC}$. The introduction of the linear term modifies the cliff height estimator (6) so that its posterior mean and covariance become:

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y,D} &= [\boldsymbol{K}_{\mathcal{B}T} \;\; -\boldsymbol{K}_{\mathcal{B}C}]\, \boldsymbol{\Sigma}_Y^{-1} \boldsymbol{Y} \,, \text{ and} \\ \Sigma_{\boldsymbol{b}_{1:R}|Y,D} &= 2\boldsymbol{K}_{\mathcal{B}\mathcal{B}} - [\boldsymbol{K}_{\mathcal{B}T} \;\; -\boldsymbol{K}_{\mathcal{B}C}]\, \boldsymbol{\Sigma}_Y^{-1}\, [\boldsymbol{K}_{\mathcal{B}T} \;\; -\boldsymbol{K}_{\mathcal{B}C}]^\intercal \,. \end{aligned} \tag{8}$$

To avoid the complexity caused by the correlation between $\boldsymbol{Y}_T$ and $\boldsymbol{Y}_C$ that the linear term induces, we suggest first obtaining an estimate $\hat{\boldsymbol{\beta}}$ of the coefficients. We show how to obtain the posterior mean of $\boldsymbol{\beta}$ in Appendix B. We can then proceed with the GeoRDD analysis on the residuals, which are decorrelated conditionally on $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. This is an approximation, as it ignores the uncertainty in the estimate of $\boldsymbol{\beta}$, but if the number of samples in the treatment and control areas is high (even away from the border), the approximation has negligible effect on the estimate of the treatment effect, and simplifies the subsequent GeoRDD analysis.

# 3 Estimating the Local Average Treatment Effect

Once we have obtained the posterior of $\tau$ (6), estimating the local average treatment effect (LATE) along the border will often be of interest. We consider the class of weighted means of the functional treatment effect $\tau(\boldsymbol{b})$, with weight function $w_{\mathcal{B}}(\boldsymbol{b})$ defined everywhere on the border $\mathcal{B}$. The weighted mean integral can be approximated as a weighted sum at the sentinels $\boldsymbol{b}_{1:R}$:

$$\tau^w = \frac{\oint_{\mathcal{B}} w_{\mathcal{B}}(\boldsymbol{b})\tau(\boldsymbol{b})\, d\boldsymbol{b}}{\oint_{\mathcal{B}} w_{\mathcal{B}}(\boldsymbol{b})\, d\boldsymbol{b}} \approx \frac{\sum_{r=1}^R w_{\mathcal{B}}(\boldsymbol{b}_r)\tau(\boldsymbol{b}_r)}{\sum_{r=1}^R w_{\mathcal{B}}(\boldsymbol{b}_r)} \,. \tag{9}$$

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

$$\mu_{\tau^w|Y} = \frac{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^\intercal \boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}}{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^\intercal\, \mathbf{1}_R} \quad \text{and} \quad \Sigma_{\tau^w|Y} = \frac{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^\intercal \Sigma_{\boldsymbol{b}_{1:R}|Y} w_{\mathcal{B}}(\boldsymbol{b}_{1:R})}{\left(w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^\intercal\, \mathbf{1}_R\right)^2} \,, \tag{10}$$

where $w_{\mathcal{B}}(\boldsymbol{b}_{1:R})$ is the $R$-vector of weights evaluated at the sentinels, and $\mathbf{1}_R$ is an $R$-vector of ones. For each estimator obtained in (10) as a weighted mean of $\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}$, we consider the "natural" estimand to be the same weighted mean applied to the true $\tau$, given by (9).

An alternative perspective on these estimators is given by the weights induced on the observations. Indeed, combining equations (5), (6), and (10), we obtain that the posterior mean of $\tau^w$ is a linear combination

$$\mathbb{E}(\tau^w \mid \boldsymbol{Y}) = \boldsymbol{w}_T^\intercal \boldsymbol{Y}_T + \boldsymbol{w}_C^\intercal \boldsymbol{Y}_C \tag{11}$$

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

$$\boldsymbol{w}_T = \frac{\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{K}_{\mathcal{B}T}^{\intercal} w_{\mathcal{B}}(\boldsymbol{b}_{1:R})}{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^{\intercal}\,\mathbf{1}_R} \quad \text{and} \quad \boldsymbol{w}_C = \frac{-\boldsymbol{\Sigma}_{CC}^{-1}\boldsymbol{K}_{\mathcal{B}C}^{\intercal} w_{\mathcal{B}}(\boldsymbol{b}_{1:R})}{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^{\intercal}\,\mathbf{1}_R}\,, \tag{12}$$

for treatment and control units respectively.

The question remains: what is the most appropriate choice of weights? We next motivate and consider four possible choices of $w_{\mathcal{B}}(\boldsymbol{b})$, and explore interpretations, advantages, and drawbacks. In Section S-2 of the Supplementary Materials, we discuss two further choices, the projected land LATE $\tau^{\text{GEO}}$, and the projected superpopulation LATE $\tau^{\text{POP}}$. We also provide a simulation study to better understand the characteristics of the different LATE choices. A summary of their properties is provided in Table 1.

## 3.1 Uniform LATE

The simplest choice is uniform weights $w_{\mathcal{B}}(\boldsymbol{b}) = 1$, a seemingly reasonable and unopinionated decision. The uniformly weighted LATE $\tau^{\text{UNIF}}$ is estimated by averaging the entries of the mean posterior at the sentinels. Following (9) and (10):

$$\begin{aligned}
&\tau^{\text{UNIF}} = \oint_{\mathcal{B}} \tau(\boldsymbol{b})\,d\boldsymbol{b} \;\Big/\; \oint_{\mathcal{B}} d\boldsymbol{b}\,,\\
&\tau^{\text{UNIF}} \mid \boldsymbol{Y}, \boldsymbol{\theta} \sim \mathcal{N}\big(\mu_{\tau^{\text{UNIF}}|Y}, \Sigma_{\tau^{\text{UNIF}}|Y}\big)\,, \text{ with}\\
&\mu_{\tau^{\text{UNIF}}|Y} = \big(\mathbf{1}_R^{\intercal}\,\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}\big)/R \quad \text{and} \quad \Sigma_{\tau^{\text{UNIF}}|Y} = \big(\mathbf{1}_R^{\intercal}\,\Sigma_{\boldsymbol{b}_{1:R}|Y}\,\mathbf{1}_R\big)/R^2\,.
\end{aligned} \tag{13}$$

The uniformly weighted estimand takes on a geometric interpretation: equal-length segments of the border are given equal weight. Unfortunately, uniform weights suffer from several issues that we describe and address in Section 3.2 and Section 3.3.

## 3.2 Density Weighted LATE

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, such as a lake or a public park, then the treatment effect there has little meaning and importance. Furthermore, $\tau(\boldsymbol{b})$ in those empty areas will have large posterior variances, which will dominate the posterior variance of $\tau^{\text{UNIF}}$, potentially jeopardizing the successful detection of otherwise strong treatment effects.

We can address this issue by weighting the treatment effect at each sentinel location by the local population density $\rho$, i.e. choosing $w_{\mathcal{B}}(\boldsymbol{b}) = \rho(\boldsymbol{b})$. Attractively, the estimand is interpretable as the average treatment effect for the superpopulation of units that live on the border:

$$\tau^{\rho} = \mathbb{E}[Y_{iT} - Y_{iC} \mid \boldsymbol{s}_i \in \mathcal{B}]\,. \tag{14}$$

It therefore better captures the "typical" treatment effect received by a unit than the uniformly weighted estimand. 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 $\tau^\rho$, 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: its susceptibility to the topology of the border. If a section of the border has more twists and turns—for example if it follows the course of a meandering river—then that section will receive disproportionately more sentinels. The unweighted and density-weighted mean treatment estimands are both affected by this effect, which gives higher weight to wigglier sections of the border. See Section S-2.3 of the Supplementary Materials for a simulation demonstrating this susceptibility to border topology. Consider, for illustration purposes, the border separating the two American states of Louisian and Mississippi, depicted in Figure S-6 of the Supplementary Materials. From North to South, it 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 follows until it reaches the Gulf of Mexico. Consequently, sentinels placed at equal distance intervals along this border will be more densely packed along the rivers, and sparsest along the straight segment. When averaging a function over the border, those sections become 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.

## 3.3 Inverse-variance Weighted LATE

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

$$\begin{aligned} &\tau^{\text{INV}} \mid \boldsymbol{Y}, \boldsymbol{\theta} \sim \mathcal{N}\left(\mu_{\tau^{\text{INV}}|Y}, \Sigma_{\tau^{\text{INV}}|Y}\right), \text{ with} \\ &\mu_{\tau^{\text{INV}}|Y} = \left(\mathbf{1}_R^\intercal \Sigma^{-1}_{\boldsymbol{b}_{1:R}|Y} \boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}\right) \Big/ \left(\mathbf{1}_R^\intercal \Sigma^{-1}_{\boldsymbol{b}_{1:R}|Y} \mathbf{1}_R\right) \quad \text{and} \quad \Sigma_{\tau^{\text{INV}}|Y} = 1 \Big/ \left(\mathbf{1}_R^\intercal \Sigma^{-1}_{\boldsymbol{b}_{1:R}|Y} \mathbf{1}_R\right). \end{aligned} \tag{15}$$

This estimator efficiently extracts the information from the posterior treatment effect, as it can be shown to minimize the posterior variance amongst weighted averages of the form (9). It automatically gives more weight to sentinels in dense areas (as the variance will be lower there), and to sentinels in straight sections of the border (as the correlations between sentinels will be lower).

The estimand is still a weighted mean, with weights for the sentinels given by $w_{\mathcal{B}}(\boldsymbol{b}_{1:R}) =$

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

## 3.4 Projected Finite-Population LATE

All LATE 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 that are within a distance $\Delta$ of the border onto the border, and use those projected unit locations without weights (see Figure S-2 of the Supplementary Materials for an illustration). For any point $\boldsymbol{s}$, we use the notation $\text{proj}_{\mathcal{B}}(\boldsymbol{s})$ to give the coordinates of the point on the border $\mathcal{B}$ that is closest to $\boldsymbol{s}$ (assuming uniqueness), and $\text{dist}_{\mathcal{B}}(\boldsymbol{s})$ for the distance between the point and the border. Let $\mathbb{I}^{\Delta}(\boldsymbol{s}) = \mathbb{I}\{\text{dist}_{\mathcal{B}}(\boldsymbol{s}) \leq \Delta\}$ indicate inclusion in the border vicinity. The projected finite-population $\tau^{\text{PROJ}}$ is then the uniformly weighted mean applied with the projected unit locations instead of the evenly spaced sentinels. We can therefore modify (13), replacing the cliff height mean vector $\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}$ and covariance matrix $\Sigma_{\boldsymbol{b}_{1:R}|Y}$ with their equivalents obtained at the projected unit locations, to obtain the posterior mean and covariance of $\tau^{\text{PROJ}}$:

$$
\begin{aligned}
&\tau^{\text{PROJ}} \mid \boldsymbol{Y}, \boldsymbol{\theta} \sim \mathcal{N}\big(\mu_{\tau^{\text{PROJ}}|Y}, \Sigma_{\tau^{\text{PROJ}}|Y}\big)\,, \text{ with} \\
&\mu_{\tau^{\text{PROJ}}|Y} = \sum_{i=1}^{n} \mathbb{I}^{\Delta}(\boldsymbol{s}_i)\,\mathbb{E}[\tau(\text{proj}_{\mathcal{B}}(\boldsymbol{s}_i)) \mid \boldsymbol{Y}, \boldsymbol{\theta}] \;\Big/\; \sum_{i=1}^{n} \mathbb{I}^{\Delta}(\boldsymbol{s}_i)\,, \text{ and} \\
&\Sigma_{\tau^{\text{PROJ}}|Y} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{I}^{\Delta}(\boldsymbol{s}_i)\,\mathbb{I}^{\Delta}(\boldsymbol{s}_j)\,\text{Cov}[\tau(\text{proj}_{\mathcal{B}}(\boldsymbol{s}_i)), \tau(\text{proj}_{\mathcal{B}}(\boldsymbol{s}_j)) \mid \boldsymbol{Y}, \boldsymbol{\theta}]}{\big(\sum_{i=1}^{n} \mathbb{I}^{\Delta}(\boldsymbol{s}_i)\big)^2}\,.
\end{aligned}
\tag{16}
$$

The posterior expectations and covariances in (16) are easily derived and computed analogously to the procedure of Section 2.2. Note that $\tau^{\text{PROJ}}$ is in the class of weighted mean estimands (9), with weight function $w_{\mathcal{B}}(\boldsymbol{b}) = \sum_{i=1}^{n} \mathbb{I}^{\Delta}(\boldsymbol{s}_i)\delta(\boldsymbol{b} - \text{proj}_{\mathcal{B}}(\boldsymbol{s}_i))$, where $\delta$ is the Dirac delta function.

The resulting estimator has desirable properties: densely populated regions receive proportionately more projected units, but wigglier segments of the border do not. While it lacks the information efficiency of the inverse-variance estimator, the projected estimand is easier to understand and interpret, and may feel more familiar to practitioners used to finite-population inference. The averaging is over the observed units in the vicinity of the border, after they have been moved to the nearest point on the border.

In our experience, the choice of $\Delta$ does not have a large effect on the estimate yielded

| Notation | Description | $\mathcal{B}$ Topology | Sentinels | Principle | Variance |
|---|---|---|---|---|---|
| $\tau^{\text{UNIF}}$ | Uniform | Sensitive | Equispaced | Geometry | High |
| $\tau^{\rho}$ | Density-weighted | Sensitive | Equispaced | Population | Low |
| $\tau^{\text{INV}}$ | Inverse-var. weighted | Robust | Equispaced | Information | Lowest |
| $\tau^{\text{PROJ}}$ | Projected finite pop. | Robust | Projected | Finite pop. | Low |
| $\tau^{\text{GEO}}$ | Proj. land | Robust | Proj. Grid | Geography | High |
| $\tau^{\text{POP}}$ | Proj. superpop. | Robust | Proj. Grid | Population | Low |

Table 1: Summary of local average treatment effect estimator and estimand properties.

by (16). A reasonable heuristic is to set $\Delta$ to a small multiple of the Gaussian process lengthscale $\ell$. It should be noted that this choice only affects the location and density of projected units on the border; the $\tau^{\text{PROJ}}$ estimator assigns non-zero unit weights (11) to all units, whether or not they fall within $\Delta$ of the border.

### 3.5 Summary

The properties of the four LATE definitions proposed in this paper, and two additional choices presented in the Supplementary Materials, are summarized in Table 1. 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 density weighted LATE $\tau^{\rho}$ it does not require estimating population density. Meanwhile, the inverse-variance estimator is the most efficient (lowest posterior variance) weighted mean estimator, and sidesteps the choice of a distance cutoff for projected units.

## 4 Testing for Non-Zero Effect

Once we have obtained the "cliff height" estimate (6) and estimated a LATE, we might also naturally wonder whether we can claim to have detected a significant treatment effect at the border. In the hypothesis testing framework, we have two possible choices of null hypotheses. The sharp null specifies that the treatment effect is zero everywhere along the border: $\tau(\boldsymbol{b}) = 0$ for all $\boldsymbol{b} \in \mathcal{B}$. Meanwhile, the weak null only requires the LATE to be zero. We focus on a test of the weak null hypothesis here, but also provide two tests of the sharp null hypothesis based on the marginal liklihood and a chi-squared statistic in Section S-3 of the Supplementary Materials. We found through simulations and in our applied example that the test presented in this paper has superior power and robustness to model misspecification, and therefore recommend its use.

As we saw in Section 3, the LATE estimand can be defined in multiple ways. If we choose the inverse-variance weighted mean, then $\tau^{\text{INV}}$ has posterior given by (15). While the posterior is a Bayesian object, we can use it heuristically to derive a pseudo-$p$-value $\tilde{p}^{\text{INV}} = 2\Phi\big(-\big|\mu_{\tau^{\text{INV}}|Y}\big|\big/\sqrt{\Sigma_{\tau^{\text{INV}}|Y}}\big)$. However, this pseudo-$p$-value obtained from the Bayesian

posterior may not have good frequentist properties. In particular, there is no guarantee that under the null hypothesis, $p^{\text{INV}}$ is below 0.05 less than 5% of the time.

To turn it into a valid frequentist test, it can be calibrated using a parametric bootstrap under the null. We specify a parametric null model $\mathscr{M}_0$ as a single Gaussian process spanning the control and treatment regions, with the same kernel and hyperparameters values obtained through the procedure of Section 2.2. $\mathscr{M}_0$ is smooth and continuous at the border, and therefore accords with both the sharp and weak null hypotheses. We now choose the posterior mean of the inverse-variance LATE $\mu_{\tau^{\text{INV}}|Y}$ as a test statistic. For $b = 1, \ldots, B$ iterations, we draw $\boldsymbol{Y}^{(b)}$ from $\mathscr{M}_0$, using the same spatial locations as the original data, and compute $\mu_{\tau^{\text{INV}}|Y^{(b)}}$ according to (15) applied to the simulated data rather than the true data. The proportion of $\mu_{\tau^{\text{INV}}|Y^{(b)}}$ with absolute value greater than the observed $\mu_{\tau^{\text{INV}}|Y^{obs}}$ estimates the $p$-value:

$$p^{\text{INV}} = \mathbb{P}\left(\left|\mu_{\tau^{\text{INV}}|Y}\right| \geq \left|\mu_{\tau^{\text{INV}}|Y^{obs}}\right| \mid \mathscr{M}_0\right) \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left\{\left|\mu_{\tau^{\text{INV}}|Y^{(b)}}\right| \geq \left|\mu_{\tau^{\text{INV}}|Y^{obs}}\right|\right\}. \tag{17}$$

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.

The calibration can also be achieved analytically, since $\mu_{\tau^{\text{INV}}|Y}$ is normally distributed under the null hypothesis. We derive the analytical calibration of hypothesis tests based on any LATE estimand in Appendix C. Note that the $p$-value for this test is derived under the parametric null model $\mathscr{M}_0$, which accords with both the sharp null and weak null hypotheses, but is not the only possible model that satisfies the weak null. The calibrated inverse-variance test targets the weak null hypothesis in the sense that the test statistic is an estimate of the LATE, and thus the test is sensitive to deviations of the LATE from zero, rather than its $p$-value being derived directly under the weak null (such as the classical $t$-test).

## 4.1 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 practitioners to probe the validity of the hypothesis test 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 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. Here we drew lines that split the placebo units in half at a sequence of angles $1°, 2°, 3°, \ldots, 180°$ counterclockwise from horizontal, each positioned so that half of the units fall on either side of the line in order to maximize power. Because the border was chosen arbitrarily by us, without reference to the outcomes, we should not expect to see a discontinuous jump in outcomes at

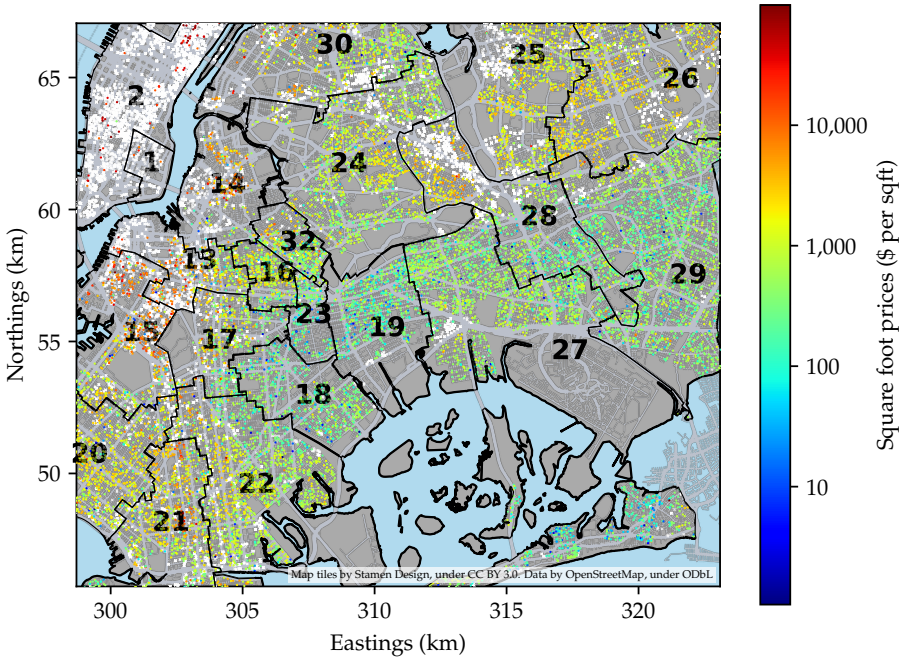


Figure 1: Map of property sales in New York City. Each dot is a sale, and its color indicates the price per square foot. White crosses indicate sales of properties with missing square footage, which are therefore excluded from the analysis. School district boundaries are shown, and each district is labeled by its number.

this border. We apply the calibrated inverse-variance test procedure described above to this arbitrarily divided data, store the results, and hope to obtain a roughly uniform distribution of $p$-values. The resulting $p$-values will obviously be highly correlated, so we should only expect a very roughly uniform distribution (because of the small effective sample size), but at the very least, this procedure allows us to visually verify that the $p$-values are not blatantly biased.

# 5 Application: NYC School Districts

We illustrate the analysis of a GeoRDD with house sales data in New York City. The city publishes information pertaining to property sales within the city in the last 12 months on a rolling basis, available at https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page, which we downloaded on September 15, 2016. The dataset includes columns for 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. It is a common belief that school districts 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 the borders separating school districts?

In order to model the property sale prices, we first need to obtain their locations. We

geocode the address of each sale by merging the sales with NYC's Pluto database, which contains X and Y coordinates for each house, identified by its borough, zip code, block and lot. These coordinates are given in the EPSG:2263 projection, which we also adopt. 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 onto EPSG:2263.

We then filter the 56,815 sales, by removing 36,448 sales outside of the family homes building class categories (one, two, and three family dwellings), 4 remaining sales missing the square footage information, and 785 remaining sales with outlier log price per square foot less than 3 or more than 8. We exclude condos and coops because only very few sales report square footage alongside the price. The resulting dataset of 19,578 sales is displayed in Figure 1. The 27,394 residential properties with missing square footage information are also shown; these are almost all coops and condos, which explains the clustering of missing data in areas of higher density.

## 5.1 Model for Property Prices

The outcome of interest is price per square foot. As is commonly done in analyses of real estate prices, we take its logarithm to reduce the skew of the outcome. The complete model is then a Gaussian process within each district (indexed by $j = 1, \ldots, J_{\texttt{Distr}}$) over the spatial covariates $\boldsymbol{s}$, super-imposed with a linear regression on the property covariates (which are $L_{\texttt{BuildClass}}$ building categories encoded as dummy variables). This model can be written:

$$\begin{aligned} Y_i &= m_{\texttt{Distr}[i]} + \beta_{\texttt{BuildClass}[i]} + f_{\texttt{Distr}[i]}(\boldsymbol{s}_i) + \epsilon_i\,, \quad \epsilon_i \sim \mathcal{N}\big(0, \sigma_y^2\big)\,, \\ \beta_l &\sim \mathcal{N}\big(0, \sigma_\beta^2\big)\,, \qquad \text{for } l = 1, \ldots, L_{\texttt{BuildClass}}\,, \\ m_j &\sim \mathcal{N}\big(0, \sigma_m^2\big)\,, \qquad \text{for } j = 1, \ldots, J_{\texttt{Distr}}\,, \\ f_j &\sim \mathcal{GP}(0, k(\boldsymbol{s}, \boldsymbol{s}'))\,, \quad \text{for } j = 1, \ldots, J_{\texttt{Distr}}\,, \end{aligned} \tag{18}$$

where $k$ is the squared exponential covariance function as in (3).

A visual inspection of the house sales map in Figure 1 drew our attention towards the border between districts 19 and 27. We arbitrarily designate district 19 as the "treatment" area and district 27 as the "control" area. Importantly, the border between the two districts is also part of the border between Brooklyn and Queens, so we will not be able to attribute a difference in price solely to the causal effect of the school districts. This is an instance of what Keele and Titiunik (2015) term "compound treatments," a frequent concern in GeoRDDs. Therefore, we are mainly *measuring* a discontinuity in the house prices at the border. Attributing the discontinuity to a particular cause (school district or borough) is not directly supported by the data.

Another concern is units sorting around the border, which would violate the identification assumptions for GeoRDDs. If people move across the border to live in a better school district, does this invalidate the analysis? We take the view that the unit of analysis here is the tract of land on which houses are built, rather than the residents themselves. If a district becomes more attractive, people may move to it, whereas land does not move but its price adjusts. A sale gives a snapshot of the price of the land, made more accurate by correcting in our model (18) for covariates that pertain to the building rather than land. Note that of course,

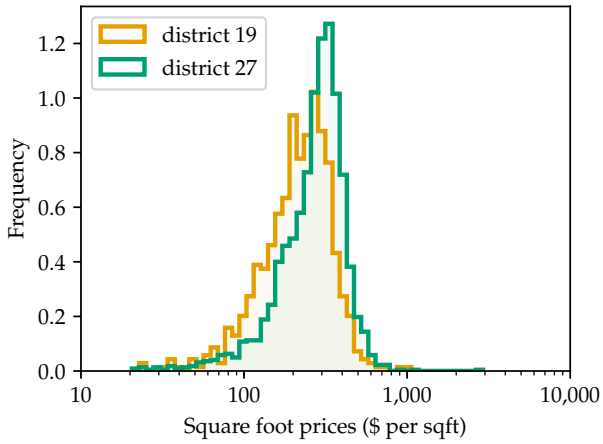


Figure 2: Histogram of log sale prices per square foot of houses sold in NYC school districts 19 and 27.

the limited covariates provided by the data cannot fully capture the value of the building. For example, the wealthier residents who inhabit the more desirable school districts may also have more funds available to maintain and enhance their home, which will drive up the property's resale value. Since it is not captured by the available covariates, this added value is folded into the treatment effect by our analysis.

The histogram in Figure 2 of log prices per square foot for sales in both districts shows that marginally the house prices are very different. Our goal is to establish whether this difference is measurable at the border, and not merely an underlying trend that spans both districts.

We fit the hyperparameters $\sigma_\beta$, $\sigma_{\text{GP}}$, $\ell$ and $\sigma_\epsilon$ by optimizing the marginal log-likelihood of the data within neighboring school districts 18, 19, 23, 24, 25, 26, 27, 28, and 29. We hold $\sigma_m$ fixed to 20 to give the district means $m_j$ a weak prior. The fitted hyperparameters were $\widehat{\sigma_\epsilon} = 0.40$, $\widehat{\sigma_{\text{GP}}} = 0.20$, $\widehat{\sigma_\beta} = 0.15$, and $\widehat{\ell} = 1.4$ km.

## 5.2 Cliff Height Estimator

We seek to estimate the treatment effect function $\tau$ on the border between the two districts. We could proceed by computing the cliff height estimator with covariates (8). But to simplify the analysis as discussed in Section 2.3, we can instead obtain the posterior means of the $\beta$ coefficients (following the procedure outlined in Appendix B, but extended to $J_{\texttt{Distr}}$ rather than just two areas), and extract the residuals $\boldsymbol{Y} - \boldsymbol{D}\hat{\boldsymbol{\beta}}$. We then treat the residuals as the observed outcomes in a GeoRDD analysis with no non-spatial covariates. In this example, we find that the posterior variance of $\boldsymbol{\beta}$ is low, and therefore the two approaches yield very similar results, but conditioning on the estimate of $\boldsymbol{\beta}$ is computationally convenient.

Following the inference procedure outlined in Section 2.1, we obtain the posterior distribution of the cliff height $\tau(\boldsymbol{b}_{1:R})$ obtained at $R = 100$ sentinel locations evenly spaced along the border. The cliff height is shown in Figure 3, and shows that $\tau$ is estimated as negative everywhere along the border, which corresponds to higher property prices in district

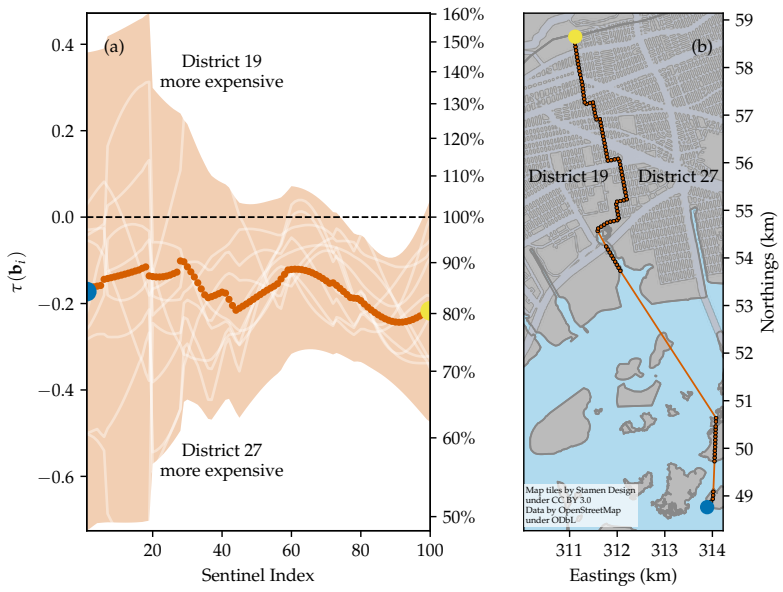


Figure 3: (a) Cliff height estimator (6) for the school district effect on house prices per square foot between district 27 and district 19, with 95% credible envelope. The left $y$-axis shows the difference in log prices per square foot; positive values mean prices are higher in district 19. The right $y$-axis shows the corresponding ratio of the price of a house in district 19 over its price in district 27. A few draws from the posterior are shown in lighter color to show the posterior correlations between sentinels. Note the decorrelation from sentinels 6 to 7, and 19 to 20, where the border crosses the water between sparsely populated islands in Jamaica Bay and then onto Long Island. (b) A map of sentinel locations, evenly spaced along the border between school districts 27 and 19. The southernmost sentinel (shown as a blue circle in both plots) has index 1, while the northernmost sentinel (shown in yellow) has index 100.

27. However, the credible envelope is fairly wide, especially in the southern section of the border, so we cannot visually rule out the null hypothesis that $\tau = 0$.

## 5.3 Average Log-Price Increase

The cliff height Figure 3 shows a negative treatment effect everywhere along the border, which can be averaged by the estimators we developed in Section 3. Our two recommended estimators, based on inverse-variance weighting and finite-population projection of units within $\Delta = 2\ell$ of the border, yield LATE estimates of $-0.19$ and $-0.18$ respectively, which corresponds to an almost 20% increase in property prices going from district 19 to district 27. By contrast, treating each district and building class as a fixed effect in an ordinary least squares (OLS) model yields a treatment effect estimate (the difference between the district 19 and 27 coefficients) of -0.12. This smaller estimate could be explained by an overall East to West positive spatial trend in prices, visible between districts 29 and 15 in Figure 1, which would confound the OLS estimate of the treatment effect. All LATE estimators from Section 3 applied to this setting are shown in Table 2. In this example the

| Estimand | Mean | Posterior Standard Dev. | Tail Prob. |
|---|---|---|---|
| $\tau^{\text{UNIF}}$ | -0.17 | 0.08 | 1.98% |
| $\tau^{\rho}$ | -0.19 | 0.06 | 0.04% |
| $\tau^{\text{INV}}$ | -0.19 | 0.06 | 0.03% |
| $\tau^{\text{PROJ}}$ $(\Delta = 2\ell)$ | -0.18 | 0.06 | 0.13% |
| $\tau^{\text{GEO}}$ $(\Delta = 2\ell)$ | -0.16 | 0.09 | 3.80% |
| $\tau^{\text{POP}}$ $(\Delta = 2\ell)$ | -0.18 | 0.06 | 0.15% |

Table 2: Average difference in log price per square foot between school districts 19 and 27. For each LATE estimand, we show the mean and standard deviation of its posterior distribution, and the tail probability $\mathbb{P}(\tau > 0 \mid \boldsymbol{Y}, \hat{\beta}, \boldsymbol{\theta})$. Negative LATEs correspond to district 27 being more expensive.

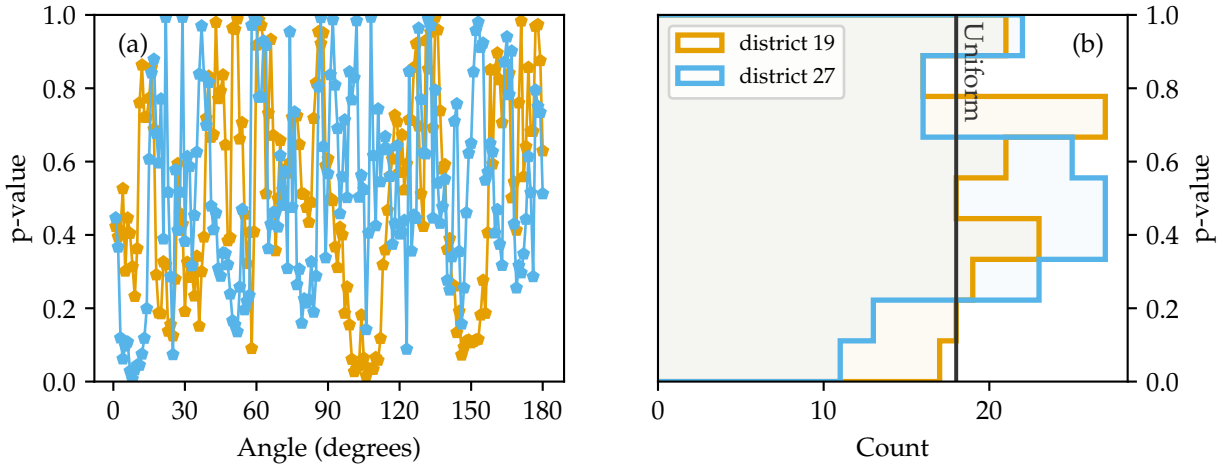


Figure 4: Placebo tests for calibrated inverse-variance test of difference in house prices at the border between NYC school districts 19 and 27. (a) the placebo $p$-value as a function of the border angle; (b) histogram of the placebo $p$-values, with the black vertical line indicating the uniform distribution.

different estimators yield similar answers, as the border is fairly straight and short relative to the fitted lengthscale.

## 5.4 Significant Difference in Price?

The estimated inverse-variance weighted mean treatment effect is suggestive of a significant treatment effect. But the posterior tail probability cannot be interpreted as a $p$-value. For this, we turn to the test developed in Section 4, which yields a $p$-value of $p^{\text{INV}} = 0.002$, thus rejecting the null hypothesis that there is no difference in house prices at the border between districts 19 and 27.

To assess the validity of the test, we apply the placebo tests devised in Section 4.1, the results of which are shown in Figure 4. Within each district, we split the data in half by a line at angles $1°$, $2°$, ..., $180°$. Because these lines were drawn arbitrarily, we do not

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—there is in fact noticeable autocorrelation in the graph of the placebo $p$-value as a function of angle—and so the low effective sample size could lead to some apparent departures from uniformity. Nonetheless, we do not observe a flagrant bias in Figure 4(b), which therefore does not discredit the calibrated inverse-variance test, and confirms the significance of the difference in price at the border between the two districts.

# 6 Conclusion

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

Previous research has focused on extending methods developed for 1D RDDs to GeoRDDs. In applied settings, some have used the signed distance from the border as the forcing variable in a 1D RDD, but the resulting estimator is spatially confounded. In this paper, we emphasize the importance of the spatial aspect of the design, and therefore draw from the spatial statistics literature, which brings a rich set of tools designed to model and exploit spatial correlations. We used Gaussian process regression, known as kriging in the spatial statistics literature, to fit the smooth surfaces to the outcomes in step (1) of our framework. Our approach yields a multivariate normal posterior distribution of the treatment effect for a collection of "sentinel" locations along the border.

Averaging the treatment effect along the border turns out to have surprising pitfalls. Simply integrating the treatment effect uniformly along the border yields an estimand that is inefficient and undesirably sensitive to the topology of the border. More sophisticated estimands, summarized in Table 1, are robust to this effect, and use the information available in the data more efficiently.

To test against the null hypothesis of zero treatment effect along the border, we develop a test based on the posterior distribution of the LATE. We use the inverse-variance weighted LATE to attain high power, but the other LATE estimates of Section 3 could be used similarly. To ensure good frequentist properties we calibrate the test, obtaining its distribution under the null model, either using a parametric bootstrap or analytically.

We applied our method to a publicly available dataset of one year of New York City property sales, to examine whether school district cause difference in property prices. Focusing on the border between school districts 19 and 27, we estimated a roughly 20% average increase in house prices per square foot when crossing the border from district 19 to district

27. However, the border between these two districts is also the border between the NYC boroughs of Brooklyn and Queens, so we cannot attribute this difference to the causal effect of the school districts. In the Supplementary Materials, we extend the GeoRDD analysis to other pairs of adjacent school districts, and find significant effects between many of these pairs. However, in some of these cases, physical barriers like parks, commercial zones, railways, and major roads can separate neighborhoods, keep data away from the borders, break the stationarity assumption of the spatial model, and increase the amount of extrapolation performed by the model, which casts doubt on the legitimacy of the estimated treatment effects. Missing data from condo sales which do not report square footage can also distort estimated effects. Overall, it seems that school district borders in Brooklyn and Queens are often accompanied by a discontinuity in house prices, but the causal attribution of this difference to the reputation of the school districts can be questionable.

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. As is common in spatial statistics, we use Gaussian process regression as a non-parametric smoothing device that flexibly captures spatial correlations, but do not claim that our model is a true representation of the stochastic mechanism generating the data. We believe care must therefore be taken not to lean heavily on modeling assumptions. In particular, we recommend that hypothesis tests always be accompanied by placebo tests: by applying the same procedure with arbitrary borders where no treatment was applied, we can verify that the test behaves appropriately under the null hypothesis, despite any potential model misspecification. We also assumed a stationary covariance structure, with hyperparameters equal in the treatment and control regions, and in particular we chose the squared exponential kernel. This kernel makes smoothness assumptions that are often considered unrealistic in geostatistical settings; the Matérn covariance family is often recommended as a more robust alternative. The assumption of equal covariance parameters in the two areas can also be relaxed, by separately tuning the parameters within each area.

Because of the need to extrapolate the fitted processes a short distance to the border, our GeoRDD method may be vulnerable to the limitations of Gaussian processes when extrapolating. The distinction between interpolation and extrapolation of spatial models is explored in some depth in Stein (2012). We expect that methodological advances that improve the extrapolating behavior of Gaussian processes would also improve the robustness of our method. For example, Wilson and Adams (2013) develop spectral mixture (SM) covariance kernels with good extrapolating behavior, which could be applied beneficially to GeoRDDs. However, SM kernels are motivated by time series 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 GPR 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 instead of the normal likelihood used in this paper.

Furthermore, in some applications, it may be of substantive interest to know whether the treatment effect is constant (homogenous) or variable (heterogenous). Hypothesis tests targeting the homogeneity of the treatment effect along the border would be an interesting possible extension of our framework.

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 Gaussian process model would need to be augmented with a temporal component, for example with an anisotropic squared exponential covariance function. We leave spatio-temporal RDDs using Gaussian process models to future research.

# References


Antonelli, J., M. Cefalu, and L. Bornn (2016). The positive effects of population-based preferential sampling in environmental epidemiology. *Biostatistics 17*(4), 764–778.

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, 04). A nonparametric bayesian methodology for regression discontinuity designs.

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), 12936–12941.

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

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika 96*(1), 187–199.

Ding, P. (2014, 02). A paradox from randomization-based causal inference.

Hahn, J., P. Todd, and W. Van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica 69*(1), 201–209.

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

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

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

Keele, L., S. Lorch, M. Passarella, D. Small, and R. Titiunik (2017). *An Overview of Geographically Discontinuous Treatment Assignments with an Application to Children's Health Insurance*, Chapter 4, pp. 147–194. Emerald Publishing Limited.

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.

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

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

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

Rasmussen, C. E. and C. K. Williams (2006). *Gaussian processes for machine learning*, Volume 1. MIT press Cambridge.

Rencher, A. C. (2003). *Methods of multivariate analysis*, Volume 492. John Wiley & Sons.

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

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

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

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association 107*(500), 1360–1371.

# Appendix A Covariances for Gaussian Process Model

All covariances below are conditional on the hyperparameters $\boldsymbol{\theta} = (\ell, \sigma_{\text{GP}}, \sigma_\epsilon, \sigma_m)$, omitted for concision.

$$\begin{aligned}
m_T, m_C &\sim \mathcal{N}\big(0, \sigma_m^2\big) \\
\text{Cov}(Y_{iT}, m_T) &= \text{Cov}(Y_{iC}, m_C) = \sigma_m^2 \\
\text{Cov}(Y_{iT}, m_C) &= \text{Cov}(Y_{iC}, m_T) = 0 \\
\text{Cov}(Y_{iT}, f_T(\boldsymbol{s}')) &= \text{Cov}(Y_{iC}, f_C(\boldsymbol{s}')) = k(\boldsymbol{s}_i, \boldsymbol{s}') \\
\text{Cov}(Y_{iT}, f_C(\boldsymbol{s}')) &= \text{Cov}(Y_{iC}, f_T(\boldsymbol{s}')) = 0 \\
\text{Cov}(Y_{iT}, Y_{jT}) &= \text{Cov}(Y_{iC}, Y_{jC}) = \sigma_m^2 + k(\boldsymbol{s}_i, \boldsymbol{s}_j) + \delta_{ij}\sigma_\epsilon^2 \\
\text{Cov}(Y_{iT}, Y_{jC}) &= 0
\end{aligned} \tag{19}$$

We further define some shorthand notation, found in Table 3.

# Appendix B Estimating Linear Regression Coefficients

We present the posterior mean of the linear regression coefficients vector $\boldsymbol{\beta}$ for the model specified in (7).

$$\begin{aligned}
&\text{Cov}(\boldsymbol{Y} \mid \boldsymbol{\beta}) = \begin{bmatrix} \boldsymbol{\Sigma}_{TT} & 0 \\ 0 & \boldsymbol{\Sigma}_{CC} \end{bmatrix}, \ \text{Cov}(\boldsymbol{\beta}) = \sigma_\beta^2 \mathbf{I}_p, \ \text{Cov}(\boldsymbol{Y}, \boldsymbol{\beta}) = \sigma_\beta^2 \boldsymbol{D} \\
&\boldsymbol{\Sigma}_Y = \text{Cov}(\boldsymbol{Y}) = \text{Cov}(\boldsymbol{Y} \mid \boldsymbol{\beta}) + \sigma_\beta^2 \boldsymbol{D}\boldsymbol{D}^\intercal \\
&\hat{\boldsymbol{\beta}} = \mathbb{E}(\beta \mid \boldsymbol{Y}) = \text{Cov}(\boldsymbol{\beta}, \boldsymbol{Y}) \, \text{Cov}(\boldsymbol{Y}, \boldsymbol{Y})^{-1} \boldsymbol{Y} = \sigma_\beta^2 \boldsymbol{D}^\intercal \boldsymbol{\Sigma}_Y^{-1} \boldsymbol{Y}
\end{aligned} \tag{20}$$

The treatment and control residuals can then be obtained as $\boldsymbol{R} = \boldsymbol{Y} - \boldsymbol{D}\hat{\boldsymbol{\beta}}$. Conditionally on $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, the residuals of the treatment and control units then have independent multivariate normal distributions with covariances $\boldsymbol{\Sigma}_{TT}$ and $\boldsymbol{\Sigma}_{CC}$ respectively.

| Symbol | Size | $ij^{\text{th}}$ entry |
|---|---|---|
| $\boldsymbol{K}_{\mathcal{B}\mathcal{B}}$ | $R \times R$ | $\sigma_m^2 + k(\boldsymbol{b}_i, \boldsymbol{b}_j)$ |
| $\boldsymbol{K}_{\mathcal{B}T}$ | $R \times n_T$ | $\sigma_m^2 + k(\boldsymbol{b}_i, \boldsymbol{s}_{jT})$ |
| $\boldsymbol{K}_{\mathcal{B}C}$ | $R \times n_C$ | $\sigma_m^2 + k(\boldsymbol{b}_i, \boldsymbol{s}_{jC})$ |
| $\boldsymbol{K}_{TT}$ | $n_T \times n_T$ | $\sigma_m^2 + k(\boldsymbol{s}_{iC}, \boldsymbol{s}_{jC})$ |
| $\boldsymbol{K}_{CC}$ | $n_C \times n_C$ | $\sigma_m^2 + k(\boldsymbol{s}_{iT}, \boldsymbol{s}_{jT})$ |
| $\boldsymbol{\Sigma}_{TT}$ | $n_T \times n_T$ | $\sigma_m^2 + k(\boldsymbol{s}_{iT}, \boldsymbol{s}_{jT}) + \delta_{ij}\sigma_\epsilon^2$ |
| $\boldsymbol{\Sigma}_{CC}$ | $n_C \times n_C$ | $\sigma_m^2 + k(\boldsymbol{s}_{iC}, \boldsymbol{s}_{jC}) + \delta_{ij}\sigma_\epsilon^2$ |

Table 3: Shorthand notation for covariance matrices. The spatial coordinates of the $i^{\text{th}}$ treatment unit are denoted by $\boldsymbol{s}_{iT}$, and those of the $j^{\text{th}}$ control unit by $\boldsymbol{s}_{jC}$, while $\boldsymbol{b}_i$ denotes the $i^{\text{th}}$ sentinel location along the border.

# Appendix C Calibration of Inverse-variance Test

We seek to obtain a valid hypothesis test against the null hypothesis of zero treatment effect everywhere along the border by using the inverse-variance weighted LATE estimate obtained in Section 3.3 as a test statistic.

Under the parametric null hypothesis $\mathscr{M}_0$, $\boldsymbol{Y}_T$ and $\boldsymbol{Y}_C$ are drawn from a single Gaussian process, with no discontinuity at the border. Their joint covariance is

$$\operatorname{Cov}\left(\begin{pmatrix}\boldsymbol{Y}_T\\ \boldsymbol{Y}_C\end{pmatrix} \mid \mathscr{M}_0\right) = \begin{bmatrix}\boldsymbol{\Sigma}_{TT} & \boldsymbol{K}_{TC}\\ \boldsymbol{K}_{TC}^\intercal & \boldsymbol{\Sigma}_{CC}\end{bmatrix}, \tag{21}$$

where $\boldsymbol{K}_{TC}$ is the $n_T \times n_C$ matrix with $ij^{\text{th}}$ entry equal to $k(\boldsymbol{s}_{iT}, \boldsymbol{s}_{jC})$. The predicted mean outcomes (5) at the sentinels $\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|T}$ and $\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|T}$ are obtained by left-multiplying $\boldsymbol{Y}_T$ and $\boldsymbol{Y}_C$ by matrices $\boldsymbol{W}_T$ and $\boldsymbol{W}_C$ (respectively) that are deterministic functions of the unit locations and the hyperparameters:

$$\boldsymbol{W}_T = \boldsymbol{K}_{\mathcal{B}T}\boldsymbol{\Sigma}_{TT}^{-1} \quad \text{and} \quad \boldsymbol{W}_C = \boldsymbol{K}_{\mathcal{B}C}\boldsymbol{\Sigma}_{CC}^{-1}. \tag{22}$$

Under $\mathscr{M}_0$, the joint distribution of $\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|T}$ and $\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|T}$ is consequently also multivariate normal with mean zero and covariance given by:

$$\operatorname{Cov}\left(\begin{pmatrix}\boldsymbol{W}_T\boldsymbol{Y}_T\\ \boldsymbol{W}_C\boldsymbol{Y}_C\end{pmatrix} \mid \mathscr{M}_0\right) = \begin{bmatrix}\boldsymbol{W}_T\boldsymbol{\Sigma}_{TT}\boldsymbol{W}_T^\intercal & \boldsymbol{W}_T\boldsymbol{K}_{TC}\boldsymbol{W}_C^\intercal\\ \boldsymbol{W}_C\boldsymbol{K}_{TC}^\intercal\boldsymbol{W}_T^\intercal & \boldsymbol{W}_C\boldsymbol{\Sigma}_{CC}\boldsymbol{W}_C^\intercal\end{bmatrix}. \tag{23}$$

Continuing in this fashion, the cliff height (6) estimate $\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y} = \boldsymbol{W}_T\boldsymbol{Y}_T - \boldsymbol{W}_C\boldsymbol{Y}_C$ is yet another zero-mean multivariate normal with covariance given by:

$$\operatorname{Cov}\left(\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y} \mid \mathscr{M}_0\right) = \boldsymbol{W}_T\boldsymbol{\Sigma}_{TT}\boldsymbol{W}_T^\intercal + \boldsymbol{W}_C\boldsymbol{\Sigma}_{CC}\boldsymbol{W}_C^\intercal - \boldsymbol{W}_T\boldsymbol{K}_{TC}\boldsymbol{W}_C^\intercal - \boldsymbol{W}_C\boldsymbol{K}_{TC}^\intercal\boldsymbol{W}_T^\intercal. \tag{24}$$

Weighted LATE estimators of the form defined in (10) are linear transformations of $\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}$ and so under $\mathscr{M}_0$, they are normally distributed with mean zero. For a given weight function $w_{\mathcal{B}}$, its variance is given by

$$\operatorname{var}\left(\mu_{\tau^w|Y} \mid \mathscr{M}_0\right) = \operatorname{Cov}\left(\frac{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^\intercal\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}}{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^\intercal\,\mathbf{1}_R}\right) = \frac{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^\intercal\operatorname{Cov}\left(\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}\right)w_{\mathcal{B}}(\boldsymbol{b}_{1:R})}{\left(w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^\intercal\,\mathbf{1}_R\right)^2}. \tag{25}$$

The $p$-value follows from treating the LATE estimate as a test statistic. Under the null hypothesis, the probability of $\mu_{\tau^w|Y}$ exceeding in magnitude its observed value $\mu_{\tau^w|Y^{obs}}$ is:

$$\mathbb{P}\left(\left|\mu_{\tau^w|Y}\right| \geq \left|\mu_{\tau^w|Y^{obs}}\right| \mid \mathscr{M}_0\right) = 2\Phi\left(-\left|\mu_{\tau^w|Y^{obs}}\right| \Big/ \sqrt{\operatorname{var}(\mu_{\tau^w|Y} \mid \mathscr{M}_0)}\right). \tag{26}$$

The calibrated inverse-variance test of Section 4 is the special case of this procedure with weights $w_{\mathcal{B}}(\boldsymbol{b}_{1:R}) = \Sigma^{-1}_{\boldsymbol{b}_{1:R}|Y}\,\mathbf{1}_R$.

# Supplementary Materials

## A Bayesian Nonparametric Approach to Geographic Regression Discontinuity Designs: Do School Districts Affect NYC House Prices?

## S-1 Spatial Confounding of Projected 1D RDD

Analysing GeoRDDs by using the signed distance from the border as a forcing variable in a 1D RDD can lead to spatial confounding. We demonstrate this with a simple artificial example, depicted in Figure S-1. Suppose we have units in a 2D square, with spatial coordinates $\boldsymbol{s}_1 \in [0, 2]$, and $\boldsymbol{s}_2 \in [-1, 1]$, and with a straight border at $\boldsymbol{s}_2 = 0$ separating a treatment region from a control region. Let us impose the null hypothesis, with outcomes driven only by a linear spatial trend running parallel to the border:

$$Y_i = \alpha \boldsymbol{s}_{1i} + \epsilon_i \,, \quad \epsilon_i \sim \mathcal{N}\big(0, \sigma_\epsilon^2\big) \,. \tag{S-1}$$

Let us consider the situation where the density $\rho(\boldsymbol{s})$ of units is different in each quadrant of the square:

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

The projection RDD then considers a 1D RDD along $\boldsymbol{s}_2$. The usual RDD estimand (1) can be obtained analytically:

$$\tau = \frac{\int_0^1 2\rho\alpha\boldsymbol{s}_1 \, d\boldsymbol{s}_1 + \int_1^2 \rho\alpha\boldsymbol{s}_1 \, d\boldsymbol{s}_1}{\int_0^1 2\rho \, d\boldsymbol{s}_1 + \int_1^2 \rho \, d\boldsymbol{s}_1} - \frac{\int_0^1 \rho\alpha\boldsymbol{s}_1 \, d\boldsymbol{s}_1 + \int_1^2 2\rho\alpha\boldsymbol{s}_1 \, d\boldsymbol{s}_1}{\int_{-1}^0 \rho \, d\boldsymbol{s}_1 + \int_0^1 2\rho \, d\boldsymbol{s}_1} = \frac{-\alpha}{3} \,, \tag{S-3}$$

and is non-zero even though the treatment effect is zero everywhere along the border. This is because $\boldsymbol{s}_1$ acts as a hidden confounder whose distribution changes discontinuously at the border, which leads to bias and inconsistency in the projected 1D 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 body of water, giving zero population density on one side of the border. A visual inspection of Figure 1 showing the locations of units in a New York City property sales dataset reveals examples of this.

## S-2 Additional LATE Estimands and Simulations

In the paper, we presented and characterized four choices of local average treatment effect (LATE) estimands (and corresponding estimators): the uniformly-weighted $\tau^{\text{UNIF}}$, population-weighted $\tau^{\rho}$, inverse-weighted $\tau^{\text{INV}}$, and finite population projected $\tau^{\text{PROJ}}$. We here present two other estimands of interest not directly presented in the paper, which extend the pro-

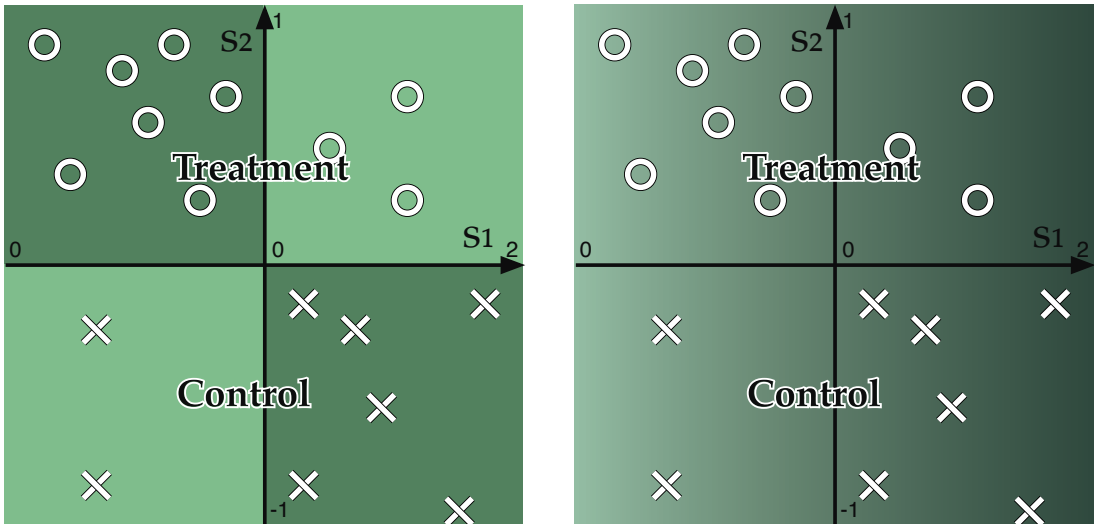


Figure S-1: A theoretical example illustrating the susceptibility of the projected 1D RDD method to spatial confounding. On the left, the background shows the density of units, while on the right it shows the linear spatial trend in outcomes. The locations of treatment and control units are shown with circles and crosses respectively, separated by a border at $\boldsymbol{s}_2 = 0$. Notice how treatment units are densest in an area with low outcomes, while control units units are densist in an area with high outcomes.

jection population idea of $\tau^{\text{PROJ}}$ to superpopulations. We then compare all the estimators in a simulated demonstration to illustrate their differences.

## S-2.1 Projected Land LATE

In certain applications, population-based estimands can be undesirable, especially if the locations at which measurements are made are not representative of the population of interest. In such cases, geography-weighted estimands can be more natural. See Antonelli et al. (2016) for a discussion of this distinction in the context of preferential sampling. Remember that the "geometry-based" estimand $\tau^{\text{UNIF}}$ places uniform weights along the border. Instead, the "geography-based" projected land LATE estimand $\tau^{\text{GEO}}$, illustrated in Figure S-2(b), begins by placing uniform weights on the treatment and control areas $\mathcal{A}_T$ and $\mathcal{A}_C$ that are within distance $\Delta$ of the border $\mathcal{B}$, but then projects them onto the border to derive border weights. In other words, the projection method from $\tau^{\text{PROJ}}$ is applied to an infinite population of uniform density on both sides of the border, instead of the finite population of observed units.

We denote the border vicinity area by $\mathcal{A}_\Delta$, defined as all points $\boldsymbol{s}$ such that $\boldsymbol{s} \in \mathcal{A}_T \cup \mathcal{A}_C$, and $\text{dist}_{\mathcal{B}}(\boldsymbol{s}) < \Delta$. To estimate $\tau^{\text{GEO}}$, a tight grid $G^\nu$ of evenly spaced points separated by $\nu$ is first generated covering $\mathcal{A}_\Delta$. Denote the number of grid points by $L_\nu$. Each point $G^\nu_l$, $l = 1, \ldots, L_\nu$ in $G^\nu$ is then projected onto the border to become a sentinel. The treatment effect at these positions is then estimated as before, yielding a mean vector and covariance matrix akin to (6). The mean of the mean vector then gives an estimate of $\tau^{\text{GEO}}$. In other words, $\tau^{\text{GEO}}$ is estimated by applying the $\tau^{\text{UNIF}}$ procedure with sentinels obtained by

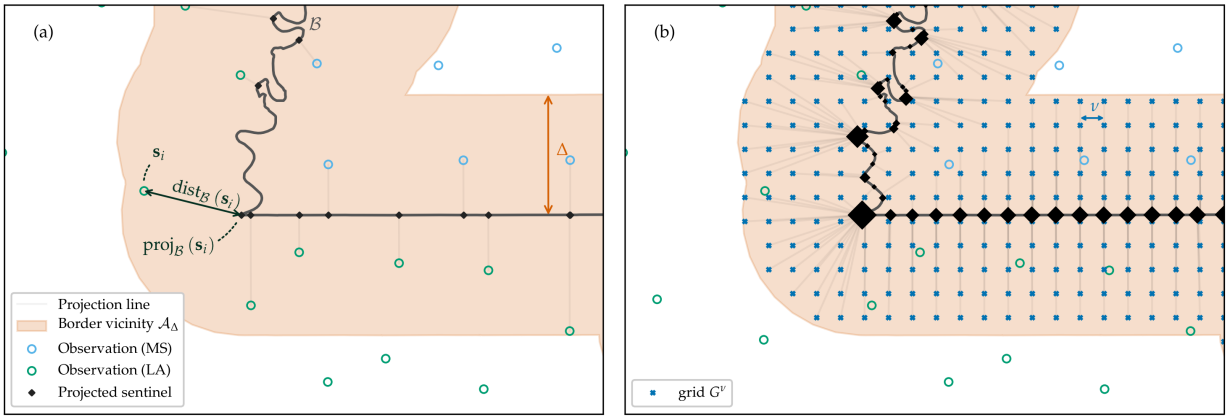


Figure S-2: Illustration of (a) projected finite-population LATE $\tau^{\text{PROJ}}$, and (b) projected land LATE $\tau^{\text{GEO}}$, using the border separating Mississippi and Louisiana near Baton Rouge, with units at the centroid of each county. The border vicinity $\mathcal{A}_\Delta$ is defined as all land within $\Delta = 50\,\text{km}$ of the border. With both methods, every projected sentinel has equal weight in the LATE, but the tight grid in (b) causes sentinels to coincide or nearly coincide, which we depict by scaling up the size of the marker by the number of coinciding sentinels.

projecting the grid points, instead of equispaced sentinels. $\tau^{\text{GEO}}$ remains in the category of weighted-mean estimands, with the weight function $w_{\mathcal{B}}(\boldsymbol{b})$ in (9) proportional to the area of $\mathcal{A}_T$ and $\mathcal{A}_C$ that $\boldsymbol{b}$ is nearest to, which can be written as the limit as the grid spacing goes to zero of point masses at the grid locations projected onto the border:

$$w_{\mathcal{B}}(\boldsymbol{b}) = \lim_{\nu \to 0} \frac{1}{L_\nu} \sum_{l=1}^{L_\nu} \delta(\boldsymbol{b} - \text{proj}_{\mathcal{B}}(G_l^\nu)) \,. \tag{S-4}$$

For certain applications, it may be desirable to further restrict $\mathcal{A}_\Delta$ to only certain types of land, for example residential areas in social studies, or farmland in agricultural studies. It is important to note that $\tau^{\text{GEO}}$ is never interpretable as the average treatment effect in the vicinity of the border, that is $\tau^{\text{GEO}} \neq \int_{\mathcal{A}_\Delta} \tau(\boldsymbol{s})\, d\boldsymbol{s}$. Estimating the latter estimand would require predicting the conditional regression function at grid locations within the treatment or control region using only observations on the *other* side of the border, which increases the extent of extrapolation required and thus makes the analysis more vulnerable to model misspecification.

## S-2.2 Projected Super-Population LATE

The purely geographical estimand $\tau^{\text{GEO}}$ can be modified by weighting the grid points $G_l^\nu$, $l = 1, \ldots, L_\Delta$ by the population density $\rho(G_l^\nu)$. This gives the projected superpopulation LATE $\tau^{\text{POP}}$. Similarly to the density-weighted LATE $\tau^\rho$, estimating $\tau^{\text{POP}}$ requires an estimate of the density $\rho(G_l^\nu)$ at every grid point. As before, the uncertainty in the estimate of $\rho$ should in principle 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.

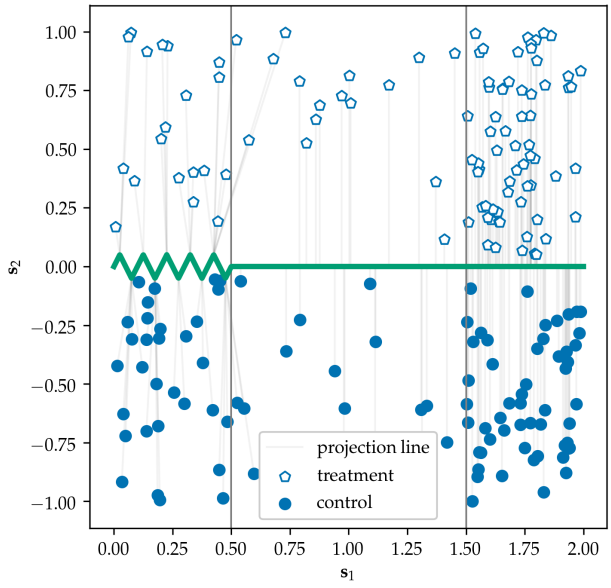


Figure S-3: Spatial positions of units and border for the wiggly border simulation of Section S-2.3. Projection lines for the projected finite population LATE are shown in light gray.

The estimand $\tau^{\text{POP}}$ can be interpreted as giving equal weight to each unit in the superpopulation of units within the border vicinity $\mathcal{A}_\Delta$, but then moving each unit from its original location to the nearest point on the border (where the GeoRDD is best able to estimate the treatment effect without undue extrapolation) and then averaging the treatment effect of each unit in this displaced superpopulation. The resulting weight function is:

$$w_{\mathcal{B}}(\boldsymbol{b}) = \lim_{\nu \to 0} \frac{1}{L_\nu} \sum_{l=1}^{L_\nu} \delta(\boldsymbol{b} - \text{proj}_{\mathcal{B}}(G_l^\nu))\rho(G_l^\nu)\,. \tag{S-5}$$

## S-2.3 Wiggly Border Simulation

We illustrate all of our LATE estimators with a simulation. 200 units are placed in a square area delimited by spatial coordinates $\boldsymbol{s}_1 \in \{0, 2\}$ and $\boldsymbol{s}_2 \in \{-1, 1\}$. A border at $\boldsymbol{s}_2 = 0$ divides units vertically into a control and treatment region, which are then further divided horizontally at $\boldsymbol{s}_1 = 0.5$ and $\boldsymbol{s}_1 = 1.5$ into three blocks:

- The leftmost block $\boldsymbol{s}_1 < 0.5$ has a weak treatment effect and density defined to be equal to $\rho = 1$.
- The middle block $0.5 \geq \boldsymbol{s}_1 < 1.5$ has a much lower population density $\rho = 0.3$, and a stronger treatment effect.
- The rightmost block $\boldsymbol{s}_1 \geq 1.5$, has a much higher population density $\rho = 2$, and a very strong treatment effect.

| | Left $\boldsymbol{s}_1 < 0.5$ | Middle $0.5 \geq \boldsymbol{s}_1 < 1.5$ | Right $1.5 \geq \boldsymbol{s}_1$ |
|---|---|---|---|
| **Border** | wiggly | straight | straight |
| **Density** | $\rho = 1$ | very low $\rho = 0.3$ | high $\rho = 2$ |
| $\boldsymbol{\tau}$ | weak | medium | strong |

Table S-1: Summary of wiggly border simulation setup.

Furthermore, the border in the leftmost block 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 S-1. We draw a single set of spatial coordinates, shown in Figure S-3, then draw 10,000 simulations of the outcomes $Y$ from a Gaussian process with squared exponential kernel ($\ell = 0.4$, $\sigma = 0.5$). To units above the border we add a treatment effect $\tau(\boldsymbol{s}_1, \boldsymbol{s}_2) = \boldsymbol{s}_1$.

We fit the Gaussian process model (3), using the known hyperparameters of the covariance kernel and a weak prior on the mean parameter of each region, and estimate the LATE using the six methods proposed above. For projection-based methods, the buffer distance $\Delta$ is infinite, so all of $\mathcal{A}$ is included. In Figure S-4(a) we show, for each estimator, the corresponding estimand and average posterior mean estimate evolving as the number of border wiggles increases. The behavior of the posterior standard deviation is shown in Figure S-4(b). The simulations results can also be found in Table S-2.

When the border is a straight line and because $\mathcal{A}_T$ and $\mathcal{A}_C$ are rectangles, the density-weighted estimand $\tau^\rho$ equals the projected superpopulation estimand $\tau^{\text{POP}}$. They are in fact both equal to the infinite-population average treatment effect since the treatment effect does not depend on $\boldsymbol{s}_2$. Correspondingly, the posteriors of $\tau^\rho$ and $\tau^{\text{POP}}$ are identical. $\tau^{\text{POP}}$ and the finite-population projected LATE $\tau^{\text{PROJ}}$ are also similar, but the latter has the advantage of not requiring local estimates of the population density.

The geometry- and geography-based LATE $\tau^{\text{UNIF}}$ 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 exactly coincide with any others. The estimand and mean estimate change slightly from 0 to 1 wiggles, but remains stable thereafter, demonstrating the robustness of this estimator to border topology. Weighting by the inverse variance gives the lowest posterior variance within the class of LATEs under consideration, which can indeed be seen in Figure S-4(b).

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.

In Figure S-5(a-f), we illustrate the behavior of border weights $w_{\mathcal{B}}(\boldsymbol{b})$ and unit weights ($\boldsymbol{w}_T$ and $\boldsymbol{w}_C$) in this simulation setting with 3 wiggles. Note how evenly spaced sentinels (for $\tau^{\text{UNIF}}$, $\tau^\rho$, and $\tau^{\text{INV}}$) are more densely packed along $\boldsymbol{s}_1$ in the leftmost area because

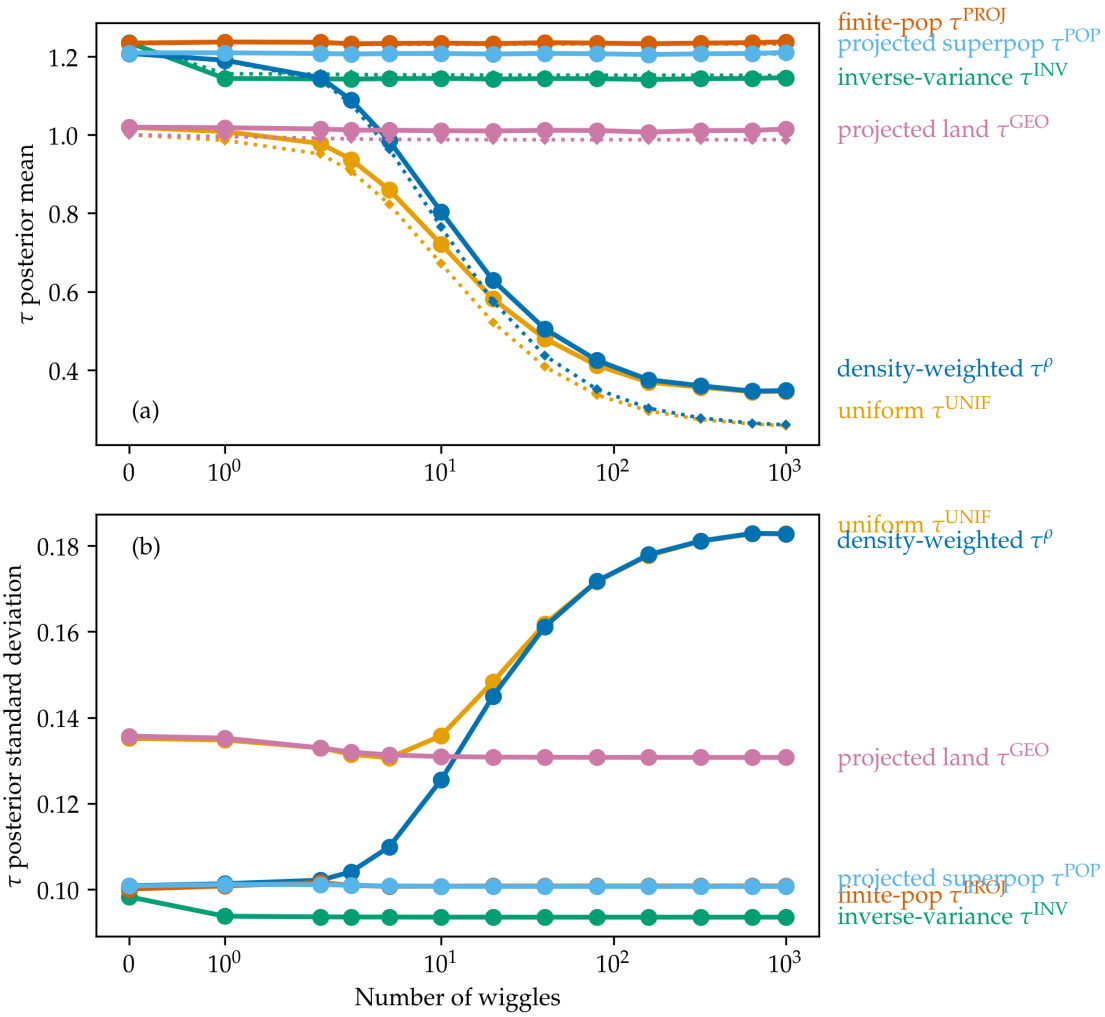


Figure S-4: Results of the simulations of Section S-2.3, showing for each LATE estimator as the leftmost section of the border gets wigglier (a) the estimate (posterior mean) averaged over 10,000 simulations with the corresponding estimand shown as a dotted line of the same color, and (b) the posterior standard deviation.

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 for $\tau^{\text{INV}}$ are well-behaved, and in fact quite similar to those of the projected finite- and infinite-population estimators. Furthermore, note that all estimators can give some small negative weights $\boldsymbol{w}_T$ to treatment units, and small positive weights $\boldsymbol{w}_C$ to control units. For Gaussian processes, this can be understood in terms of the negative side-lobes of the equivalent kernel (see Rasmussen and Williams (2006) Section 2.6). The high variance of $\tau^{\text{UNIF}}$ and $\tau^{\text{GEO}}$ 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.

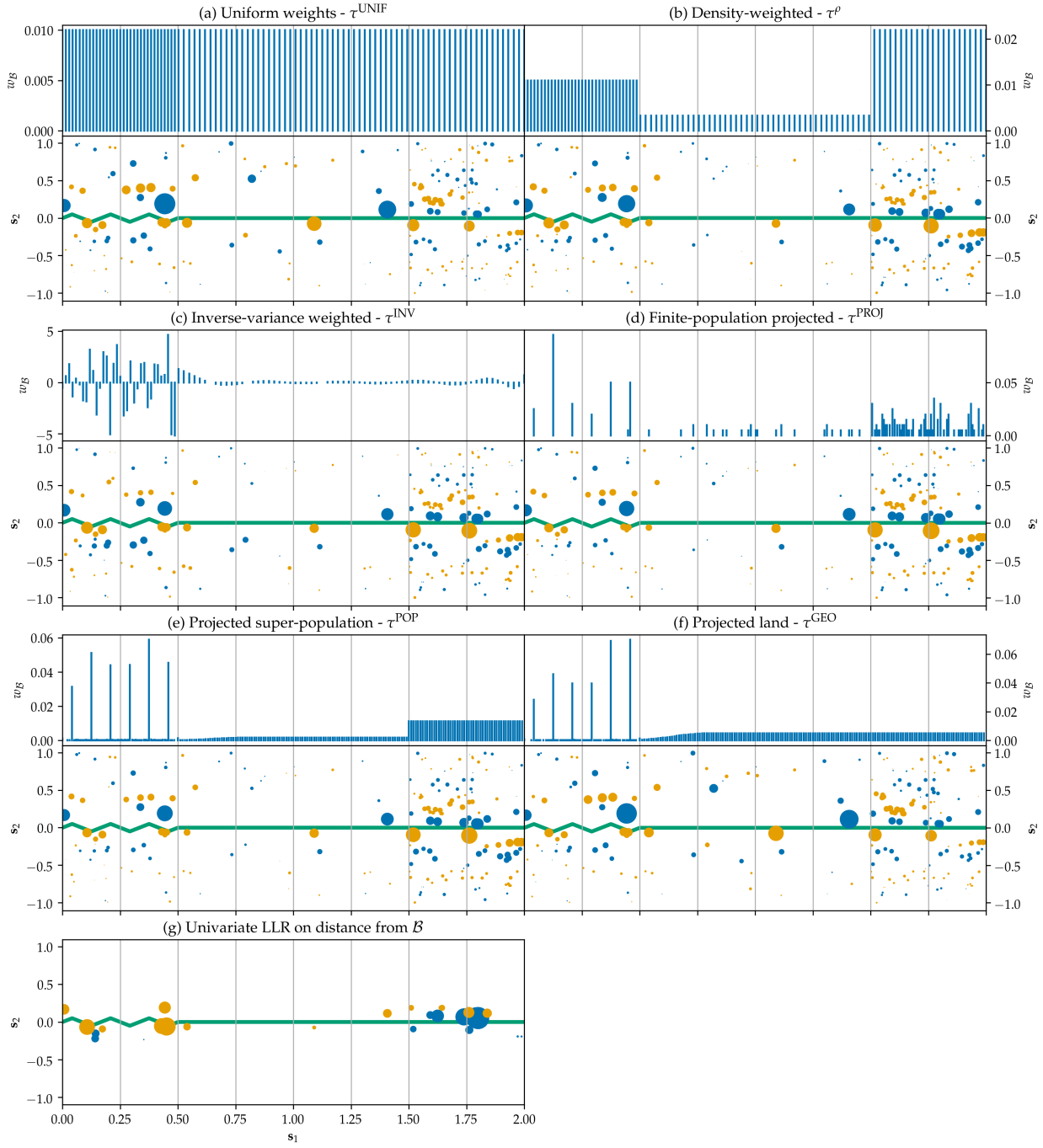


Figure S-5: Weight functions and induced weights on the observations for the six weight functions proposed in this paper. The weight function plots show the weight $w_{\mathcal{B}}(\boldsymbol{b})$ against each sentinel's $\boldsymbol{s}_1$ coordinate. Sentinels with coinciding or nearly coinciding (within 0.005 of each other) coordinate $\boldsymbol{s}_1$ 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 ($\boldsymbol{w}_T$ and $\boldsymbol{w}_C$ for treatment and control units respectively), and colored in blue for positive weights and orange for negative weights.

For comparison, the weights placed on units by the projected 1D RDD are shown in Figure S-5(g). A triangular kernel in $\boldsymbol{s}_2$ was used with bandwidth selected using the MSE-minimizing method proposed by Imbens and Kalyanaraman (2012). The Projected 1D RDD estimator can also be written as a linear combination of the observed outcomes (11), and the unit weight vectors can be derived as:

$$\boldsymbol{w}_T = \boldsymbol{X}_b(\boldsymbol{X}_T^\intercal \boldsymbol{W}_T \boldsymbol{X}_T)^{-1}\boldsymbol{X}_T^\intercal \boldsymbol{W}_T \quad \text{and} \quad \boldsymbol{w}_C = -\boldsymbol{X}_b(\boldsymbol{X}_C^\intercal \boldsymbol{W}_C \boldsymbol{X}_C)^{-1}\boldsymbol{X}_C^\intercal \boldsymbol{W}_C\,, \tag{S-6}$$

where $\boldsymbol{X}_b = (1\ 0)$, $\boldsymbol{X}_T$ is the $n_T \times 2$ design matrix with the first column filled with ones and the second column containing the distance from the border of each treatment unit, and $\boldsymbol{W}_T$ is an $n_T \times n_T$ diagonal matrix where the $i^{\text{th}}$ diagonal element is the triangular kernel evaluated on the $i^{\text{th}}$ unit's distance from the border. The $\boldsymbol{X}_C$ and $\boldsymbol{W}_C$ matrices are analogously defined for control units. By construction, the unit weights drop to zero outside of the support of the kernel. Within the support, Projected 1D RDD can also give negative weights to treatment units, and positive weights to control units. This results from the negative influence on the prediction $\widehat{y^*}$ at $x^*$ that univariate linear regression can give to an observation $Y_i$ at $X_i$ sufficiently far away on the opposite side of the mean $\overline{X}$ of all observations. 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 $\boldsymbol{s}_1$ would confound the estimated treatment effect.

| $n_{\text{wiggles}}$ | $\widehat{\tau^{\text{UNIF}}}$ | $\tau^{\text{UNIF}}$ | $\widehat{\tau^{\text{INV}}}$ | $\tau^{\text{INV}}$ | $\widehat{\tau^{\rho}}$ | $\tau^{\rho}$ | $\widehat{\tau^{\text{PROJ}}}$ | $\tau^{\text{PROJ}}$ | $\widehat{\tau^{\text{GEO}}}$ | $\tau^{\text{GEO}}$ | $\widehat{\tau^{\text{POP}}}$ | $\tau^{\text{POP}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.02 (0.14) | 1.00 | 1.24 (0.10) | 1.23 | 1.21 (0.10) | 1.21 | 1.23 (0.10) | 1.24 | 1.02 (0.14) | 1.00 | 1.21 (0.10) | 1.21 |
| 1 | 1.01 (0.13) | 0.99 | 1.14 (0.09) | 1.16 | 1.19 (0.10) | 1.19 | 1.24 (0.10) | 1.24 | 1.02 (0.14) | 1.00 | 1.21 (0.10) | 1.21 |
| 2 | 0.98 (0.13) | 0.95 | 1.14 (0.09) | 1.15 | 1.15 (0.10) | 1.14 | 1.24 (0.10) | 1.24 | 1.01 (0.13) | 0.99 | 1.21 (0.10) | 1.21 |
| 3 | 0.94 (0.13) | 0.91 | 1.14 (0.09) | 1.15 | 1.09 (0.10) | 1.08 | 1.23 (0.10) | 1.23 | 1.01 (0.13) | 0.99 | 1.21 (0.10) | 1.21 |
| 5 | 0.86 (0.13) | 0.82 | 1.14 (0.09) | 1.15 | 0.98 (0.11) | 0.96 | 1.23 (0.10) | 1.23 | 1.01 (0.13) | 0.99 | 1.21 (0.10) | 1.21 |
| 10 | 0.72 (0.14) | 0.67 | 1.14 (0.09) | 1.15 | 0.80 (0.13) | 0.76 | 1.23 (0.10) | 1.23 | 1.01 (0.13) | 0.99 | 1.21 (0.10) | 1.21 |
| 20 | 0.58 (0.15) | 0.52 | 1.14 (0.09) | 1.15 | 0.63 (0.14) | 0.58 | 1.23 (0.10) | 1.23 | 1.01 (0.13) | 0.99 | 1.21 (0.10) | 1.21 |
| 40 | 0.48 (0.16) | 0.41 | 1.14 (0.09) | 1.15 | 0.50 (0.16) | 0.44 | 1.23 (0.10) | 1.23 | 1.01 (0.13) | 0.99 | 1.21 (0.10) | 1.21 |
| 80 | 0.41 (0.17) | 0.34 | 1.14 (0.09) | 1.15 | 0.42 (0.17) | 0.35 | 1.23 (0.10) | 1.23 | 1.01 (0.13) | 0.99 | 1.21 (0.10) | 1.21 |
| 160 | 0.37 (0.18) | 0.30 | 1.14 (0.09) | 1.15 | 0.38 (0.18) | 0.30 | 1.23 (0.10) | 1.23 | 1.01 (0.13) | 0.99 | 1.20 (0.10) | 1.21 |
| 320 | 0.36 (0.18) | 0.27 | 1.14 (0.09) | 1.15 | 0.36 (0.18) | 0.28 | 1.23 (0.10) | 1.23 | 1.01 (0.13) | 0.99 | 1.21 (0.10) | 1.21 |
| 640 | 0.34 (0.18) | 0.26 | 1.14 (0.09) | 1.15 | 0.35 (0.18) | 0.26 | 1.23 (0.10) | 1.23 | 1.01 (0.13) | 0.99 | 1.21 (0.10) | 1.21 |
| 1000 | 0.35 (0.18) | 0.26 | 1.15 (0.09) | 1.15 | 0.35 (0.18) | 0.26 | 1.24 (0.10) | 1.23 | 1.01 (0.13) | 0.99 | 1.21 (0.10) | 1.21 |

Table S-2: **Wiggly Border Simulation Results.** Posterior mean averaged over 10,000 simulations, posterior standard deviation and true value for each LATE estimand as the wiggliness of the border is increased in the simulations of Section S-2.3.

# S-3 Alternate Tests for Non-Zero Treatment Effect

In our main paper, we present the calibrated inverse-variance test, which targets the weak null hypothesis of zero LATE. It can be generalized to any other choice of LATE estimand defined as a weighted mean over the border, as in (9). Tests of the sharp null hypothesis, that is $\tau(\boldsymbol{b}) = 0$ for all $\boldsymbol{b} \in \mathcal{B}$, are also of interest, and we present two such tests in this section. We provide a simulation comparing the power of the three tests when the treatment effect is constant, and apply each test to the NYC school district application of the main paper. We advocate for the use of the calibrated inverse-variance test in most situations, as it has demonstrated higher power and robustness to model misspecification than the sharp null tests.

## S-3.1 Marginal Likelihood Test

Recall the null model $\mathscr{M}_0$ defined in Section 4; the unified Gaussian process 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 $\mathscr{M}_0$ than under $\mathscr{M}_1$, the $\mathcal{GP}$ model as specified in (3). We therefore choose the difference in log-likelihoods as our test statistic

$$t = \log \mathbb{P}(\boldsymbol{Y} \mid \mathscr{M}_1) - \log \mathbb{P}(\boldsymbol{Y} \mid \mathscr{M}_0)\,, \tag{S-7}$$

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 $B$ bootstrap samples $\boldsymbol{Y}^{(b)}$ from $\mathscr{M}_0$, using the same spatial locations as the original data, and then fit the two competing models to the simulated data in order to obtain the bootstrapped test statistic

$$t^{(b)} = \log \mathbb{P}\big(\boldsymbol{Y}^{(b)} \mid \mathscr{M}_1\big) - \log \mathbb{P}\big(\boldsymbol{Y}^{(b)} \mid \mathscr{M}_0\big)\,. \tag{S-8}$$

Repeating this procedure, we obtain a distribution of $t$ under $\mathscr{M}_0$, which we can then compare to the observed $t$. More precisely, the proportion of $t^{(b)}$ drawn above $t_{obs}$ estimates the $p$-value:

$$p = \mathbb{P}(t > t_{obs} \mid \mathscr{M}_0) \approx \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\big\{t^{(b)} > t_{obs}\big\}\,. \tag{S-9}$$

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.

## S-3.2 "Chi-squared" Test

The likelihood-based sharp null above is valid and easy to understand. But it may seem odd that the test aims to detect a non-zero treatment effect at the border, without any explicit reference to the border $\mathcal{B}$. The test statistic and $p$-values can be computed without

access to the sentinel positions, using only the treatment and control indicators. If the test is significant, there is no guarantee that this is due to a discontinuity at the border.

To address this oddity, we can derive a test statistic directly from the cliff height estimator (6). We use $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as shorthand for the posterior mean $\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{b}_{1:R}|Y}$ throughout this section. If a $k$-vector $\boldsymbol{y}$ is distributed $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean vector $\boldsymbol{\mu}$ unknown and covariance $\boldsymbol{\Sigma}$ known, then under the null hypothesis that $\mu = 0$, the test statistic $\boldsymbol{y}^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{y}$ has distribution $\chi^2_k$. See for example Rencher (2003) Section 5.2.2 for a classical derivation of this test. This suggests that we could use $S^2 = \boldsymbol{\mu}^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$ as a test statistic, and obtain a $p$-value from a $\chi^2_R$ distribution function evaluated at $S^2$, where $R$ is the number of sentinels. However, we face two problems. Firstly, this test, obtained heuristically from a Bayesian posterior by analogy with the classical multivariate normal result, is not a valid frequentist test. Secondly, while $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Sigma}$ 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 always conservative in their simulations. 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 trust its validity.

We therefore again take the parametric bootstrap approach, this time using $S^2$ as the test statistic. With B bootstrap samples, the $p$-value is estimated as

$$p \approx \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\{S^2_{(b)} < S^2\} \quad \text{with} \quad S^2_{(b)} = (\boldsymbol{\mu}_{(b)})^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{(b)}\,, \tag{S-10}$$

where $\boldsymbol{\mu}_{(b)}$ is the result of applying the cliff height estimator (6) to the bootstrap sample $\boldsymbol{Y}^{(b)}$.

Because calculating $S^2$ involves inverting a matrix $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Sigma}$ 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.

### S-3.3 Comparing Power of Tests in Simulated Example

The three tests we developed leverage different aspects of the design, 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 S-6. We simulate outcomes from a single Gaussian process covering both states. For simplicity, we fix the hyperparameters to arbitrary values: $\sigma_\epsilon = \sigma_{\text{GP}} = 1.0$ and $\ell = 100\,\text{km}$. We then add a constant treatment effect $\tau$ to all the outcomes in Louisiana. The results of the three tests proposed so far are shown in the first three rows of Table S-3 for $\tau = 0$ (null hypothesis) and $\tau = 1.2$ and significance level $\alpha = 0.05$.

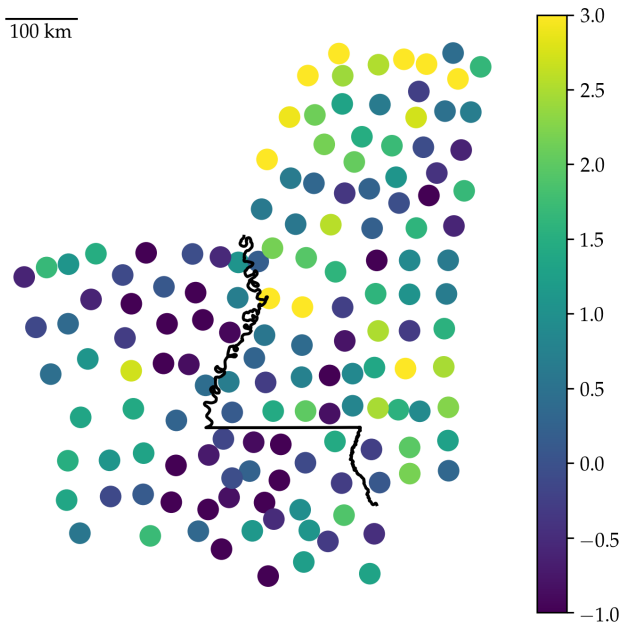


Figure S-6: Set-up of the imaginary experiment in Louisiana and Mississippi. Each unit is at the centroid of a county. The colors indicated the observed outcomes in one draw of the simulation under $\tau = 1.5$. In this particular run, the $p$-values were 0.0016, 0.0018, and 0.0013 for the mLL, $\chi^2$, and inverse-variance test respectively.

We see that under the null, the $\chi^2$ 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 9% 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.

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

## S-3.4 Additional Tests for NYC School Districts Application

We now compare the results of the three hypothesis tests applied to the NYC house prices application. The three $p$-values are provided in Table S-4, and show agreement between the

| Test | Power under $\tau = 0$ | Power under $\tau = 1.2$ |
|---|---|---|
| Marginal log-likelihood bootstrap | 0.05 | 0.72 |
| $\chi^2$ bootstrap | 0.05 | 0.63 |
| $\tau^{\text{INV}}$ uncalibrated | 0.09 | 0.87 |
| $\tau^{\text{INV}}$ calibrated | 0.05 | 0.80 |

Table S-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.

| Test | $p$-value |
|---|---|
| Marginal log-likelihood bootstrap | 0.003 |
| $\chi^2$ bootstrap | 0.022 |
| $\tau^{\text{INV}}$ uncalibrated | 0.0007 |
| $\tau^{\text{INV}}$ calibrated | 0.002 |

Table S-4: Results of hypothesis tests for New York school district house prices. The marginal log-likelihood and $\chi^2$ test were both performed with 10,000 bootstrap samples.

three tests, though the $\chi^2$ test returns a considerably higher $p$-value, which is unsurprising considering the lower power of this test seen in simulations.

To assess the validity of the three tests, we apply the placebo tests devised in Section 4.1. Within each district, we split the data in half by a line at angles $1°$, $3°$, $5°$, $6°$, ..., $179°$. 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 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 concern in spatial models, makes the interpretation of the mLL test unreliable. Based on this vulnerability, and its manifestation in this application, we do not recommend relying on the likelihood-ratio test.

The $\chi^2$ test shows more robustness, with Figure S-7(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 $\chi^2$ 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 calibrated inverse-variance placebo $p$-values display no obvious bias, with Figure S-7(f) close to uniformly distributed, and Figure S-7(e) showing a lower auto-correlation than the mLL and $\chi^2$ tests. The high power and robustness of the inverse-variance test make a strong case for its use in most applications.

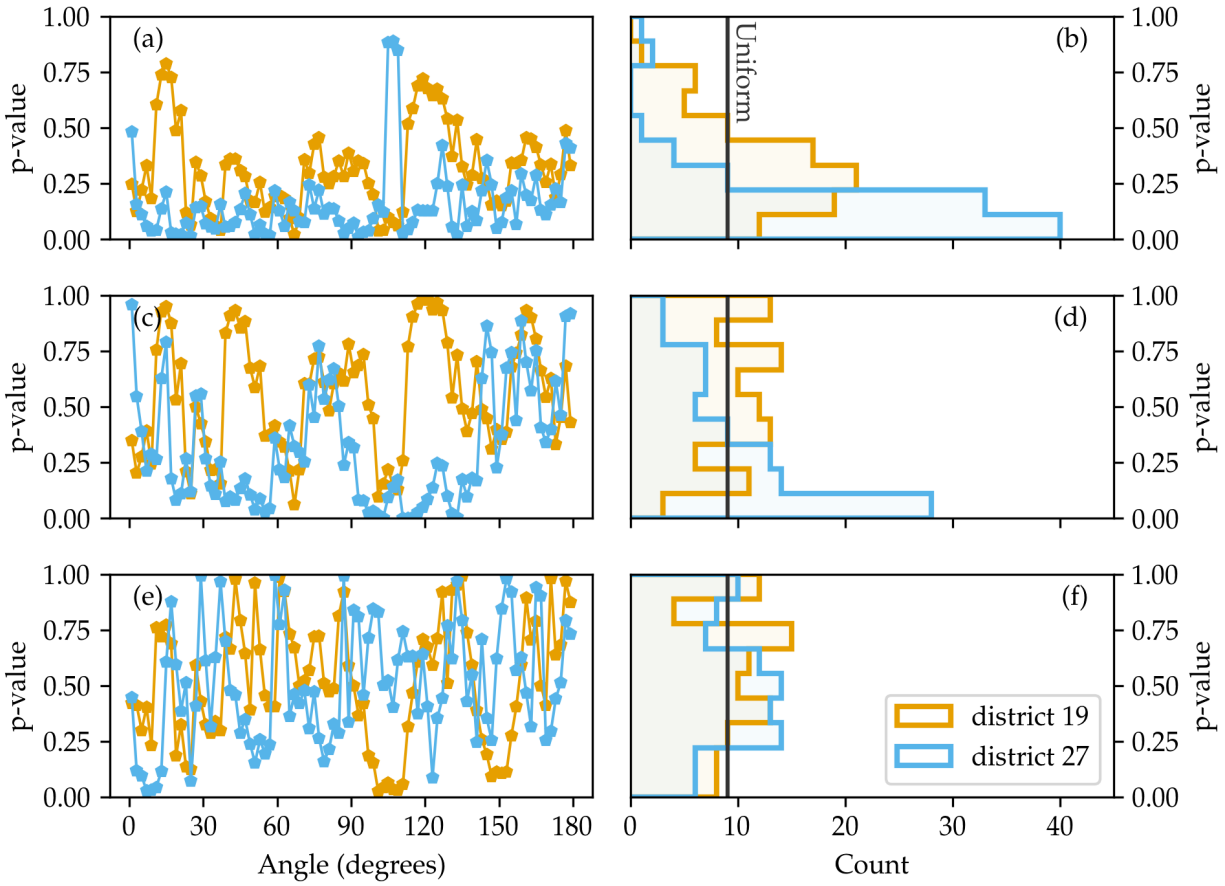


Figure S-7: Placebo tests for significance tests applied to NYC school district house prices, applied within districts 19 and 27. From top to bottom, results are shown for the marginal log-likelihood bootstrap test, chi-squared bootrap test, and calibrated inverse-variance test. The first column shows the placebo $p$-value as a function of the border angle; the second column shows histograms of the placebo $p$-values, with the black vertical line indicating the uniform distribution.

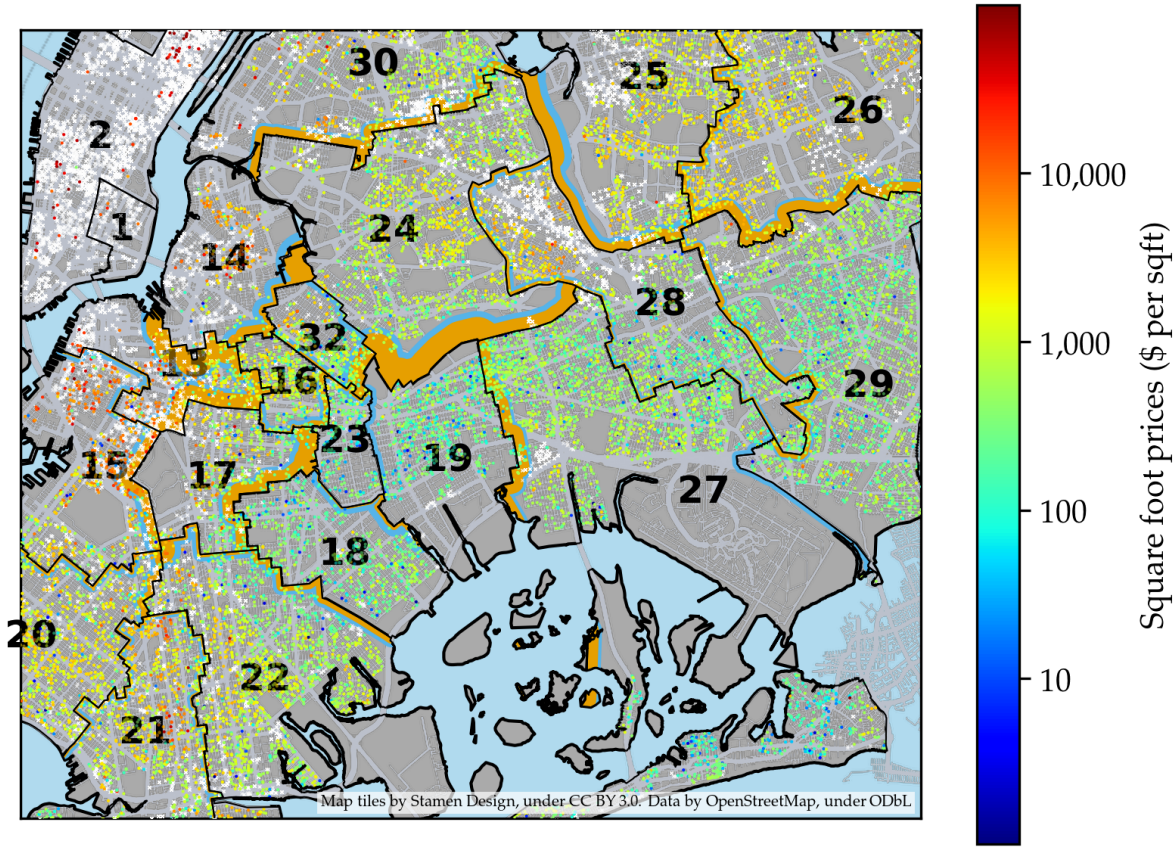


Figure S-8: Pairwise estimates of the inverse variance LATE between adjacent districts. The thickness of the orange buffer adjacent to borders is proportional to the posterior mean of the inverse variance LATE, and the blue buffer beyond it is proportional to the posterior standard deviation of the LATE. The buffers are drawn on the side of the border that is estimated to have higher house prices.

## S-4 Full Analysis: NYC School Districts

The GeoRDD analysis can be repeated for each pair of adjacent districts. Figure S-8 and Table S-5 give an overview of the results by showing the posterior mean and standard deviation of the inverse variance LATE estimated at each border. Significant effects are found between many districts, but interpreting the results requires some caution. We have already mentioned the issue of compound treatments for borders between school districts that overlap with the border between boroughs. School districts 19, 32, and 14 are in Brooklyn, while districts 30, 24, and 27 are in Queens.

Some school districts are separated by parks (or other non-residential zones), for example districts 15 & 17 or 19 & 24, so that house sales do not extend all the way to the border on one or both sides. A significant treatment effect between these pairs cannot be interpreted as the detection of a discontinuity in prices at the border, let alone any kind of causal interpretation, but rather it means that the difference in prices between the two sides of the park exceeds

the typical spatial variation of house prices expected over the same distance. This is not unsurprising, and one may speculate that physical barriers like parks, rivers, railways and major roads can separate neighborhoods with distinct character, demographics and thus house prices. This in turn challenges the stationarity assumption of the spatial model (3). The higher distance between data and the border also stretches the spatial model's ability to extrapolate, which makes it more vulnerable to model misspecification.

Other pairs of district, like 13 & 14, 13 & 17, and 25 & 28 have clusters of missing data (condo sales with unknown square footage) near the border that cast doubt on the interpretation of the estimated effect. Nonetheless, significant effects are also found between pairs of school districts without issues due to compound treatments, physical barriers, or missing data. House prices increase going across the border from districts 16 to 13, 18 to 17, 24 to 30, 23 to 17, 25 to 26, 28 to 29, and 29 to 26. Overall, it seems that school district borders in Brooklyn and Queens can correspond to measurable jumps in house prices per square foot. The estimated size of this effect varies: zero or negligible in some cases, such as between districts 15, 20, 21, and 22; and quite pronounced in others, such as a 20% price increase from 29 to 26, or 22% from 18 to 17.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **13** | **14** : $-0.29 \pm 0.09$ | **15** : $+0.03 \pm 0.07$ | **16** : $+0.13 \pm 0.07$ | **17** : $+0.26 \pm 0.08$ | | | |
| **14** | **13** : $-0.29 \pm 0.09$ | **16** : $+0.16 \pm 0.10$ | **24** : $+0.38 \pm 0.15$ | **32** : $+0.07 \pm 0.12$ | | | |
| **15** | **13** : $+0.03 \pm 0.07$ | **17** : $+0.18 \pm 0.10$ | **20** : $-0.05 \pm 0.06$ | **22** : $-0.28 \pm 0.11$ | | | |
| **16** | **13** : $+0.13 \pm 0.07$ | **14** : $+0.16 \pm 0.10$ | **17** : $-0.04 \pm 0.07$ | **23** : $-0.10 \pm 0.07$ | **32** : $-0.05 \pm 0.06$ | | |
| **17** | **13** : $+0.26 \pm 0.08$ | **15** : $+0.18 \pm 0.10$ | **16** : $-0.04 \pm 0.07$ | **18** : $+0.20 \pm 0.07$ | **22** : $+0.06 \pm 0.07$ | **23** : $-0.29 \pm 0.10$ | |
| **18** | **17** : $+0.20 \pm 0.07$ | **19** : $-0.06 \pm 0.12$ | **22** : $+0.10 \pm 0.07$ | **23** : $-0.03 \pm 0.09$ | | | |
| **19** | **18** : $-0.06 \pm 0.12$ | **23** : $-0.00 \pm 0.08$ | **24** : $+0.39 \pm 0.11$ | **27** : $+0.19 \pm 0.06$ | **32** : $-0.27 \pm 0.12$ | | |
| **20** | **15** : $-0.05 \pm 0.06$ | **21** : $-0.04 \pm 0.05$ | **22** : $+0.11 \pm 0.08$ | | | | |
| **21** | **20** : $-0.04 \pm 0.05$ | **22** : $-0.04 \pm 0.05$ | | | | | |
| **22** | **15** : $-0.28 \pm 0.11$ | **17** : $+0.06 \pm 0.07$ | **18** : $+0.10 \pm 0.07$ | **20** : $+0.11 \pm 0.08$ | **21** : $-0.04 \pm 0.05$ | | |
| **23** | **16** : $-0.10 \pm 0.07$ | **17** : $-0.29 \pm 0.10$ | **18** : $-0.03 \pm 0.09$ | **19** : $-0.00 \pm 0.08$ | **32** : $+0.04 \pm 0.08$ | | |
| **24** | **14** : $+0.38 \pm 0.15$ | **19** : $+0.39 \pm 0.11$ | **25** : $-0.26 \pm 0.13$ | **27** : $-0.22 \pm 0.10$ | **28** : $+0.06 \pm 0.06$ | **30** : $-0.14 \pm 0.05$ | **32** : $-0.02 \pm 0.08$ |
| **25** | **24** : $-0.26 \pm 0.13$ | **26** : $+0.08 \pm 0.04$ | **28** : $-0.15 \pm 0.08$ | **29** : $-0.06 \pm 0.10$ | **30** : $+0.28 \pm 0.15$ | | |
| **26** | **25** : $+0.08 \pm 0.04$ | **29** : $-0.18 \pm 0.05$ | | | | | |
| **27** | **19** : $+0.19 \pm 0.06$ | **24** : $-0.22 \pm 0.10$ | **28** : $-0.04 \pm 0.04$ | **29** : $+0.01 \pm 0.08$ | | | |
| **28** | **24** : $+0.06 \pm 0.06$ | **25** : $-0.15 \pm 0.08$ | **27** : $-0.04 \pm 0.04$ | **29** : $-0.09 \pm 0.04$ | | | |
| **29** | **25** : $-0.06 \pm 0.10$ | **26** : $-0.18 \pm 0.05$ | **27** : $+0.01 \pm 0.08$ | **28** : $-0.09 \pm 0.04$ | | | |
| **30** | **24** : $-0.14 \pm 0.05$ | **25** : $+0.28 \pm 0.15$ | | | | | |
| **32** | **14** : $+0.07 \pm 0.12$ | **16** : $-0.05 \pm 0.06$ | **19** : $-0.27 \pm 0.12$ | **23** : $+0.04 \pm 0.08$ | **24** : $-0.02 \pm 0.08$ | | |

Table S-5: **Estimated Treatment Effects Between Adjacent NYC School Districts.** Each row gives the posterior (mean $\pm$ standard deviation) of the inverse-variance LATEs for one district (row header) compared to its neighbors. For example the first cell indicates an estimated average change in log house prices per square foot of -0.29 when crossing the border from district 13 to 14.