Journal of the
American
Statistical
Association

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Do School Districts Affect NYC House Prices? Identifying Border Differences Using a Bayesian Nonparametric Approach to Geographic Regression Discontinuity Designs

Maxime Rischard, Zach Branson, Luke Miratrix & Luke Bornn

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 534, 619–631: Applications and Case Studies
https://doi.org/10.1080/01621459.2020.1817749

Taylor & Francis
Taylor & Francis Group

Do School Districts Affect NYC House Prices? Identifying Border Differences Using a Bayesian Nonparametric Approach to Geographic Regression Discontinuity Designs

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

[a]Department of Statistics, Harvard University, Cambridge, MA; [b]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA; [c]Graduate School of Education, Harvard University, Cambridge, MA; [d]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada

ABSTRACT
What is the premium on house price for a particular school district? To estimate this in New York City we use a novel implementation of a geographic regression discontinuity design (GeoRDD) built from Gaussian processes regression (kriging) to model spatial structure. With a GeoRDD, we specifically examine price differences along borders between "treatment" and "control" school districts. GeoRDDs extend RDDs to multivariate settings; location is the forcing variable and the border between school districts constitutes the discontinuity threshold. We first obtain a Bayesian posterior distribution of the price difference function, our nominal treatment effect, along the border. We then 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). We test for nonzero LATE with a calibrated hypothesis test with good frequentist properties, which we further validate using a placebo test. Using our methodology, we identify substantial differences in price across several borders. In one case, a border separating Brooklyn and Queens, we estimate a statistically significant 20% higher price for a house on the more desirable side. We also find that geographic features can undermine some of these comparisons. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

ARTICLE HISTORY
Received October 2018
Accepted August 2020

KEYWORDS
Bayesian testing; Causal inference; Gaussian processes; Housing market; Kriging; Regression discontinuity

1. Introduction

How much do New York City (NYC) house prices vary as a function of different school zones? In NYC, the city-wide public school district (the largest school district in the United States) is divided into 32 subdistricts. Many complex rules and systems control school access, with some of these systems depending on these school zones. More broadly, it is commonly believed that quality of schools can have significant effects on the price of nearby housing, as at least some parents will make efforts to live in locations that increase the chances of access to perceived higher quality schools. And like any city in the United States, NYC public schools are said to vary in quality. We then ask, does something like this mechanism play a role in determining NYC housing costs? Of course, NYC house prices vary for a multitude of reasons, many not related to the quality of public education in a given subdistrict at all. Our goal is to measure the cost, in terms of housing, of being in one school subdistrict versus another beyond these other varying factors.

Economists have long been interested in measuring and using variation in house price across district lines to estimate the implicit price of school quality. Black and Machin (2011) reviewed the existing literature on this topic, and classify the various identification strategies and methodologies that have been proposed. Typically, the theoretical underpinning of this literature is the hedonic valuation model (Rosen 1974; Sheppard 1999), whereby the price of a house is decomposed into (usually log-linear) contributions from characteristics of the house, of the neighborhood, and of the schools that local residents can access. The estimation of the causal effect of a unit of school quality on the price of a house is made difficult by unobserved neighborhood characteristics which correlate with both school quality and house prices. Consequently, ordinary least squares (OLS) estimates will be confounded by these unobserved factors (Black and Machin 2011).

However, by paying attention to the geographic boundaries of public school subdistricts within NYC, we can potentially identify the causal effect that one subdistrict versus another could have on house prices. Starting with Black (1999), economists have recognized the opportunity offered by public school systems that rely on attendance districts to assign children to schools. If houses on opposite sides of the boundaries between districts have access to the same neighborhood amenities (public transport, parks, etc.) except for schools, then the price difference between otherwise similar houses can plausibly be attributed to difference in the quality of the schools that their residents can access. This strategy to identify and estimate the causal effect of school quality can be understood as a regression discontinuity design (RDD) along the entire district boundary. Given New York's complex methods for assigning schools to

CONTACT Maxime Rischard maxime.rischard@gmail.com Department of Statistics, Harvard University, 1 Oxford St, Cambridge, MA 02138.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

students, it is an open question whether these subdistrict lines will similarly result in systematically different prices.

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) traced 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.

In our application, the forcing variable is geographic location of a house sale, and the threshold, no longer a single point, is the border between school subdistricts. Until relatively recently, most theory and applications of RDDs were for univariate cases, but, beginning with Papay, Willett, and Murnane (2011), methods have been developed to analyze multivariate RDDs. Imbens and Zajonc (2011) extended the local linear regression methods (see Imbens and Lemieux 2008) that are popular for analyzing univariate RDDs (1D RDDs) to settings with multiple forcing variables. When these forcing variables are spatial, that is, treatment and control units are separated by a geographical border—as is the case in our application—this becomes a geographical regression discontinuity design (GeoRDD). A convenient approach to GeoRDDs sometimes seen in applied work (e.g., Holmes 1998; Magruder 2012; Chen et al. 2013; MacDonald, Klick, and Grunwald 2015) is to reduce it to a 1D RDD by using the signed minimum distance to the boundary (positive for treatment and negative for control) as the forcing variable, a method we refer to as "projected 1D RDD." But this can fail to capture the spatial variation in outcomes, resulting in a confounded estimator: see Section 1 of the supplementary materials and Section 4.2 of Keele and Titiunik (2015). Firmer theoretical foundations for GeoRDDs are built by Keele and Titiunik (2015), who extended the identification assumptions that were formalized by Hahn, Todd, and Van der Klaauw (2001) for 1D RDDs, and by Imbens and Zajonc (2011) for multivariate RDDs. To estimate the treatment effect, Keele and Titiunik (2015) and Keele et al. (2017) applied the projected 1D RDD method locally around points on the border, thus alleviating the problem of spatial confounding. For valuing school quality, Gibbons, Machin, and Silva (2013) and Fack and Grenet (2010) proposed matching methodologies to address the issue of spatial confounding. By matching similar units on opposite sides of the border that are near each other geographically, the difference in their outcomes can plausibly be attributed to the presence of the border. Similarly, Keele, Titiunik, and Zubizarreta (2015) used the matching methods of Zubizarreta (2012) in a GeoRDD to estimate the effect of ballot initiatives on voter turnout in Milwaukee, Wisconsin.

Spatial statistics, the branch of statistics dedicated to inference for geographical units with spatially correlated outcomes, has been mostly absent from this literature. In this article, to take advantage of the geographic information available in our NYC application, we develop a framework for analyzing GeoRDDs that is a spatial analogue of 1D RDD methods. Broadly, 1D RDD methodologies (Imbens and Lemieux 2008) are composed of three steps: (i) fit a smooth *function* to the outcomes against the forcing variable on each side of the threshold, (ii) extrapolate the functions to the *threshold point*, and (iii) 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 our geographical RDD application, our framework proceeds analogously: (i) fit a smooth *surface* to the outcomes (in our case, house prices) against the geographical covariates in each region, (ii) extrapolate the surfaces to the *border curve* (in our case, the boundary between two subdistricts), and (iii) take the *pointwise* difference between the two extrapolations to estimate the treatment effect along the border. The usefulness of spatial models is then evident; we use kriging, also known as Gaussian process regression (GPR), to fit and extrapolate the outcomes, but other spatial methods could also be suitable. For 1D RDDs, Branson et al. (2019) proposed a GPR methodology that exhibits promising coverage and MSE properties compared to local linear regression. We believe this approach to be particularly suitable to GeoRDDs, as GPR is a well-established tool in spatial statistics for fitting smoothly varying spatial processes. See Banerjee, Carlin, and Gelfand (2014) for a textbook introduction to kriging for spatial data, and Rasmussen and Williams (2006) for a machine learning perspective.

Our implementation of the methods proposed in this article uses the `GaussianProcesses.jl` package (Fairbrother et al. 2018) for the julia programming language (Bezanson et al. 2017). All replication materials for our analysis are available on the first author's GitHub account.

Section 2 explains our GeoRDD methodology. In Section 2.1, we use GPR to estimate the treatment effect along the border by extending the model of Branson et al. (2019) 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 2.3, where we obtain Bayesian estimators for multiple possible LATE estimands and discuss their properties. In Section 2.4, we turn to hypothesis testing, and propose a method to test against the null hypothesis of no treatment effect along the border.

In Section 3, we apply our methodology to the problem of valuing school quality, using publicly available data of property sales in NYC to determine whether school districts affect property prices. Initially focusing on a single border between two school districts, we estimate the treatment effect everywhere along the border, obtain estimates of the LATE, and perform and validate a hypothesis test. For that border, we find a statistically significant difference in price across the border with a p-value of 0.003, 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 cannot be attributed solely to the quality or reputation of the schools, as this border also separates the boroughs of Brooklyn and Queens, thus confounding the causal effect of the districts. We then extend to the other borders between subdistricts in NYC, and talk about cross-cutting themes in our evaluation.

2. GeoRDD Estimation 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 two-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 nonoverlapping, intersecting only at the border $\mathcal{B}$ between them. Throughout this article, points on the border are denoted by $\boldsymbol{b}$. Under Neyman's potential outcomes framework for causal inference (Rubin 1974; Splawa-Neyman et al. 1923/1990, but see, e.g., Rosenbaum 2010 for a good discussion and overview), 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 concatenating $\boldsymbol{Y}_T$ 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

$$\begin{aligned}\tau &= \mathbb{E}\left[Y_{iT} \mid X_i = b\right] - \mathbb{E}\left[Y_{iC} \mid X_i = b\right] \\ &= \lim_{x \downarrow b} \mathbb{E}\left[Y \mid X = x\right] - \lim_{x \uparrow b} \mathbb{E}\left[Y \mid X = x\right]\,, \qquad (1)\end{aligned}$$

where the second equality requires the assumption that the conditional regression functions $\mathbb{E}\left[Y_{iT} \mid X_i = x\right]$ and $\mathbb{E}\left[Y_{iC} \mid X_i = x\right]$ 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:

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

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 $\mathbb{E}\left[Y_{iT} \mid \boldsymbol{s}_i = \boldsymbol{s}\right]$ and $\mathbb{E}\left[Y_{iC} \mid \boldsymbol{s}_i = \boldsymbol{s}\right]$ 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).

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 article, we use GPR for this purpose. GPR, known as kriging in the spatial statistics literature, is a Bayesian nonparametric method for fitting smooth functions. Recently, Branson et al. (2019) 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. (2019) 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 (GP) $f(\boldsymbol{s})$, and iid normal noise ϵ. The GP 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, Carlin, and Gelfand (2014) and Rasmussen and Williams (2006) for examples. In this article, we use the Matérn $\nu = 1/2$ covariance (also known as the exponential kernel) for its ease of understanding and its prevalence in applied spatial statistics. This yields the outcomes model:

$$\begin{aligned}&Y_{iT} = \underbrace{m_T + f_T(\boldsymbol{s}_i)}_{g_T(\boldsymbol{s}_i)} + \epsilon_i \quad \text{and} \quad Y_{iC} = \underbrace{m_C + f_C(\boldsymbol{s}_i)}_{g_C(\boldsymbol{s}_i)} + \epsilon_i\,, \\ &\quad \text{with } \epsilon_i \overset{\perp\!\!\!\perp}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)\,; \qquad (3) \\ &f_T, f_C \overset{\perp\!\!\!\perp}{\sim} \mathcal{GP}(0, k(\boldsymbol{s}, \boldsymbol{s}')) \quad \text{with} \\ &\quad k(\boldsymbol{s}, \boldsymbol{s}') = \sigma_{\mathrm{GP}}^2 \exp(-\left\|\boldsymbol{s} - \boldsymbol{s}'\right\|_2 / \ell)\,.\end{aligned}$$

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

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

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. $\tau(\boldsymbol{b})$ is a conditional average treatment effect (see, e.g., Hill 2011), giving the expected treatment effect at a given location, not including idiosyncratic, unit-specific noise that could include individual-level treatment effect heterogeneity.

In this specification, the hyperparameters ℓ, σ_{GP}, and σ_ϵ are the same in the treatment and control regions, so we assume that the spatial smoothness of the responses 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. (2019).

2.2. Inference of the Treatment Effect

If m_T and m_C are given normal priors with variance σ_m^2, then the model specification (3) can be used to obtain covariances between the observations, the GPs, and the mean parameters. Given hyperparameters $\boldsymbol{\theta} = (\ell, \sigma_{\mathrm{GP}}, \sigma_\epsilon, \sigma_m)$, 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 GPs 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 $\boldsymbol{b}_{1:R} = \{\boldsymbol{b}_1, \dots, \boldsymbol{b}_R\}$ of R "sentinel" units dotted along $\mathcal{B}$. The extrapolation step then follows mechanically through multivariate normal theory. On the treatment side, for example, we have a posterior for the price curve at the sentinel points of

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

with the various covariance matrices $\boldsymbol{K}_{\mathcal{B}\mathcal{B}}, \boldsymbol{K}_{\mathcal{B}T}, \Sigma_{TT}$, etc., derived from the model specification (see Appendix A for their derivations and expressions). 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

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

Our $\tau(\boldsymbol{b}_{1:R})$ is an R-vector with the rth entry $\tau(\boldsymbol{b}_r)$ being the treatment effect evaluated at $\boldsymbol{b}_r$. The posterior mean and covariance of $\tau(\boldsymbol{b}_{1:R})$ are the primary output of our GeoRDD analysis; we refer to (6) as the "cliff height" estimator. For a frequentist view, we would take the $\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}$ as our point estimates.

This leaves the choice of the hyperparameters: $\boldsymbol{\theta} = \ell$, σ_{GP}, σ_ϵ, and σ_m. For σ_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 $p(\boldsymbol{Y} \mid \ell, \sigma_{\mathrm{GP}}, \sigma_\epsilon)$, which is available analytically and easily computed for GPR. This empirical Bayes approach is common in spatial and machine learning applications of GPs. An alternative would be to also specify a prior on the hyperparameters, which would be preferable to fully account for the uncertainty in the model, but fully Bayesian inference of large GP models tends to be computationally expensive.

2.3. Estimating the Local Average Treatment Effect

Our $\tau(\boldsymbol{b})$ gives the treatment impact along the full border, but we often want to summarize it into an overall LATE. Given a weight function $w_{\mathcal{B}}(\boldsymbol{b})$ defined everywhere on the border $\mathcal{B}$ we can calculate the LATE as the weighted average of $\tau(\boldsymbol{b})$ using a weighted mean integral. We approximate this integral as a weighted sum at the sentinels $\boldsymbol{b}_{1:R}$:

$$\tau^w = \frac{\oint_{\mathcal{B}} w_{\mathcal{B}}(\boldsymbol{b}) \tau(\boldsymbol{b}) d\boldsymbol{b}}{\oint_{\mathcal{B}} w_{\mathcal{B}}(\boldsymbol{b}) d\boldsymbol{b}} \approx \frac{\sum_{r=1}^{R} w_{\mathcal{B}}(\boldsymbol{b}_r) \tau(\boldsymbol{b}_r)}{\sum_{r=1}^{R} w_{\mathcal{B}}(\boldsymbol{b}_r)} \,. \tag{7}$$

This approximation assumes the sentinels are evenly spaced; if they are not, each term in the sum needs to be reweighted by the length of the border the sentinels represent.

The posterior distribution of $\tau(\boldsymbol{b}_{1:R})$ is multivariate normal (see (6)). Since τ^w is a linear combination of $\tau(\boldsymbol{b}_{1:R})$, its posterior is also normal, with mean $\mu_{\tau^w|Y}$ and variance $\Sigma_{\tau^w|Y}$ approximated by

$$\begin{aligned} \mu_{\tau^w|Y} &\approx \frac{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^{\intercal} \boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}}{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^{\intercal} \boldsymbol{1}_R} \quad \text{and} \\ \Sigma_{\tau^w|Y} &\approx \frac{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^{\intercal} \Sigma_{\boldsymbol{b}_{1:R}|Y} w_{\mathcal{B}}(\boldsymbol{b}_{1:R})}{\left(w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^{\intercal} \boldsymbol{1}_R\right)^2} \,, \end{aligned} \tag{8}$$

where $w_{\mathcal{B}}(\boldsymbol{b}_{1:R})$ is the R-vector of the weight function evaluated at the sentinels, and $\boldsymbol{1}_R$ is an R-vector of ones. Given a weight function, the "natural" estimand in (7) for the estimator in (8) is the same weighted mean applied to the true τ.

An alternative perspective on these estimators is given by the weights induced on the observations. Combining Equations (5), (6), and (8), we obtain that the posterior mean of τ^w is a weighted difference in means between the treatment and control units:

$$\mathbb{E}\left(\tau^w \mid \boldsymbol{Y}\right) = \boldsymbol{w}_T^{\intercal} \boldsymbol{Y}_T - \boldsymbol{w}_C^{\intercal} \boldsymbol{Y}_C, \tag{9}$$

with vectors of "unit weights" given by

$$\boldsymbol{w}_T = \frac{\Sigma_{TT}^{-1} \boldsymbol{K}_{\mathcal{B}T}^{\intercal} w_{\mathcal{B}}(\boldsymbol{b}_{1:R})}{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^{\intercal} \boldsymbol{1}_R} \quad \text{and} \quad \boldsymbol{w}_C = \frac{\Sigma_{CC}^{-1} \boldsymbol{K}_{\mathcal{B}C}^{\intercal} w_{\mathcal{B}}(\boldsymbol{b}_{1:R})}{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^{\intercal} \boldsymbol{1}_R} \,, \tag{10}$$

for treatment and control units, respectively.

We next motivate and consider four possible choices of $w_{\mathcal{B}}(\boldsymbol{b})$, and explore interpretations, advantages, and drawbacks. Section 3 of the supplementary materials gives two further choices, the projected land LATE τ^{GEO}, and the projected superpopulation LATE τ^{POP}. In that section, we also provide a simulation study to better understand the different LATE choices. A summary of their properties is provided in Table 1.

2.3.1. Uniform Weighting

The simplest choice is uniform weights $w_{\mathcal{B}}(\boldsymbol{b}) = 1$, a seemingly reasonable and unopinionated decision. The uniformly weighted LATE τ^{UNIF} is estimated by averaging the entries of the mean posterior at the sentinels. Following (7) and (8):

$$\begin{aligned} \tau^{\mathrm{UNIF}} &= \oint_{\mathcal{B}} \tau(\boldsymbol{b}) d\boldsymbol{b} \Big/ \oint_{\mathcal{B}} d\boldsymbol{b} \,, \\ \tau^{\mathrm{UNIF}} \mid \boldsymbol{Y}, \boldsymbol{\theta} &\sim \mathcal{N}\left(\mu_{\tau^{\mathrm{UNIF}}|Y}, \Sigma_{\tau^{\mathrm{UNIF}}|Y}\right) \text{ , with} \\ \mu_{\tau^{\mathrm{UNIF}}|Y} &= \left(\boldsymbol{1}_R^{\intercal} \boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}\right) / R \quad \text{and} \\ \Sigma_{\tau^{\mathrm{UNIF}}|Y} &= \left(\boldsymbol{1}_R^{\intercal} \Sigma_{\boldsymbol{b}_{1:R}|Y} \boldsymbol{1}_R\right) / R^2 \,. \end{aligned} \tag{11}$$

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

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 these empty areas will have large posterior variances, which will dominate the posterior variance of τ^{UNIF}, potentially jeopardizing the successful detection of otherwise strong treatment effects.

2.3.2. Density-Weighted

We can address this issue by weighting the treatment effect at each sentinel location by the local population density ρ, that is, 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:

$$\tau^{\rho} = \mathbb{E}\left[Y_{iT} - Y_{iC} \mid \boldsymbol{s}_i \in \mathcal{B}\right] . \tag{12}$$

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 followed 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, or any spatial point process model. Strictly speaking, the uncertainty of the local density estimate should then be propagated to the estimate of τ^{ρ}, which may therefore no longer have a normally distributed or analytically tractable posterior.

Both the uniform and density-weighted estimators are undesirably susceptible 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. These regions will therefore get disproportionately more weight, purely as an artifact of the border shape: the more wiggly the border, the more weight, regardless of the number and placement of actual units in the region. See Section 3.3 of the supplementary materials for a simulation demonstrating this susceptibility to border topology.

2.3.3. Inverse-Variance Weighted

The unwelcome dependence of the τ^{UNIF} and τ^{ρ} estimands on the border topology is a symptom of the geometry of the GeoRDD: the border treatment effect function (2) is defined on a one-dimensional manifold $\mathcal{B}$, which itself is embedded in a Euclidean two-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 τ^{INV}:

$$\begin{aligned} \tau^{\text{INV}} \mid \boldsymbol{Y}, \boldsymbol{\theta} &\sim \mathcal{N}\left(\mu_{\tau^{\text{INV}}|Y}, \Sigma_{\tau^{\text{INV}}|Y}\right) \text{, with} \\ \mu_{\tau^{\text{INV}}|Y} &= \left(\mathbf{1}_R^{\intercal} \Sigma_{\boldsymbol{b}_{1:R}|Y}^{-1} \boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}\right) \Big/ \left(\mathbf{1}_R^{\intercal} \Sigma_{\boldsymbol{b}_{1:R}|Y}^{-1} \mathbf{1}_R\right) \quad \text{and} \\ \Sigma_{\tau^{\text{INV}}|Y} &= 1 \Big/ \left(\mathbf{1}_R^{\intercal} \Sigma_{\boldsymbol{b}_{1:R}|Y}^{-1} \mathbf{1}_R\right) . \end{aligned} \tag{13}$$

This estimator efficiently extracts the information from the posterior treatment effect, as it can be shown to minimize the posterior variance among weighted averages of the form (7). 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 $w_{\mathcal{B}}(\boldsymbol{b}_{1:R}) = \Sigma_{\boldsymbol{b}_{1:R}|Y}^{-1} \mathbf{1}_R$. 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, Morgan, and Zaslavsky 2018). Regression adjustment of multisite trials with fixed effects is similarly a precision average impact estimate (see, e.g., Angrist and Pischke 2008; Miratrix, Weiss, and Henderson 2020). The 1D RDD could be said to provide yet another example, as the estimand (1) focuses on the treatment effect near the threshold not because these units are of particular substantive interest, but because the available data restricts estimation of the treatment effect elsewhere.

2.3.4. Projected Finite Population Weighted

All LATE estimators considered so far presuppose evenly spaced sentinel points, which are then given weights. Alternatively, we can project onto the border the treatment and control units that are within a prechosen distance Δ of the border, and use these projected unit locations without weights (see Figure 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 $\mathbb{I}^{\Delta}(\boldsymbol{s}) = \mathbb{I}\left\{\text{dist}_{\mathcal{B}}(\boldsymbol{s}) \leq \Delta\right\}$ indicate inclusion in the border vicinity. The projected finite-population τ^{PROJ} is then the uniformly weighted mean applied with the projected unit locations instead of the evenly spaced sentinels. We can therefore modify (11), 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 τ^{PROJ}:

$$\begin{aligned} \tau^{\text{PROJ}} \mid \boldsymbol{Y}, \boldsymbol{\theta} &\sim \mathcal{N}\left(\mu_{\tau^{\text{PROJ}}|Y}, \Sigma_{\tau^{\text{PROJ}}|Y}\right) \text{, with} \\ \mu_{\tau^{\text{PROJ}}|Y} &= \sum_{i=1}^{n} \mathbb{I}^{\Delta}(\boldsymbol{s})_i \, \mathbb{E}\left[\tau\left(\text{proj}_{\mathcal{B}}(\boldsymbol{s}_i)\right) \mid \boldsymbol{Y}, \boldsymbol{\theta}\right] \\ &\qquad \Big/ \sum_{i=1}^{n} \mathbb{I}^{\Delta}(\boldsymbol{s})_i \text{ , and} \\ \Sigma_{\tau^{\text{PROJ}}|Y} &= \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{I}^{\Delta}(\boldsymbol{s})_i \, \mathbb{I}^{\Delta}(\boldsymbol{s})_j \operatorname{cov} \left[\tau\left(\text{proj}_{\mathcal{B}}(\boldsymbol{s}_i)\right), \tau\left(\text{proj}_{\mathcal{B}}(\boldsymbol{s}_j)\right) \mid \boldsymbol{Y}, \boldsymbol{\theta}\right]}{\left(\sum_{i=1}^{n} \mathbb{I}^{\Delta}(\boldsymbol{s})_i\right)^2} . \end{aligned} \tag{14}$$

The posterior expectations and covariances in (14) are easily derived and computed analogously to the procedure of Section 2.2. Note that τ^{PROJ} is in the class of weighted mean estimands (7), with weight function $w_{\mathcal{B}}(\boldsymbol{b}) = \sum_{i=1}^{n} \mathbb{I}^{\Delta}(\boldsymbol{s})_i \, \delta\left(\boldsymbol{b} - \text{proj}_{\mathcal{B}}(\boldsymbol{s}_i)\right)$, where δ 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

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

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

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 width Δ does not have a large effect on the estimate yielded by (14). A reasonable heuristic is to set Δ to a small multiple of the GP lengthscale ℓ. Regardless, the choice of Δ only affects the location and density of projected units on the border; the τ^{PROJ} estimator assigns nonzero unit weights (9) to all units, whether or not they fall within Δ of the border.

2.3.5. Selecting an Estimand

The properties of the four LATE definitions proposed in this article, along with two additional choices presented in Section 3 of the supplementary materials, are summarized in Table 1. In most applications, we recommend the use of the projected 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 τ^{ρ}, 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.

2.4. Testing for Nonzero 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 distinguish two possible choices of null hypotheses: the sharp null specifies that the treatment effect is zero everywhere along the border, $\tau(\boldsymbol{b}) = 0\ \forall\, \boldsymbol{b} \in \mathcal{B}$, whereas 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 likelihood and a chi-squared statistic in Section 4 of the supplementary materials. We found through simulations and in our applied example that the test presented in this article has superior power and robustness to model misspecification, and therefore recommend its use.

As we saw in Section 2.3, the LATE estimand can be defined in multiple ways. If we choose the inverse-variance weighted mean, then τ^{INV} has posterior given by (13). While the posterior is a Bayesian object, we can use it heuristically to derive a pseudo-p-value $\tilde{p}^{\text{INV}} = 2\Phi(|\mu_{\tau^{\text{INV}}|Y}| / \sqrt{\Sigma_{\tau^{\text{INV}}|Y}})$. 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, $\tilde{p}^{\text{INV}}$ 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 M_0 as a single GP spanning the control and treatment regions, with the same kernel and hyperparameters values obtained through the procedure of Section 2.2. Under M_0, the expected outcomes surface 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 M_0, using the same spatial locations as the original data, and compute $\mu_{\tau^{\text{INV}}|Y^{(b)}}$ according to (13) 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 $\mu_{\tau^{\text{INV}}|Y^{\text{obs}}}$ estimates the p-value:

$$\begin{aligned} p^{\text{INV}} &= p\left(\left|\mu_{\tau^{\text{INV}}|Y}\right| \geq \left|\mu_{\tau^{\text{INV}}|Y^{\text{obs}}}\right| \mid M_0\right) \\ &\approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left\{\left|\mu_{\tau^{\text{INV}}|Y^{(b)}}\right| \geq \left|\mu_{\tau^{\text{INV}}|Y^{\text{obs}}}\right|\right\}. \end{aligned} \tag{15}$$

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 B. Note that the p-value for this test is derived under the parametric null model 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).

2.4.1. Placebo Tests

GP models are almost always misspecified. We do not believe that the GP with stationary Matérn 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"), 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°, ..., 180°

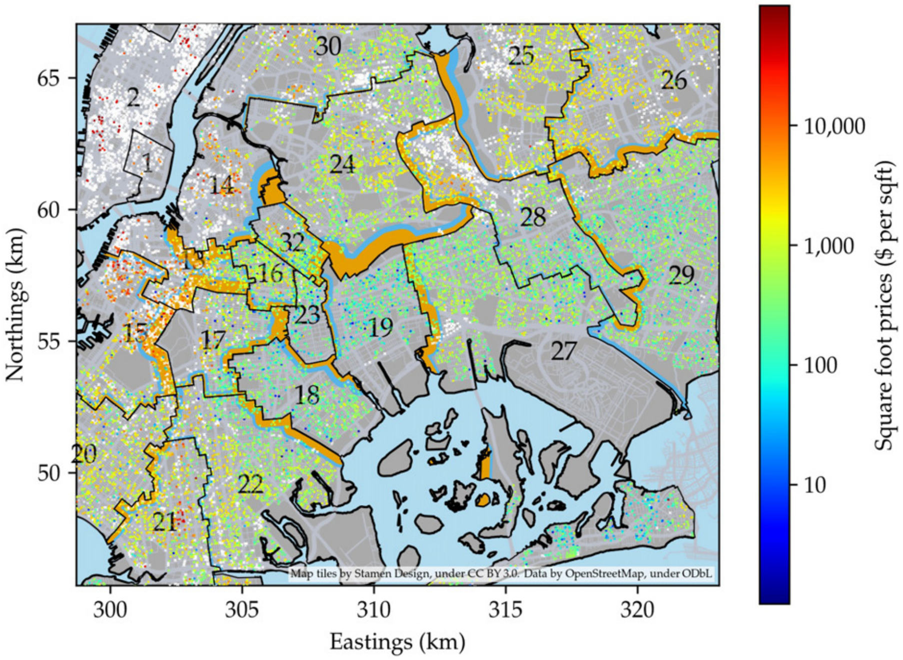
30
25
26
2
1
14
24
28
32
16
29
23
19
15
17
27
18
20
22
21
Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL
Northings (km)
65
60
55
50
Eastings (km)
300
305
310
315
320
Square foot prices ($ per sqft)
10,000
1,000
100
10

Figure 1. Map of property sales in New York City along with estimated pairwise border impacts. Each dot is a sale, and its color indicates the price per square foot. White indicate sales with missing square footage. All pairwise estimates of the inverse variance LATE between adjacent school districts shown by the orange and blue buffers along their borders, with the thickness of the orange buffer proportional to the posterior mean, and the blue beyond it proportional to the posterior standard deviation. All buffers drawn on the side estimated to have higher house prices. Each district is labeled by its number.

counter-clockwise from horizontal, each positioned so that half of the units fall on either side of the line to maximize power. Because the borders are drawn arbitrarily, without reference to the outcomes, we should not expect to observe a discontinuous jump in outcomes. We apply the calibrated inverse-variance test procedure described above to the data arbitrarily divided by each placebo border, and hope to obtain a roughly uniform distribution of p-values. The placebo p-values are highly correlated, resulting in a small effective sample size, but this procedure nonetheless allows us to visually verify that the p-values are not blatantly biased. Gibbons, Machin, and Silva (2013) performed a falsification test that is similar in spirit to our procedure. They shift the locations of the housing transactions 10 km North and East, and show that their matching method, unlike OLS, no longer yields a significant estimate of the effect of school quality on house prices.

3. Valuing NYC School Districts

We now turn to our NYC public schools application that we discussed in Section 1. Specifically, we will use our methodology to study the effect of school districts on house prices in NYC. The city publishes information pertaining to property sales within the city in annualized datasets, available at *https://www1.nyc.gov/site/finance/taxes/property-annualized-sales-update.page*, and in this section we focus on the 2015 dataset. 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 subdistricts. 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 whether we can measure a discontinuous jump in house prices across the borders separating school districts.

We first geocode the address of each sale by merging the sales with the NYC Department of Finance's Digital Tax Map, which contains X and Y coordinates for the centroid of every parcel in NYC, identified by its borough, zip code, block, and lot. These coordinates are given in the EPSG:2263 projection, which we also adopt. Details and code for this preprocessing step are available from the first author's GitHub account.

We then filter the 83,441 sales by removing (i) 51,741 sales outside of the family homes building class categories (one, two, and three family dwellings), (ii) 11,141 remaining properties without a reported sale price, (iii) 62 remaining sales missing the square footage information, (iv) 76 remaining properties which could not be geocoded, and (v) 905 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 19,516 sales is displayed in Figure 1. The 33,331 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.

3.1. Model for Property Prices

Our application relates to the economics literature on valuing school quality (Black and Machin 2011), based on the hedonic valuation model (Rosen 1974; Sheppard 1999) which typically takes the form of a linear model for the log of the sale price p of a property at location $\boldsymbol{s}$ (see, e.g., Gibbons, Machin, and Silva 2013):

$$p = s(\boldsymbol{s})\beta + x(\boldsymbol{s})\gamma + g(\boldsymbol{s}) + \epsilon\,, \tag{16}$$

where $s(\boldsymbol{s})$ is the expected quality of the schools that residents near $\boldsymbol{s}$ can access, x is a set of observed property and neighborhood covariates, $g(\boldsymbol{s})$ captures spatially correlated unobserved covariates, and ϵ represents unobserved characteristics of the property and errors that are independent of x and $\boldsymbol{s}$. "School quality" is variously defined and estimated (Black and Machin 2011): average standardized test scores, survey responses, funding levels, pass rates, publicly, etc.

Most research focuses on estimating β, the effect of a unit of school quality on the log-price, while addressing confounding due to $g(\boldsymbol{s})$. Imagine moving a property sale an infinitesimal distance from one side of the border to the other; the difference in prices is then:

$$\Delta(p) = \Delta(s(\boldsymbol{s})\beta) + \Delta(x(\boldsymbol{s})\gamma) + \Delta g(\boldsymbol{s}) + \Delta\epsilon\,. \qquad (17)$$

The last three terms on the right-hand-side equal zero: the property has not changed, and x and g are assumed smooth and continuous, so only the change in attendance districts can explain the change in price. The GeoRDD uses this idea to identify $\Delta(s(\boldsymbol{s})\beta)$, the jump in price attributable to the difference between school districts. However, to estimate β, the hedonic model requires the further assumptions that school quality is well-defined and measured, and that its effect on log-prices is linear and constant. In this article, we avoid these assumptions by seeking to directly estimate the jump in prices at the border, without attempting to attribute the jump to a specific measure of school quality.

In our application, the outcome of interest is price per square foot of a property sale. As is commonly done in analyses of real estate prices, we take its logarithm to reduce the skew in the outcomes. The complete model is then a GP 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):

$$\begin{aligned} Y_i &= m_{\texttt{Distr}[i]} + \gamma_{\texttt{BuildClass}[i]} + f_{\texttt{Distr}[i]}(\boldsymbol{s}_i) + \epsilon_i\,, \\ &\qquad \epsilon_i \overset{\perp\!\!\perp}{\sim} \mathcal{N}\left(0, \sigma_\epsilon^2\right)\,, \\ \gamma_l &\sim \mathcal{N}\left(0, \sigma_\gamma^2\right)\,, \\ &\qquad \text{for } l = 1, \ldots, L_{\texttt{BuildClass}}\,, \\ m_j &\sim \mathcal{N}\left(0, \sigma_m^2\right)\,, \; f_j \sim \mathcal{GP}\left(0, k(\boldsymbol{s}, \boldsymbol{s}')\right)\,, \\ &\qquad \text{for } j = 1, \ldots, J_{\texttt{Distr}}, \end{aligned} \qquad (18)$$

where k is the Matérn covariance function as in (3).

A visual inspection of the house sales map in Figure 1 initially drew our attention toward the border between districts 19 and 27, which we arbitrarily designate as "treatment" and "control," respectively. Importantly, the border between the two districts is also part of the border between Brooklyn and Queens. This is an instance of what Keele and Titiunik (2015) termed "compound treatments," a frequent concern in GeoRDDs. Therefore, we are *measuring* a discontinuity in the house prices at the border, but 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. 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; this snapshot is made more accurate by correcting for covariates that pertain to the building rather than land.

3.2. Cliff Height Estimator

We fit the hyperparameters σ_γ, σ_{GP}, ℓ, and σ_ϵ 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 σ_m fixed to 20 to give the district means m_j a weak prior. The fitted hyperparameters were $\widehat{\sigma_\epsilon} = 0.40$, $\widehat{\sigma_{\mathrm{GP}}} = 0.29$, $\widehat{\sigma_\gamma} = 0.14$, and $\widehat{\ell} = 5.7$ km.

We seek to estimate the treatment effect function τ on the border between the two districts, adjusting for measured building and site characteristics. We could proceed by computing the cliff height estimator with covariates (5). But to simplify the analysis we instead residualize the prices by estimating γ and obtaining residuals $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\gamma}}$. We then treat these residuals as the observed outcomes in a GeoRDD analysis with no nonspatial covariates. In our context, the posterior variance of $\boldsymbol{\gamma}$ is low, and therefore the two approaches yield very similar results, but conditioning on the estimate of $\boldsymbol{\gamma}$ is computationally convenient. See Section 2 of the supplementary materials for further details.

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 2, and shows that τ is estimated as negative everywhere along the border, which corresponds to higher property prices in district 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 $\tau = 0$.

3.3. Average Log-Price Increase

The cliff height Figure 2 shows a negative treatment effect everywhere along the border, which can be averaged by the estimators we developed in Section 2.3. Our two recommended estimators, based on inverse-variance weighting and finite-population projection of units within $\Delta = \ell$ of the border, yield LATE estimates of -0.22 and -0.18, respectively, which corresponds to a roughly 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 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 2.3 applied to this setting are shown in Table 2. In this example, the different estimators yield similar answers, as the border is fairly straight and short relative to the fitted lengthscale.

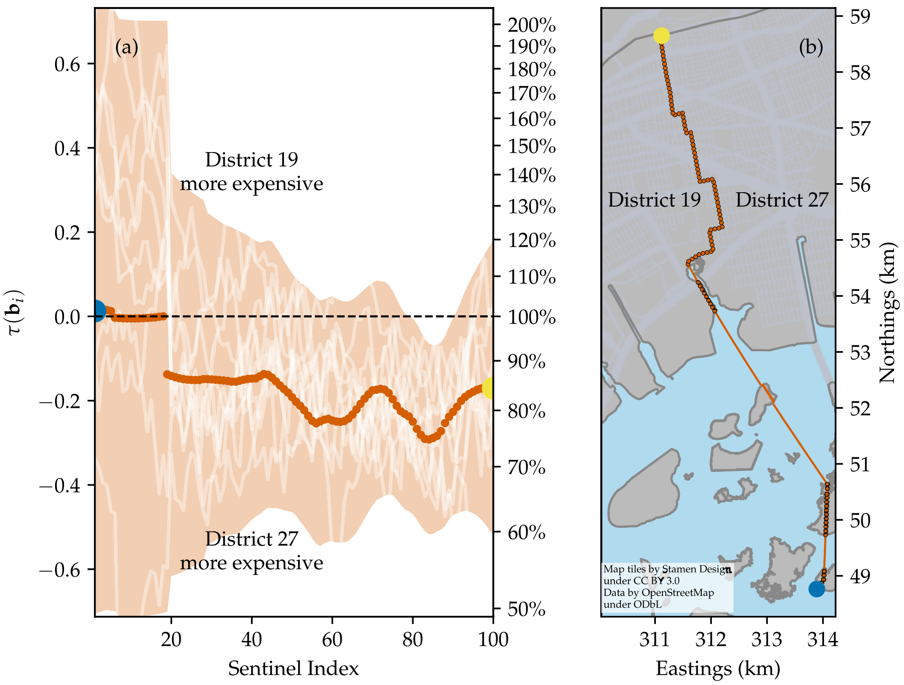

Figure 2. (a) Cliff height estimator (6) for the school district effect on house prices per square foot between district 27 and district 19, with 95% credible envelope. The left y-axis shows the difference in log prices per square foot; positive values mean prices are higher in district 19. The right 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 (skipping regions of water). The southernmost sentinel (shown as a blue circle in both plots) has index 1, while the northernmost sentinel (shown in yellow) has index 100.

Table 2. Average difference in log price per square foot between school districts 19 and 27.

	Posterior		
Estimand	Mean	SD	Tail prob.
τ^{UNIF}	−0.16	0.10	5.25%
τ^{ρ}	−0.22	0.07	0.04%
τ^{INV}	−0.22	0.06	0.04%
τ^{PROJ}	−0.18	0.09	2.04%
τ^{GEO}	−0.11	0.15	22.24%
τ^{POP}	−0.18	0.08	1.03%

NOTE: For each LATE estimand, we show the mean and standard deviation of its posterior distribution, and the tail probability $p(\tau > 0 \mid \boldsymbol{Y}, \hat{\gamma}, \boldsymbol{\theta})$. Negative LATEs correspond to district 27 being more expensive.

3.4. 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 2.4, which yields a p-value of $p^{\mathrm{INV}} = 0.003$, 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 2.4.1. Within each district, we split the data in half by a line at angles 1°, 3°, 5°, . . . , 179°. 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 a noticeable autocorrelation in the graph of the placebo p-value as a function of angle (see Section 5 of the supplementary materials)—and so the low effective sample size could lead to some apparent departures from uniformity. Nonetheless, our placebo test gives a roughly uniform distribution of p-values, 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. See Section 5 of the supplementary materials for further discussion, as well as the results of the placebo test applied to the other testing approaches.

3.5. Extending to All Borders

Our GeoRDD analysis can be repeated for each pair of adjacent districts. Figure 1 gives an overview of the results by showing the posterior mean and standard deviation of the inverse variance LATE estimated at each border. See Table 5 of the supplementary materials for a table of numerical results. 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. In particular, 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 nonresidential 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 surprising, 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 (e.g., 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 23 to 17, 28 to 29, 29 to 26. The results also show an increase in price at the border from 24 to 28, but this could be confounded by gaps in the sales data due to Forest Park, St. John Cemetery and condos near Queens Blvd. Also note that we report comparisons between 40 pairs of districts, so some false positives would be expected at the 5% significance level. 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 17% price increase from 29 to 26.

4. Discussion

The aim of this article was to estimate the shift in house prices across school districts borders in NYC. Measuring the effect of school quality on house prices has a long history in economics, but most existing methods are vulnerable to unobserved factors—such as neighborhood characteristics—that are correlated with school quality and house prices and thus confound causal effects. For our application, an effective way to identify our causal effects of interest was to frame the application as a GeoRDD and take advantage of methods from the spatial statistics literature to account for spatially varying unobserved factors.

GeoRDDs arise when a treatment is assigned to one region but not to an adjacent region. In our application, "treatment" can be defined as one school district and "control" can be defined as an adjacent school district, thus forming a GeoRDD, where the geographic boundary between districts constitutes the threshold in the RDD. Under smoothness assumptions, houses 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: (i) fit a smooth surface on either side of the border, (ii) extrapolate the surfaces to the border, and (iii) take the difference of the two extrapolations to estimate the treatment effect along the border.

In this article, 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 GPR (kriging) to fit the smooth surfaces to the outcomes in step (i) of our framework. Our approach yields a multivariate normal posterior distribution of the treatment effect for a collection of "sentinel" locations along the border.

We investigated, using a publicly available dataset of one year of NYC property sales, whether school districts can explain systematic differences in property prices. Initially focusing on a single border, we estimated a roughly 20% average increase in house prices per square foot when crossing the border from district 19 to district 27. In contrast to the literature on valuing school quality through house prices, our focus is on inferring the discontinuity in house prices at the border, without attempting to attribute it to a difference in a measured metrics of school quality. In our case, 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. Across all the borders, we see that physical barriers like parks, commercial zones, railways, and major roads can separate neighborhoods. This keeps data away from the borders, breaks the stationarity assumption of the spatial model, and increases the extent of extrapolation performed by the model, which casts doubt on the legitimacy of the estimated treatment effects. Nonetheless, we do find significant effects in several pairs of school districts without such confounding factors.

We also found that averaging the treatment effect along a border has 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. We therefore use more sophisticated estimands, summarized in Table 1, that 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 had to 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 2.3 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.

While our framework is intuitive and well-motivated by the 1D RDD literature, it does have drawbacks. It does not specify a prior directly on the treatment effect along the border; instead it can be shown that our GP model implicitly gives it a wide prior for a constant effect plus a GP prior with double the covariance function k in (3). Such a wide prior can lead to regularization induced confounding (RIC) as defined and demonstrated by Hahn, Murray, and Carvalho (2017) and Hahn et al. (2018). RIC can be understood as the tendency of a Bayesian or regularized model to recruit a treatment effect variable that has a weak prior to explain away a more strongly regularized trend in the control variables. The calibration of the p-values in Section 2.4 safeguards the validity of our proposed hypothesis tests—further validated by placebo tests—but RIC could bias the cliff-face and LATE estimates. Unfortunately Hahn et al.'s solution of first regressing the treatment variable on control variables (spatial covariates in GeoRDDs) to estimate the propensity score cannot be used: the assumption of overlap (Hahn, Murray, and Carvalho 2017, eq. (2)) is violated, because the propensity scores are known to be 0 or 1 in the control and treatment regions, respectively. That said, it could be beneficial to place a direct prior on the treatment effect; this could for example be accomplished by specifying a baseline smooth spatial process that spans both regions, and an independent treatment effect surface with lower prior variance that also spans both regions but is multiplied by 0.5 in the treatment region and −0.5 in the control region. We hope that our work will encourage further exploration of this and other Bayesian nonparametric model specifications for GeoRDDs.

Another limitation of our approach to GeoRDDs is the reliance on modeling assumptions. We modeled the response surfaces as two independent GPs, with iid normal noise for each observation. As is common in spatial statistics, we use GPR as a nonparametric 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.

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 GPs 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 GPs would also improve the robustness of our method. For example, Wilson and Adams (2013) developed spectral mixture (SM) covariance kernels with good extrapolating behavior. These 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 analyze 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 GPs. Banerjee, Carlin, and Gelfand (2014) provided 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 article.

Furthermore, in some applications, it may be of substantive interest to know whether the treatment effect is constant (homogeneous) or variable (heterogeneous). 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 article 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. We leave spatio-temporal RDDs using GP models to future research.

Table A.1. Shorthand notation for covariance matrices.

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

NOTE: The spatial coordinates of the *i*th treatment unit are denoted by $\boldsymbol{s}_{iT}$, and those of the *j*th control unit by $\boldsymbol{s}_{jC}$, while $\boldsymbol{b}_i$ denotes the *i*th sentinel location along the border.

The calibrated inverse-variance test of Section 2.4 is the special case of this procedure with weights $w_{\mathcal{B}}(\boldsymbol{b}_{1:R}) = \Sigma^{-1}_{\boldsymbol{b}_{1:R}|Y}\mathbf{1}_R$.

Appendix A: Covariances for Gaussian Process Model

All covariances below are conditional on the hyperparameters $\boldsymbol{\theta} = (\ell, \sigma_{\mathrm{GP}}, \sigma_\epsilon, \sigma_m)$, omitted for concision. We further define some shorthand notation, found in Table A.1.

$$\begin{aligned}
m_T, m_C &\sim \mathcal{N}\left(0, \sigma_m^2\right),\\
\mathrm{cov}(Y_{iT}, m_T) = \mathrm{cov}(Y_{iC}, m_C) &= \sigma_m^2,\\
\mathrm{cov}(Y_{iT}, m_C) = \mathrm{cov}(Y_{iC}, m_T) &= 0,\\
\mathrm{cov}\left(Y_{iT}, f_T(\boldsymbol{s}')\right) = \mathrm{cov}\left(Y_{iC}, f_C(\boldsymbol{s}')\right) &= k(\boldsymbol{s}_i, \boldsymbol{s}'),\\
\mathrm{cov}\left(Y_{iT}, f_C(\boldsymbol{s}')\right) = \mathrm{cov}\left(Y_{iC}, f_T(\boldsymbol{s}')\right) &= 0,\\
\mathrm{cov}(Y_{iT}, Y_{jT}) = \mathrm{cov}(Y_{iC}, Y_{jC}) &= \sigma_m^2 + k(\boldsymbol{s}_i, \boldsymbol{s}_j) + \delta_{ij}\sigma_\epsilon^2,\\
\mathrm{cov}(Y_{iT}, Y_{jC}) &= 0.
\end{aligned} \tag{A.1}$$

Appendix B: 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 (13) as a test statistic.

Under the parametric null hypothesis M_0, $\boldsymbol{Y}_T$ and $\boldsymbol{Y}_C$ are drawn from a single GP, with no discontinuity at the border. Their joint covariance is

$$\mathrm{cov}\left(\begin{pmatrix}\boldsymbol{Y}_T\\ \boldsymbol{Y}_C\end{pmatrix} \mid M_0\right) = \begin{bmatrix}\Sigma_{TT} & \boldsymbol{K}_{TC}\\ \boldsymbol{K}_{TC}^\intercal & \Sigma_{CC}\end{bmatrix}, \tag{B.1}$$

where $\boldsymbol{K}_{TC}$ is the $n_T \times n_C$ matrix with *ij*th entry equal to $k\left(\boldsymbol{s}_{iT}, \boldsymbol{s}_{jC}\right)$. 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:

$$\boldsymbol{W}_T = \boldsymbol{K}_{\mathcal{B}T}\Sigma_{TT}^{-1} \quad \text{and} \quad \boldsymbol{W}_C = \boldsymbol{K}_{\mathcal{B}C}\Sigma_{CC}^{-1}. \tag{B.2}$$

Under 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

$$\mathrm{cov}\left(\begin{pmatrix}\boldsymbol{W}_T\boldsymbol{Y}_T\\ \boldsymbol{W}_C\boldsymbol{Y}_C\end{pmatrix} \mid M_0\right) = \begin{bmatrix}\boldsymbol{W}_T\Sigma_{TT}\boldsymbol{W}_T^\intercal & \boldsymbol{W}_T\boldsymbol{K}_{TC}\boldsymbol{W}_C^\intercal\\ \boldsymbol{W}_C\boldsymbol{K}_{TC}^\intercal\boldsymbol{W}_T^\intercal & \boldsymbol{W}_C\Sigma_{CC}\boldsymbol{W}_C^\intercal\end{bmatrix}. \tag{B.3}$$

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

$$\begin{aligned}\operatorname{cov}\left(\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y} \mid M_0\right) &= \boldsymbol{W}_T\Sigma_{TT}\boldsymbol{W}_T^{\intercal} + \boldsymbol{W}_C\Sigma_{CC}\boldsymbol{W}_C^{\intercal} \\ &\quad - \boldsymbol{W}_T\boldsymbol{K}_{TC}\boldsymbol{W}_C^{\intercal} - \boldsymbol{W}_C\boldsymbol{K}_{TC}^{\intercal}\boldsymbol{W}_T^{\intercal}. \qquad \text{(B.4)}\end{aligned}$$

Weighted LATE estimators of the form defined in (8) are linear transformations of $\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}$ and so under M_0, they are normally distributed with mean zero. For a given weight function $w_{\mathcal{B}}$, its variance is given by

$$\begin{aligned}\operatorname{var}\left(\mu_{\tau^w|Y} \mid M_0\right) &= \operatorname{cov}\left(\frac{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^{\intercal}\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}}{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^{\intercal}\mathbf{1}_R}\right) \\ &= \frac{w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^{\intercal}\operatorname{cov}\left(\boldsymbol{\mu}_{\boldsymbol{b}_{1:R}|Y}\right)w_{\mathcal{B}}(\boldsymbol{b}_{1:R})}{\left(w_{\mathcal{B}}(\boldsymbol{b}_{1:R})^{\intercal}\mathbf{1}_R\right)^2}. \qquad \text{(B.5)}\end{aligned}$$

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 $\mu_{\tau^w|Y^{\text{obs}}}$ is

$$\begin{aligned}&p\left(\left|\mu_{\tau^w|Y}\right| \geq \left|\mu_{\tau^w|Y^{\text{obs}}}\right| \mid M_0\right) \\ &\quad = 2\Phi\left(-\left|\mu_{\tau^w|Y^{\text{obs}}}\right| \Big/ \sqrt{\operatorname{var}(\mu_{\tau^w|Y} \mid M_0)}\right). \qquad \text{(B.6)}\end{aligned}$$

The calibrated inverse-variance test of Section 2.4 is the special case of this procedure with weights $w_{\mathcal{B}}(\boldsymbol{b}_{1:R}) = \Sigma_{\boldsymbol{b}_{1:R}|Y}^{-1}\mathbf{1}_R$.

Supplementary Materials

(1) Why the Projected 1D RDD Approach Can Lead to Spatial Confounding, (2) Handling Nonspatial Covariates, (3) Additional LATE Estimands and Simulations, (4) Alternate Tests for Non-Zero Treatment Effect, and (5) Additional Testing Details for NYC School Districts Application.

Funding

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.

References

Angrist, J. D., and Pischke, J.-S. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton, NJ: Princeton University Press. [623]

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: CRC Press. [620,621,629]

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017), "Julia: A Fresh Approach to Numerical Computing," *SIAM Review*, 59, 65–98. [620]

Black, S. E. (1999), "Do Better Schools Matter? Parental Valuation of Elementary Education," *The Quarterly Journal of Economics*, 114, 577–599. [619]

Black, S. E., and Machin, S. (2011), "Housing Valuations of School Performance," in *Handbook of the Economics of Education* (Vol. 3), eds. E. A. Hanushek, S. Machin, and L. Woessmann, Amsterdam: Elsevier, pp. 485–519. [619,625,626]

Branson, Z., Rischard, M., Bornn, L., and Miratrix, L. W. (2019), "A Nonparametric Bayesian Methodology for Regression Discontinuity Designs," *Journal of Statistical Planning and Inference*, 202, 14–30. [620, 621]

Chen, Y., Ebenstein, A., Greenstone, M., and Li, H. (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 of the United States of America*, 110, 12936–12941. [620]

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, 636–654. [620]

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009), "Dealing With Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96, 187–199. [623]

Fack, G., and Grenet, J. (2010), "When Do Better Schools Raise Housing Prices? Evidence From Paris Public and Private Schools," *Journal of Public Economics*, 94, 59–77. [620]

Fairbrother, J., Nemeth, C., Rischard, M., and Brea, J. (2018), "Gaussianprocesses.jl: A Nonparametric Bayes Package for the Julia Language," arXiv no. 1812.09064. [620]

Gibbons, S., Machin, S., and Silva, O. (2013), "Valuing School Quality Using Boundary Discontinuities," *Journal of Urban Economics*, 75, 15–28. [620,625]

Hahn, J., Todd, P., and Van der Klaauw, W. (2001), "Identification and Estimation of Treatment Effects With a Regression-Discontinuity Design," *Econometrica*, 69, 201–209. [620]

Hahn, P. R., Carvalho, C. M., Puelz, D., and He, J. (2018), "Regularization and Confounding in Linear Regression for Treatment Effect Estimation," *Bayesian Analysis*, 13, 163–182. [628]

Hahn, P. R., Murray, J. S., and Carvalho, C. (2017), "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects," arXiv no. 1706.09523. [628,629]

Hill, J. L. (2011), "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 20, 217–240. [621]

Holmes, T. J. (1998), "The Effect of State Policies on the Location of Manufacturing: Evidence From State Borders," *Journal of Political Economy*, 106, 667–705. [620]

Imbens, G. W., and Lemieux, T. (2008), "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142, 615–635. [620,621]

Imbens, G. W., and Zajonc, T. (2011), "Regression Discontinuity Design With Multiple Forcing Variables," Report, Harvard University, 972. [620, 621,623]

Keele, L. J., Lorch, S., Passarella, M., Small, D., and Titiunik, R. (2017), (Chapter 4), in *An Overview of Geographically Discontinuous Treatment Assignments With an Application to Children's Health Insurance*, eds. Matias D. Cattaneo and Juan Carlos Escanciano, Bingley, UK: Emerald Publishing Limited, pp. 147–194. [620]

Keele, L. J., and Titiunik, R. (2015), "Geographic Boundaries as Regression Discontinuities," *Political Analysis*, 23, 127–155. [620,621,623,626]

Keele, L. J., Titiunik, R., and Zubizarreta, J. R. (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, 178, 223–239. [620]

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018), "Balancing Covariates via Propensity Score Weighting," *Journal of the American Statistical Association*, 113, 390–400. [623]

MacDonald, J. M., Klick, J., and Grunwald, B. (2015), "The Effect of Private Police on Crime: Evidence From a Geographic Regression Discontinuity Design," *Journal of the Royal Statistical Society*, Series A, 179, 831–846. [620]

Magruder, J. R. (2012), "High Unemployment Yet Few Small Firms: The Role of Centralized Bargaining in South Africa," *American Economic Journal: Applied Economics*, 4, 138–166. [620]

Miratrix, L. W., Weiss, M. J., and Henderson, B. (2020), "An Applied Researcher's Guide to Estimating Effects From Multisite Individually Randomized Trials: Estimands, Estimators, and Estimates," *Journal of Research on Educational Effectiveness*, to appear. [623]

Papay, J. P., Willett, J. B., and Murnane, R. J. (2011), "Extending the Regression-Discontinuity Approach to Multiple Assignment Variables," *Journal of Econometrics*, 161, 203–207. [620]

Rasmussen, C. E., and Williams, C. K. (2006), *Gaussian Processes for Machine Learning* (Vol. 1), Cambridge, MA: MIT Press. [620,621]

Rosen, S. (1974), “Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition,” *Journal of Political Economy*, 82, 34–55. [619,625]

Rosenbaum, P. R. (2010), *Design of Observational Studies*, Springer Series in Statistics, New York: Springer. [621]

Rubin, D. B. (1974), “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” *Journal of Educational Psychology*, 66, 688–701. [621]

Sheppard, S. (1999), “Hedonic Analysis of Housing Markets,” *Handbook of Regional and Urban Economics*, 3, 1595–1635. [619,625]

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. (1923/1990), “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9,” *Statistical Science*, 5, 465–472. [621]

Stein, M. L. (2012), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer. [629]

Thistlethwaite, D. L., and Campbell, D. T. (1960), “Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment,” *Journal of Educational Psychology*, 51, 309. [620,621]

Wilson, A., and Adams, R. (2013), “Gaussian Process Kernels for Pattern Discovery and Extrapolation,” in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1067–1075. [629]

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, 1360–1371. [620]