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:

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

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

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

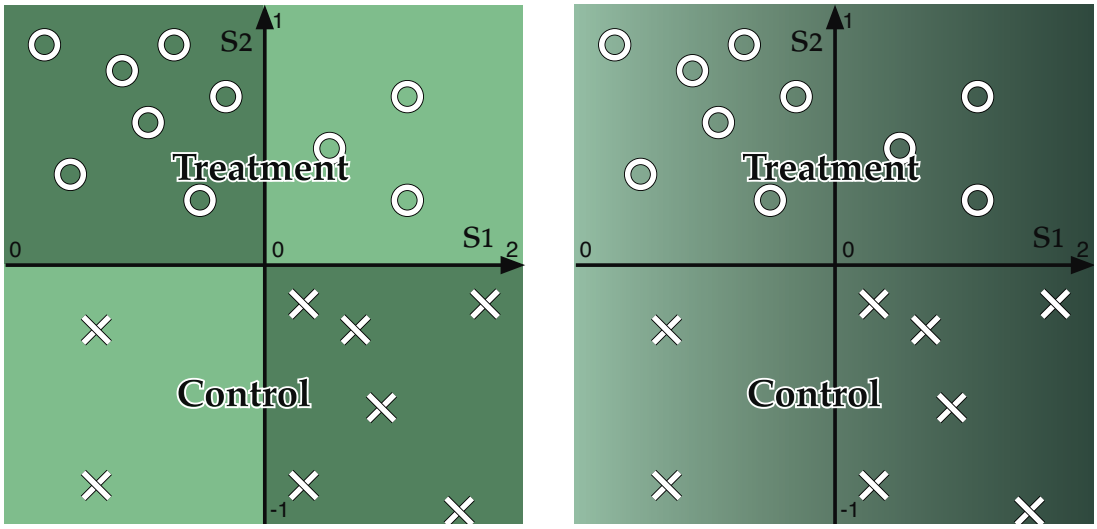

Figure S-1: A theoretical example illustrating the susceptibility of the projected 1D RDD method to spatial confounding. On the left, the background shows the density of units, while on the right it shows the linear spatial trend in outcomes. The locations of treatment and control units are shown with 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.

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 τ^{UNIF}, population-weighted τ^{ρ}, inverse-weighted τ^{INV}, and finite population projected τ^{PROJ}. We here present two other estimands of interest not directly presented in the paper, which extend the pro-

jection population idea of τ^{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 τ^{UNIF} places uniform weights along the border. Instead, the "geography-based" projected land LATE estimand τ^{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 Δ of the border $\mathcal{B}$, but then projects them onto the border to derive border weights. In other words, the projection method from τ^{PROJ} is applied to an infinite population of uniform density on both sides of the border, instead of the finite population of observed units.

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 τ^{GEO}, a tight grid G^ν of evenly spaced points separated by ν is first generated covering $\mathcal{A}_\Delta$. Denote the number of grid points by L_ν. Each point G^ν_l, $l = 1, \ldots, L_\nu$ in G^ν is then projected onto the border to become a sentinel. The treatment effect at these positions is then estimated as before, yielding a mean vector and covariance matrix akin to (6). The mean of the mean vector then gives an estimate of τ^{GEO}. In other words, τ^{GEO} is estimated by applying the τ^{UNIF} procedure with sentinels obtained by projecting the grid points, instead of equispaced sentinels. τ^{GEO} remains in the category of weighted-mean estimands, with the weight function $w_{\mathcal{B}}(\boldsymbol{b})$ in (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

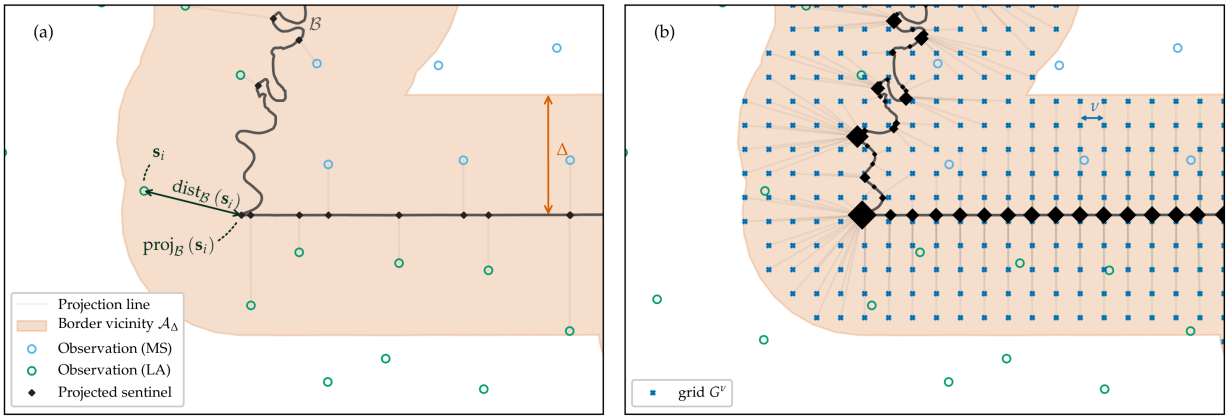

Figure S-2: Illustration of (a) projected finite-population LATE τ^{PROJ}, and (b) projected land LATE τ^{GEO}, using the border separating Mississippi and Louisiana near Baton Rouge, with units at the centroid of each county. The border vicinity $\mathcal{A}_\Delta$ is defined as all land within $\Delta = 50\,\text{km}$ 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.

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 τ^{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 τ^{GEO} can be modified by weighting the grid points G_l^ν, $l = 1, \ldots, L_\Delta$ by the population density $\rho(G_l^\nu)$. This gives the projected superpopulation LATE

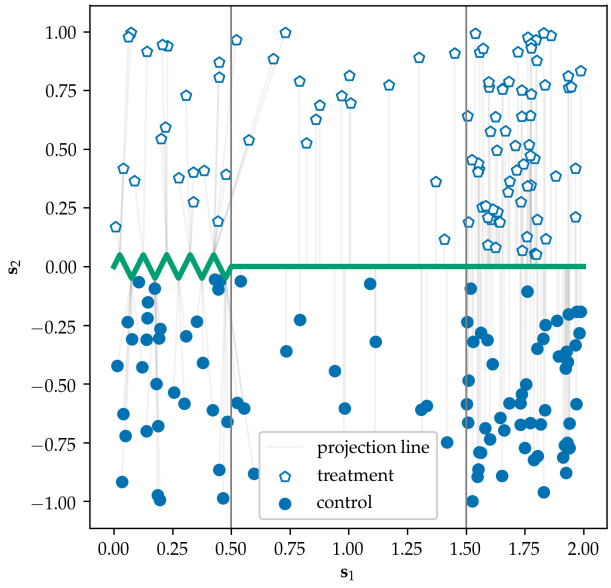

Figure S-3: Spatial positions of units and border for the wiggly border simulation of Section S-2.3. Projection lines for the projected finite population LATE are shown in light gray.

τ^{POP}. Similarly to the density-weighted LATE τ^{ρ}, estimating τ^{POP} requires an estimate of the density $\rho(G_l^{\nu})$ at every grid point. As before, the uncertainty in the estimate of ρ should in principle be propagated to the estimate of τ^{POP}, which generally will make the posterior distribution of τ^{POP} neither normal nor analytically tractable.

The estimand τ^{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 $\boldsymbol{s}_2 = 0$

	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.

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.

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 Δ 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.

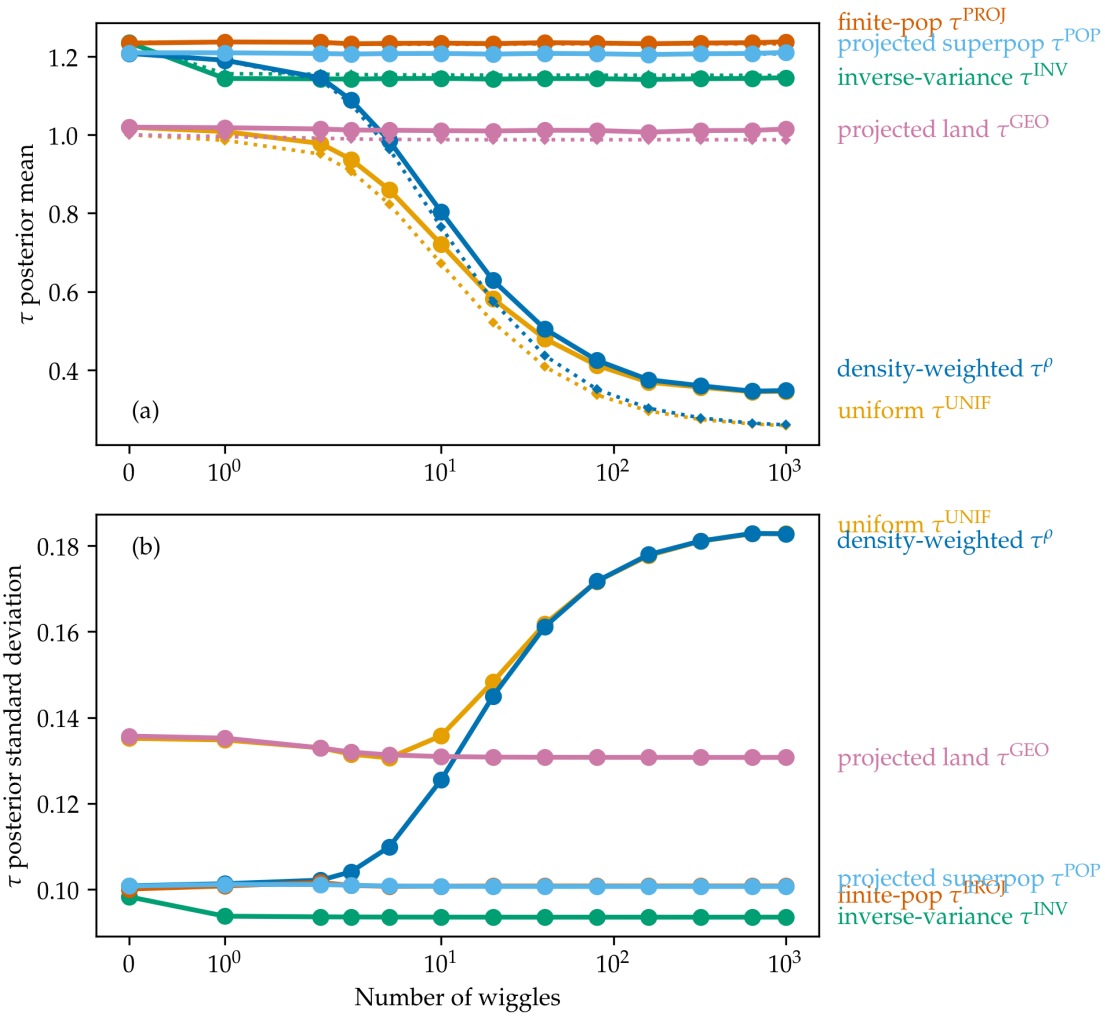

Figure S-4: Results of the simulations of Section S-2.3, showing for each LATE estimator as the leftmost section of the border gets wigglier (a) the estimate (posterior mean) averaged over 10,000 simulations with the corresponding estimand shown as a dotted line of the same color, and (b) the posterior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

When the border is a straight line and because $\mathcal{A}_T$ and $\mathcal{A}_C$ are rectangles, the density-weighted estimand τ^{ρ} equals the projected superpopulation estimand τ^{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 τ^{ρ} and τ^{POP} are identical. τ^{POP} and the finite-population projected LATE τ^{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 τ^{UNIF} and τ^{GEO} are also equivalent when the

border is a straight line. They give equal weight to the sparsely populated middle band, which produces a lower estimate with higher variance than the posteriors of τ^{ρ} and τ^{POP}.

Lastly, the information-based inverse-variance estimand τ^{INV} does not 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, τ^{ρ} and τ^{UNIF} show their susceptibility to the border topology. Proportionally more sentinels are packed into the leftmost section of the border, upweighting the lower treatment effect of that band, and resulting in a drop of the two estimates and estimands. Meanwhile, τ^{INV} remains stable despite the wiggles, because the additional sentinels in the leftmost band get automatically downweighted as their correlation rises. The estimators that rely on projection τ^{PROJ}, τ^{GEO}, and τ^{POP} also remain stable, because the projected sentinels hardly move.

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 τ^{UNIF}, τ^{ρ}, and τ^{INV}) are more densely packed along $\boldsymbol{s}_1$ in the leftmost area because of the zig-zagging border. The inverse-variance weighted estimator border weights can be seen to respond to this change in the border topology, though it is difficult to interpret their oscillating behavior. While these border-weights look unreasonable and unstable, the induced unit weights for τ^{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 τ^{UNIF} and τ^{GEO} manifests itself as large weights given to a small number of units. All other estimators spread the weights more evenly amongst the units near the

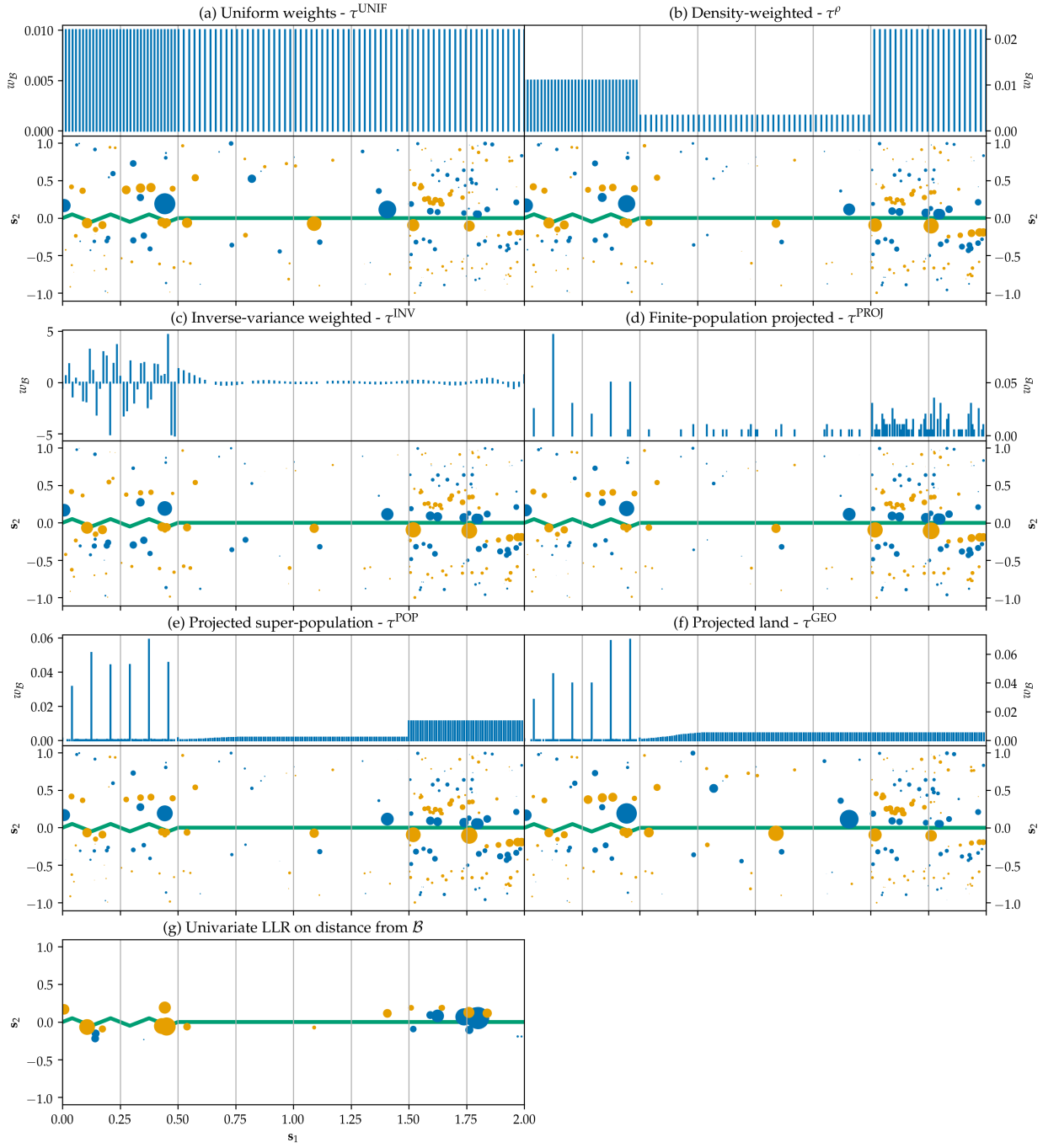

Figure S-5: Weight functions and induced weights on the observations for the six weight functions proposed in this paper. The weight function plots show the weight $w_{\mathcal{B}}(\boldsymbol{b})$ against each sentinel's $\boldsymbol{s}_1$ coordinate. Sentinels with coinciding or nearly coinciding (within 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.

border, which reduces their variance.

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^{th} diagonal element is the triangular kernel evaluated on the i^{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_{wiggles}	$\widehat{\tau^{\text{UNIF}}}$	τ^{UNIF}	$\widehat{\tau^{\text{INV}}}$	τ^{INV}	$\widehat{\tau^{\rho}}$	τ^{ρ}	$\widehat{\tau^{\text{PROJ}}}$	τ^{PROJ}	$\widehat{\tau^{\text{GEO}}}$	τ^{GEO}	$\widehat{\tau^{\text{POP}}}$	τ^{POP}
0	1.02 (0.14)	1.00	1.24 (0.10)	1.23	1.21 (0.10)	1.21	1.23 (0.10)	1.24	1.02 (0.14)	1.00	1.21 (0.10)	1.21
1	1.01 (0.13)	0.99	1.14 (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:

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

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 χ^2_k. See for example Rencher (2003) Section 5.2.2 for a classical derivation of this test. This suggests that we could use $S^2 = \boldsymbol{\mu}^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$ as a test statistic, and obtain a p-value from a χ^2_R distribution function evaluated at S^2, where R is the number of sentinels. However, we face two problems. Firstly, this test, obtained heuristically from a Bayesian posterior by analogy with the classical multivariate normal result, is not a valid frequentist test. Secondly, while $\boldsymbol{\Sigma}$ 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 $\mathbf{\Sigma}$ eigenvalues λ_i lower than $\epsilon \sum_{j=1}^{k} \lambda_j$, with ϵ a pre-specified small number (they use 0.01). They address the first problem by showing that the resulting p-value is always conservative in their simulations. However, in our work, we found the resulting p-value to be sensitive to the arbitrarily chosen ϵ tolerance parameter, which makes it difficult to trust its validity.

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

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

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 $\mathbf{\Sigma}$ 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 $\mathbf{\Sigma}$ 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 τ to all the outcomes in Louisiana. The results of the three tests proposed so far are shown in the first three rows of Table S-3 for $\tau = 0$ (null hypothesis) and $\tau = 1.2$ and significance level $\alpha = 0.05$.

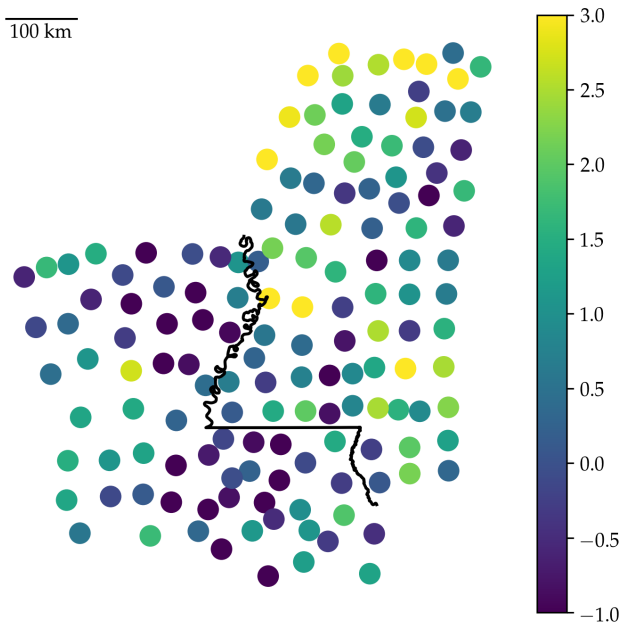

Figure S-6: Set-up of the imaginary experiment in Louisiana and Mississippi. Each unit is at the centroid of a county. The colors indicated the observed outcomes in one draw of the simulation under $\tau = 1.5$. In this particular run, the p-values were 0.0016, 0.0018, and 0.0013 for the mLL, χ^2, and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

We see that under the null, the χ^2 and likelihood ratio tests are valid (rejection of the null in 5% of simulations up to simulation error). This is enforced by the parametric bootstrap, which draws test statistics from the same null distribution to calibrate the tests. However, the p-values for the inverse-variance test are biased down, so that we will falsely reject the null 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 χ^2 tests. This can lead to a paradox: we may reject the weak null hypothesis, but fail to reject the sharp null hypothesis (using the χ^2 or likelihood test), even though rejection of the weak null 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).

Test	Power under $\tau = 0$	Power under $\tau = 1.2$
Marginal log-likelihood bootstrap	0.05	0.72
χ^2 bootstrap	0.05	0.63
τ^{INV} uncalibrated	0.09	0.87
τ^{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
χ^2 bootstrap	0.022
τ^{INV} uncalibrated	0.0007
τ^{INV} calibrated	0.002

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

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 three tests, though the χ^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 $1\,^\circ$, $3\,^\circ$, $5\,^\circ$, $6\,^\circ$, ..., $179\,^\circ$. 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

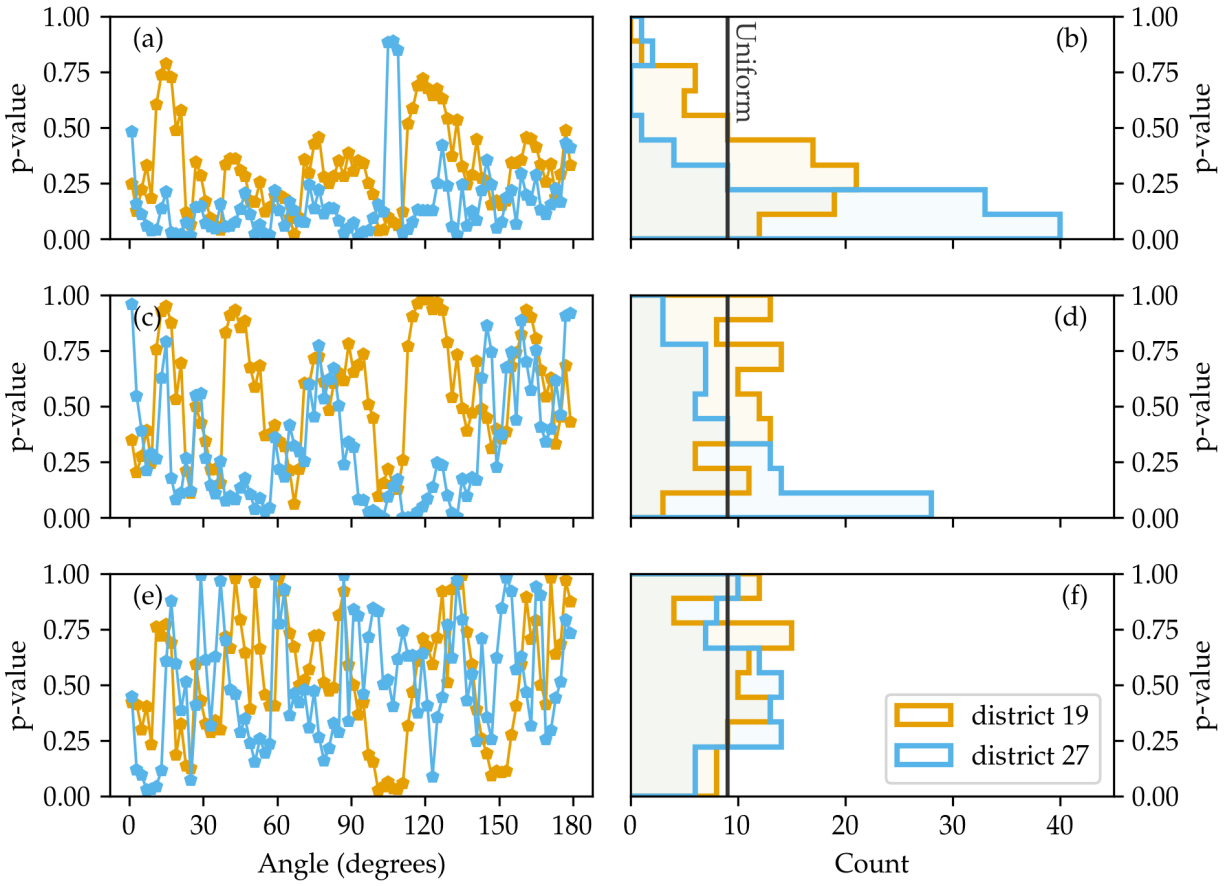

Figure S-7: Placebo tests for significance tests applied to NYC school district house prices, applied within districts 19 and 27. From top to bottom, results are shown for the marginal log-likelihood bootstrap test, chi-squared bootrap test, and calibrated inverse-variance test. The first column shows 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.

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 χ^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 χ^2 test will continue to be reliable under misspecification. It is only due to its low power that we hesitate to recommend its use in applications where the treatment effect is expected to be fairly homogenous.

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

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).

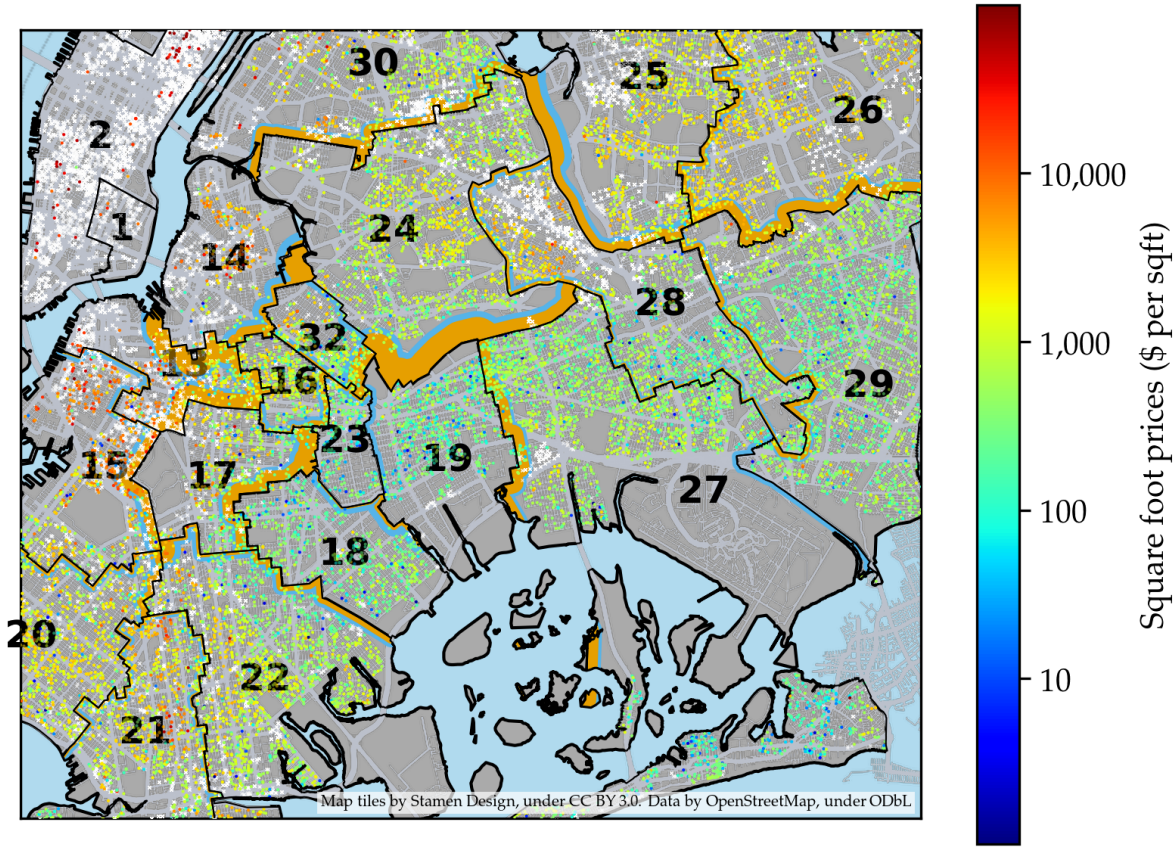

Figure S-8: Pairwise estimates of the inverse variance LATE between adjacent districts. The thickness of the orange buffer adjacent to borders is proportional to the posterior mean of the inverse variance LATE, and the blue buffer beyond it is proportional to the posterior standard deviation of the LATE.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

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

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.

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

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

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.