

GeoRDD manuscript

Maxime Rischard

April 19, 2017

Contents

1	Introduction	2
1.1	Motivation	2
1.2	Prior attempts	2
2	Model Specification	2
2.1	Notation	2
2.2	1GP solution	2
2.3	2GP solution	2
2.4	Discussion	3
3	Inference	3
3.1	1GP	3
3.2	2GP	4
4	Handling covariates	4
5	Average treatment effect	4
6	2GP: Testing for non-zero effect	6
6.1	Using the inverse-variance weighted mean treatment effect posterior to test the weak null hypothesis	6
6.2	Likelihood-based sharp null test	6
6.3	χ^2 test for the sharp null	7
6.4	Power in simulated example	8
6.5	Placebo tests	9
7	Spatial advantage	9
8	Example: NYC school districts	10
8.1	Preprocessing	10
8.2	Exploratory analysis	10
8.3	Model for property prices	10
8.4	parameter optimization	12
8.5	cliff face	12
8.5.1	surface plots	12
8.6	average effect	12
8.7	significance tests	12
8.7.1	placebo tests	12
8.8	pairwise treatment effect (all districts)	12
9	Conclusion	12

1 Introduction

1.1 Motivation

1.2 Prior attempts

2 Model Specification

2.1 Notation

- 2-dimensional coordinate space \mathcal{S}
- treatment units are in region $\mathcal{S}_T \subset \mathcal{S}$ and control units are in non-overlapping \mathcal{S}_C outside of the treatment region, so that $\mathcal{S}_C = \mathcal{S}_T^c$ and $\mathcal{S}_T \cup \mathcal{S}_C = \mathcal{S}$
- Observed outcomes for units in treatment region $s \in \mathcal{S}_T$ are labeled $Y_T(s)$, and units in control region $Y_C(s)$.
- Potential outcomes framework: Each unit has a potential outcome under treatment $Y_T(s)$ and a potential outcome under control $Y_C(s)$. If $s \in \mathcal{S}_T$, then $Y_T(s)$ is observed, otherwise $Y_C(s)$ is observed.

2.2 1GP solution

Most straightforwardly, we model the observed outcomes Y at locations S as the sum of an intercept μ , linear trend $S\beta$, a spatial Gaussian process $f(S)$, a constant treatment effect τ in the treatment region, and iid normal noise ϵ .

$$\begin{aligned}
 Y_i(s) &= \mu + s^\top \beta + f(s) + \tau \mathbb{I}\{s \in \mathcal{S}_T\} + \epsilon_i \\
 f(S) &\sim \mathcal{GP}(0, k(s, s')) \\
 k(s, s') &= \sigma_{GP}^2 \exp\left(-\frac{(s - s')^\top (s - s')}{2\ell^2}\right) \\
 \epsilon_i &\stackrel{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)
 \end{aligned} \tag{1}$$

$f(S)$ is a smooth surface covering all of \mathcal{S} , specified as a Gaussian Process with squared exponential covariance kernel k with lengthscale ℓ and variance σ_{GP}^2 . The squared exponential kernel is frequently used in spatial settings. The constant treatment effect implies the assumption that $Y_T(s) = \tau + Y_C(s)$ for all units at all locations.

2.3 2GP solution

The constant treatment effect is a strong assumption that will be hard to justify in many applications. To allow the treatment effect to vary spatially, an alternative is to specify two independent Gaussian processes for the treatment response and the control response.

$$\begin{aligned}
 Y_{T,i}(s) &= \underbrace{\mu_T + s^\top \beta_T + f_T(s)}_{g_T(s)} + \epsilon_i \\
 Y_{C,i}(s) &= \underbrace{\mu_C + s^\top \beta_C + f_C(s)}_{g_C(s)} + \epsilon_i \\
 f_T(S), f_C(S) &\stackrel{\perp}{\sim} \mathcal{GP}(0, k(s, s')) \\
 k(s, s') &= \sigma_{GP}^2 \exp\left(-\frac{(s - s')^\top (s - s')}{2\ell^2}\right)
 \end{aligned} \tag{2}$$

Here, the treatment effect τ is no longer included explicitly in the model. Instead, the treatment effect at a location s is derived as the difference between the two (noise-free) surfaces g_T and g_C .

$$\tau(\mathbf{s}) = [\mu_T + \mathbf{s}^\top \beta_T + f_T(\mathbf{s})] - [\mu_C + \mathbf{s}^\top \beta_C + f_C(\mathbf{s})] \quad (3)$$

In this specification, the kernel parameters ℓ and σ_{GP} are the same in the treatment and control regions, so we assume that the spatial smoothness of the responses isn't affected by the treatment. This assumption will be reasonable in most applications, but can be easily relaxed. Inference on the hyperparameters proceeds as in the 1GP case, using the sum of the likelihood in the treatment and control regions.

2.4 Discussion

- different assumptions
- will stick to 2GP from now on

3 Inference

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

3.1 1GP

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

$$\begin{aligned} \tau &\sim \mathcal{N}(0, \sigma_\tau^2) \\ \mu &\sim \mathcal{N}(0, \sigma_\mu^2) \\ \beta &\sim \mathcal{N}(0, \sigma_\beta^2) \\ \text{cov}(Y_i(\mathbf{s}), \tau) &= \sigma_\tau^2 \mathbb{I}\{\mathbf{s} \in \mathcal{S}_T\} \\ \text{cov}(Y_i(\mathbf{s}), \mu) &= \sigma_\mu^2 \\ \text{cov}(Y_i(\mathbf{s}), \beta) &= \sigma_\beta^2 \mathbf{s}^\top \mathbf{s} \\ \text{cov}(Y_i(\mathbf{s}), Y_i(\mathbf{s}')) &= \sigma_\mu^2 + \sigma_\tau^2 \mathbb{I}\{\mathbf{s} \in \mathcal{S}_T\} \mathbb{I}\{\mathbf{s}' \in \mathcal{S}_T\} + \sigma_\beta^2 \mathbf{s}^\top \mathbf{s}' + k(\mathbf{s}, \mathbf{s}') + \delta_{ij} \sigma_\epsilon^2 \\ \text{cov}(Y(\mathbf{s}), f(\mathbf{s}')) &= \text{cov}(f(\mathbf{s}), f(\mathbf{s}')) = k(\mathbf{s}, \mathbf{s}') \end{aligned} \quad (4)$$

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

$$\tau \mid Y, \ell, \sigma_{\text{GP}} \sim \mathcal{N}\left(\text{cov}(Y, \tau)^\top \text{cov}(Y)^{-1} Y, \sigma_\tau^2 - \text{cov}(Y, \tau)^\top \text{cov}(Y)^{-1} \text{cov}(Y, \tau)\right) \quad (5)$$

To proceed computationally, we define the treatment indicator vector \mathbb{I}_T with i th entry equal to 0 when \mathbf{s}_i is in the control region, and 1 in the treatment region, and the $n \times n$ kernel covariance matrix \mathbf{K} having entries $\mathbf{K}_{ij} = k(\mathbf{s}_i, \mathbf{s}_j)$. The posterior mean and variance are then easily computed.

$$\begin{aligned} \mathbb{E}(\tau \mid Y, \ell, \sigma_{\text{GP}}, \sigma_\epsilon) &= \sigma_\tau^2 \mathbb{I}_T^\top \left\{ \sigma_\mu^2 + \sigma_\tau^2 \mathbb{I}_T \mathbb{I}_T^\top + \sigma_\beta^2 S S^\top + \mathbf{K} + \sigma_\epsilon^2 \mathbf{I} \right\}^{-1} Y \\ \text{var}(\tau \mid Y, \ell, \sigma_{\text{GP}}, \sigma_\epsilon) &= \sigma_\tau^2 - \sigma_\tau^2 \mathbb{I}_T^\top \left\{ \sigma_\mu^2 + \sigma_\tau^2 \mathbb{I}_T \mathbb{I}_T^\top + \sigma_\beta^2 S S^\top + \mathbf{K} + \sigma_\epsilon^2 \mathbf{I} \right\}^{-1} \mathbb{I}_T \end{aligned} \quad (6)$$

What remains is the inference on the hyperparameters $\sigma_\epsilon, \sigma_{\text{GP}}$ and ℓ . The two approaches typically taken in modern spatial statistics are either to maximize the marginal likelihood of Y as a function of those

three parameters, or to assign them a prior and take a Bayesian approach, requiring that the posterior of τ be integrated over those parameters. The compromise is clear: the Bayesian approach incorporates the uncertainty in the hyperparameters, thus giving more reliable inference on τ , but maximizing the marginal likelihood has a much lower computation cost. Therefore, we recommend taking the Bayesian approach whenever computationally possible, and maximizing the marginal likelihood when the data is larger.

3.2 2GP

In the 2GP setting, we begin by modeling the treatment and control units with two independent Gaussian processes with shared hyperparameters. Because the treatment and control regions do not overlap, inference on the treatment effect is only measurable near the boundary. In the classical one-dimensional regression discontinuity design, the estimand is therefore defined at the boundary $x = b$:

$$\tau = \lim_{x \downarrow b} \mathbb{E}[y \mid X = s] - \lim_{x \uparrow b} \mathbb{E}[y \mid X = x] = \mathbb{E}[Y_T \mid X = b] - \mathbb{E}[Y_C \mid X = b] \quad (7)$$

Analogously, we focus on the treatment effect at the boundary ∂ between the treatment and control regions. ∂ is therefore a one-dimensional subset of \mathcal{S} . We will proceed by extrapolating both Gaussian processes to the boundary, and then subtracting the predictions to obtain the estimated treatment effect. Computationally, we need to represent this boundary as a set of k "sentinel" units distributed along the boundary $\partial = \{\partial_1, \dots, \partial_k\}$, $\partial_i \in \partial$. The extrapolation step then proceeds mechanically through multivariate-normal theory.

$$\begin{aligned} g_T(\partial) \mid Y_T, S_T, \ell, \sigma_{GP}, \sigma_\epsilon &\sim \mathcal{N}(\mu_{\partial|T}, \Sigma_{\partial|T}) \\ \mu_{\partial|T} &\equiv \text{cov}(g_T(\partial), Y_T) \text{cov}(Y_T)^{-1} Y_T \\ \Sigma_{\partial|T} &\equiv \text{cov}(g_T(\partial)) - \text{cov}(g_T(\partial), Y_T) \text{cov}(Y_T)^{-1} \text{cov}(Y_T, g_T(\partial)) \end{aligned} \quad (8)$$

All the covariance terms can be derived from the model similarly to what we saw in the 1GP procedure. Analogously, we also generate predictions for $g_C(\partial)$ using the data in the control region, and denote their posterior mean and covariance as $\mu_{\partial|C}$ and $\Sigma_{\partial|C}$. Since the two surfaces are modeled as independent, the treatment effect $\tau(\partial) = g_T(\partial) - g_C(\partial)$ along the boundary is also multivariate normal with posterior mean and covariance

$$\begin{aligned} \mu_{\partial|Y} &= \mathbb{E}(\tau(\partial) \mid Y_T, Y_C) = \mu_{\partial|T} - \mu_{\partial|C} \\ \Sigma_{\partial|Y} &= \text{cov}(\tau(\partial) \mid Y_T, Y_C) = \Sigma_{\partial|T} + \Sigma_{\partial|C}. \end{aligned} \quad (9)$$

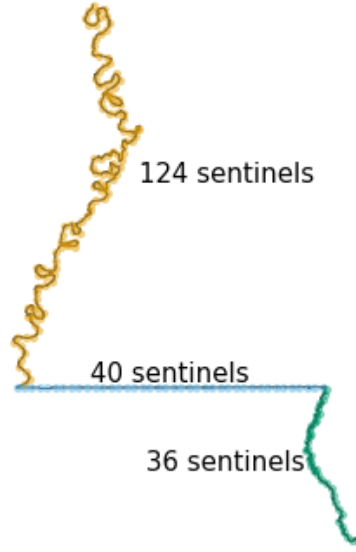
4 Handling covariates

The Gaussian Process specification makes it easy to incorporate a linear model on non-spatial covariates, both mathematically and computationally. The model is modified by the addition of the linear regression term $D\gamma$ on the $n \times p$ matrix of covariates D . In the spirit of ridge regression, we recommend placing a normal prior $\mathcal{N}(0, \sigma_\gamma^2)$ on the regression coefficients. This preserves the multivariate normality of the problem, with the simple addition of a term $\sigma_\gamma^2 D^T D$ to the covariance of Y .

With the 1GP model, covariates can therefore be handled at very little additional cost, except that the additional hyperparameter σ_γ^2 needs to be fitted.

5 Average treatment effect

Once we obtain the posterior on the treatment effect function $\tau(\partial)$, estimating the average treatment effect along the boundary will often be of interest. Most straightforwardly, if the sentinels are evenly spaced, we can estimate $\bar{\tau}$, the mean of $\tau(s)$ along the boundary, by averaging the entries of the mean posterior at the



mississippi counts

sentinels. If the sentinels are not evenly spaced, then each entry needs to be re-weighted by the length of the border that the sentinel occupies.

$$\begin{aligned}
 \bar{\tau} &\equiv \frac{\oint_{\partial} \tau(x) dx}{\oint_{\partial} dx} \\
 \bar{\tau} \mid Y_T, Y_C, \sigma_{GP}, \sigma_e, \ell &\sim \mathcal{N} \left(\mu_{\bar{\tau} \mid Y}, \Sigma_{\bar{\tau} \mid Y} \right) \\
 \mu_{\bar{\tau} \mid Y} &\approx \left(\mathbf{1}^\top \mu_{\partial \mid Y} \right) / n_{\partial} \\
 \Sigma_{\bar{\tau} \mid Y} &\approx \left(\mathbf{1}^\top \Sigma_{\partial \mid Y} \mathbf{1} \right) / n_{\partial}^2
 \end{aligned} \tag{10}$$

This procedure is mathematically sound, but the choice of the $\bar{\tau}$ estimand raises two problems. Firstly, parts of the border adjoining dense populations are given equal weights to those in sparsely populated areas. If the border goes through an unpopulated area, like a lake or a public park, then the treatment effect there has little meaning and importance. Furthermore, $\tau(s)$ in those areas will have large posterior variances, which will dominate the posterior variance of $\bar{\tau}$, making otherwise large treatment effects difficult to detect.

Secondly, the unweighted mean treatment estimand is affected by the shape of the border between the treatment and control regions. We illustrate this with the border separating two American States: Louisiana and Mississippi. From North to South, the border follows the meandering Mississippi river, then takes a sharp turn to the East and becomes a straight line, until it meets the even more sinuous Pearl river, which it then follows until it reaches the Gulf of Mexico. Sentinels placed at constant intervals along this interval will therefore be most densely packed along the Pearl River, and sparsest along the straight segment of the border (see Figure ??). When averaging a function over the border, those sections will therefore be overrepresented. Troublingly, the sinuousness of the border therefore determines the estimand, and the resolution of our map can drastically change our estimate, even though the outcomes of the treatment we are studying might have nothing to do with river topographies.

Weighing the treatment effect at each sentinel location by a local density estimate would address the first issue, but not the second. We view the unwelcome dependence of the $\bar{\tau}$ estimand on the border topography as a side effect of ignoring the fact that the 1-dimensional treatment function $\tau(\partial)$ is embedded in a Euclidean 2-dimensional space. This fact is captured by the covariance structure: sentinels in the straight segment of the border will be less strongly correlated than in the sinuous segments. The more correlated sentinels individually carry less information about the local treatment effect. This suggests that instead of averaging the treatment effect evenly along the border, we wish to average evenly the information contained therein. This motivates the use of the inverse-variance weighted mean τ^{IV} , which efficiently extracts the information from the posterior to produce the weighted average with minimum variance.

$$\begin{aligned}\tau^{IV} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell &\sim \mathcal{N}\left(\mu_{\tau^{IV}|Y}, \Sigma_{\tau^{IV}|Y}\right) \\ \mu_{\tau^{IV}|Y} &\approx \left(\mathbf{1}^\top \Sigma_{\partial|Y}^{-1} \mu_{\partial|Y}\right) / \left(\mathbf{1}^\top \Sigma_{\partial|Y}^{-1} \mathbf{1}\right) \\ \Sigma_{\tau^{IV}|Y} &\approx 1 / \left(\mathbf{1}^\top \Sigma_{\partial|Y}^{-1} \mathbf{1}\right)\end{aligned}\tag{11}$$

This estimator will automatically give more weight to sentinels in dense areas (as the variance will be lower there), and to sentinels in straight sections of the border. While the estimand is less clear, the approach is in keeping with the philosophy of regression discontinuity designs. We let information be our guide when averaging over our boundary, just like it guided the analysis of regression discontinuity designs to only focus on the treatment effect at the boundary. The estimand isn't chosen by the scientist, but it is dictated by the limitations of the data.

6 2GP: Testing for non-zero effect

Following the 2GP procedure, we might naturally wonder whether we can claim to have detected a significant treatment effect anywhere along the boundary. In the hypothesis testing framework, we have two possible choices of null hypotheses. The **sharp null** specifies that the treatment effect is zero everywhere along the boundary: $\tau(\partial) = 0$, while the **weak null** only requires the average treatment effect to be zero.

6.1 Using the inverse-variance weighted mean treatment effect posterior to test the weak null hypothesis

As we saw in the previous section, the “average” treatment effect can be defined in multiple ways. If we choose the inverse-variance weighted mean, then τ^{IV} has posterior given by (11). While the posterior is a Bayesian object, we can use it heuristically to derive a pseudo- p -value

$$\begin{aligned}Z_0 &\sim \mathcal{N}\left(0, \Sigma_{\tau^{IV}|Y}\right) \\ p^{\text{INV}} &= \mathbb{P}\left(|Z_0| > \left|\mu_{\tau^{IV}|Y}\right|\right) \\ &= 2\Phi\left(-\frac{\left|\mu_{\tau^{IV}|Y}\right|}{\sqrt{\Sigma_{\tau^{IV}|Y}}}\right)\end{aligned}\tag{12}$$

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

6.2 Likelihood-based sharp null test

We can also target the sharp null hypothesis. We first create a null model \mathcal{M}_0 , specified as a single Gaussian process spanning the control and treatment regions, with the same kernel and hyperparameters obtained

in the 2GP procedure. \mathcal{M}_0 is smooth and continuous at the boundary, and therefore accords with the sharp null hypothesis. Intuitively, if there is a treatment effect, the likelihood of the observations should be lower under \mathcal{M}_0 than under \mathcal{M}_1 , the 2GP model as specified in equation (2). We therefore choose the difference in log-likelihoods as our test statistic

$$t = \log \mathbb{P}(Y_T, Y_C \mid \mathcal{M}_1) - \log \mathbb{P}(Y_T, Y_C \mid \mathcal{M}_0) \quad (13)$$

and wish to reject the sharp null hypothesis when its observed value t_{obs} is high.

A parametric bootstrap approach is used to quantify what “high” means. We draw Y_T^*, Y_C^* from \mathcal{M}_0 , using the same spatial locations as in the original data, and then fit the two competing models to the simulated data in order to obtain the bootstrapped test statistic

$$t^* = \log \mathbb{P}(Y_T^*, Y_C^* \mid \mathcal{M}_1) - \log \mathbb{P}(Y_T^*, Y_C^* \mid \mathcal{M}_0) \quad (14)$$

Repeating this procedure, we obtain a distribution of t under \mathcal{M}_0 , which we can then compare to the observed t . More precisely, we can interpret the proportion of t^* drawn above t_{obs} as a p -value.

$$p^{\text{lik}} = \mathbb{P}(t^* > t_{obs} \mid \mathcal{M}_0) \quad (15)$$

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.

6.3 χ^2 test for the sharp null

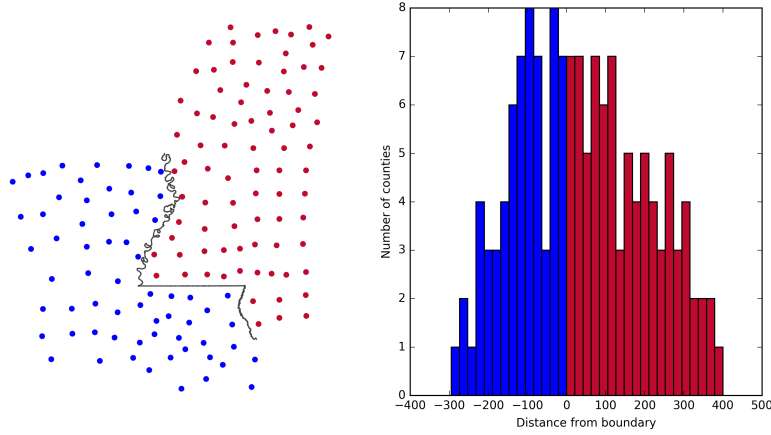
The likelihood-based sharp null above is valid and easy to understand. But it may seem odd that the test aims to detect a non-zero treatment effect at the boundary, without any explicit reference to the boundary ∂ . The test statistic and p -values can be computed without access to the sentinel positions, using only the treatment and control indicators.

To address this oddity, we can derive a test statistic directly from the posterior treatment effect along the boundary, approximated in equation (9) by its mean vector $\mu_{\partial|Y}$ and covariance matrix $\Sigma_{\partial|Y}$ at the sentinel positions ∂ . We will use μ and Σ as shorthand throughout this section. If a k -vector y has multivariate distribution $\mathcal{N}(0, \Sigma)$, then $y^T \Sigma^{-1} y$ has distribution χ_k^2 . This suggests that we could use $S = \mu^T \Sigma^{-1} \mu$ as a test statistic, and obtain a p -value by comparing it to a χ_k^2 distribution, where k is the number of sentinels. However, we face two problems. Firstly, this test derived heuristically from a Bayesian posterior is invalid from a frequentist perspective. Secondly, while Σ is mathematically full-rank, it is typically numerically rank-deficient. Therefore, k overestimates the true degrees of freedom of Σ , which invalidates the test.

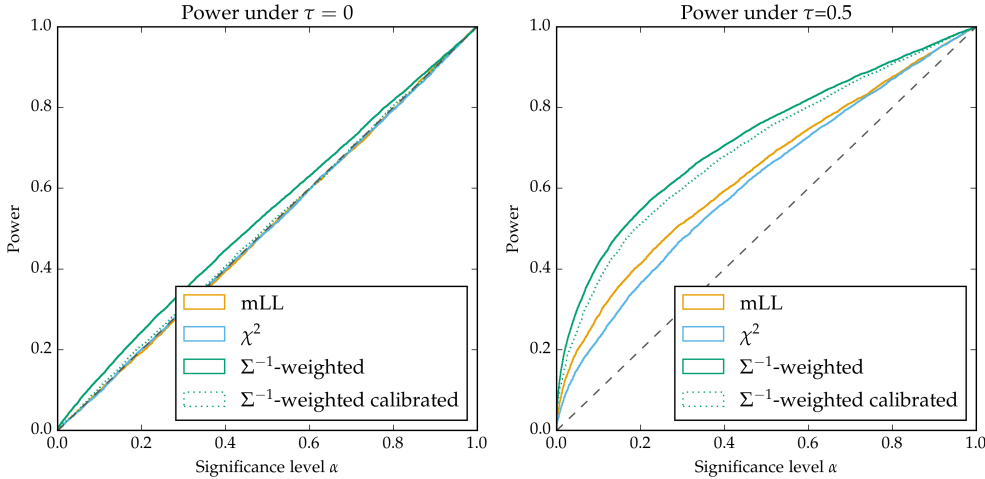
Benavoli and Mangili (2015), developing a test for function equality, address the second problem by trimming the Σ eigenvalues λ_i lower than $\epsilon \sum_{j=1}^k \lambda_j$, with ϵ a pre-specified small number (they use 0.01). They address the first problem by showing that the resulting p -value is conservative in certain simulation settings. 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 believe its validity.

We therefore again take the parametric bootstrap approach, using S as a test statistic rather than the likelihood ratio, and relying on Julia’s matrix division polyalgorithm to ensure numerical stability. We checked that adding a small constant to the diagonal of Σ does not greatly affect the computed S . Furthermore, the parametric bootstrap ensures the frequentist validity of the test.

6.4 Power in simulated example



The three tests we developed leverage different aspects of the problem, and target two different null hypotheses. One may wonder how their power compares in the presence of a treatment effect. Considering once more the boundary between Louisiana and Mississippi, we imagine an experiment where the unit of analysis is the county, located at its centroid, as shown in Figure 6.4(a). We will simulate outcomes from a single Gaussian Process covering both states. For simplicity, we fix the hyperparameters to arbitrary values: $\sigma_\epsilon = \sigma_{GP} = 1.0$ and $\ell = 50$ km. We then add a constant treatment effect τ to all the outcomes in Louisiana. The results for $\tau = 0$ (null hypothesis) and $\tau = 0.5$ are shown in Figure 6.4.



We see that under the null, the p -values of the χ^2 and likelihood ratio tests are uniformly distributed. This is enforced by the parametric bootstrap, which draws test statistics from the same null distribution to calibrate the tests. However, the p -values for the inverse-variance test are biased down, so for example if we set $\alpha = 0.05$, we will falsely reject the null 7.5% instead of 5% of the time. While unfortunate, this is unsurprising, since the inverse-variance test was derived heuristically rather than from a rigorous frequentist procedure. It can be calibrated using the same parametric bootstrap approach that was used for the likelihood and χ^2 tests. The calibration can also be achieved analytically, since $\mu_{\tau|IV|Y}$ is normally distributed under the null hypothesis.

Even after the calibration, the hypothesis test based on the inverse-variance mean has the highest power to detect the constant treatment effect. This can lead to a paradox: we may reject the weak null hypothesis, but fail to reject the sharp null hypothesis (using the χ^2 or likelihood test), even though rejection of the weak null should logically imply rejection of the sharp null. This paradox isn't specific to this setting, and is discussed in depth in the context of randomization-based inference by Ding (2014). Therefore, in scientific

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

6.5 Placebo tests

Gaussian Process models are almost always misspecified. We do not believe that the Gaussian process with stationary squared exponential kernel is the true data-generating process, although we hope that the model is sufficiently flexible to represent reality well. Under misspecification, we should be skeptical of results that rely on the truth of the model specification. We therefore encourage practitioners to probe the validity of the above hypothesis tests by running a “placebo” test. A placebo test repeatedly applies the hypothesis test on data that are known to have zero treatment effect (a “placebo”), in order to verify that the returned p-values are uniformly distributed. In our spatial setting, we will use the treatment and control regions separately as placebo groups. Within each placebo group, we repeatedly draw an arbitrary geographical boundary, creating new treatment and control groups. Because the boundary was chosen arbitrarily by us, we should not expect there to be a discontinuous jump in outcomes at this boundary. We then apply the bootstrapped likelihood test procedure described above to this arbitrarily divided data, store the results, and hope to obtain a roughly uniform distribution of p-values. In our implementation, we drew lines that split the placebo units in half at a sequence of angles $1^\circ, 2^\circ, 3^\circ, \dots, 180^\circ$. The resulting p-values will obviously be highly correlated, so we should only expect a very roughly uniform distribution (because of the small effective sample size), but at the very least, this procedure allows us to visually verify that the p-values are not blatantly biased.

7 Spatial advantage

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

We illustrate this statement once more with the Louisiana-Mississippi example. The variance of the inverse-variance weighted treatment effect τ^{IV} is thence only a function of the positions of the units, available analytically by plugging the posterior variance (9) into the inverse-variance estimator (11). Following this procedure, we obtain a posterior standard deviation of the average treatment effect of 0.31. We then create a one-dimensional regression discontinuity design for the same setting, by using each unit’s distance from the boundary as the covariate x , the distribution of which is shown in Figure 6.4(b). Following the exact same 2GP procedure with the same hyperparameters as in the spatial setting, and with a discontinuity at $x = 0$, we again compute the posterior standard deviation of the treatment effect at the boundary (now a single number rather than a continuous function), this time obtaining 0.58. This higher figure indicates that, perhaps counter-intuitively, the spatial experiment actually has more power than its one-dimensional analog.

To gain intuition about the higher power of the spatial RDD, we turn to the interpretation of regression discontinuity designs as natural experiments [need reference]. Near the discontinuity, we can reasonably claim that the side of the discontinuity that each unit fell into was largely dictated by random noise in the covariate. This in turn allows us to claim that a natural randomized experiment took place near the boundary, with treatment and control units coming from the same population. We can extend this interpretation to the spatial setting, by conceiving of multiple correlated experiments taking place all along the boundary. The average treatment effect estimator then pools the information supplied by all of these experiments. The question then becomes: do we get more powerful inference by grouping all the units into a single experi-

ment, or by spreading them along a multitude of weaker experiments? There are two sources of uncertainty in our model: the observation noise ϵ_i , and the underlying processes g_T and g_C . Adding more units to a single experiment allows us to cancel out more of the observation noise, but if the new units aren't added closer to the discontinuity, uncertainty always remains in g_T and g_C . In the spatial setting, however, we observe multiple realizations of the Gaussian process, and therefore do not suffer from the same diminishing returns.

8 Example: NYC school districts

We illustrate the analysis of geographical regression discontinuity designs using house sales data from New York City. The city publishes information pertaining to property sales within the city in the last twelve months on a rolling basis. This includes the sale price, building class, and the address of the property. Public schools in the city are all part of the City School District of the City of New York, the district is itself divided into 32 sub-districts. Within districts, schools also have attendance zones, and children living within a zone are guaranteed attendance in their zone school unless the school is full [is this true? insideschools.com gives a more complete picture]. It is commonly held [could cite [this article at cityrealtor.com](https://www.cityrealtor.com)] that school districts therefore have an impact on real estate price, as parents are willing to pay more to live in districts with better schools. We therefore ask: can we measure a discontinuous jump in house prices across school district boundaries?

8.1 Preprocessing

In order to model the property sale prices with a stationary Gaussian process, we need to obtain their location on a Euclidean grid. We geocode the address of each sale by merging the sales with NYC's Pluto database, which contains X and Y coordinates for each house, identified by its borough, zip code, block and lot. These coordinates are given in the EPSG:2263 projection in units of feet. We use this projection throughout this example. For addresses that do not find a match in Pluto, we use google's geocoding API to obtain a latitude and longitude, which we then project to EPSG:2263.

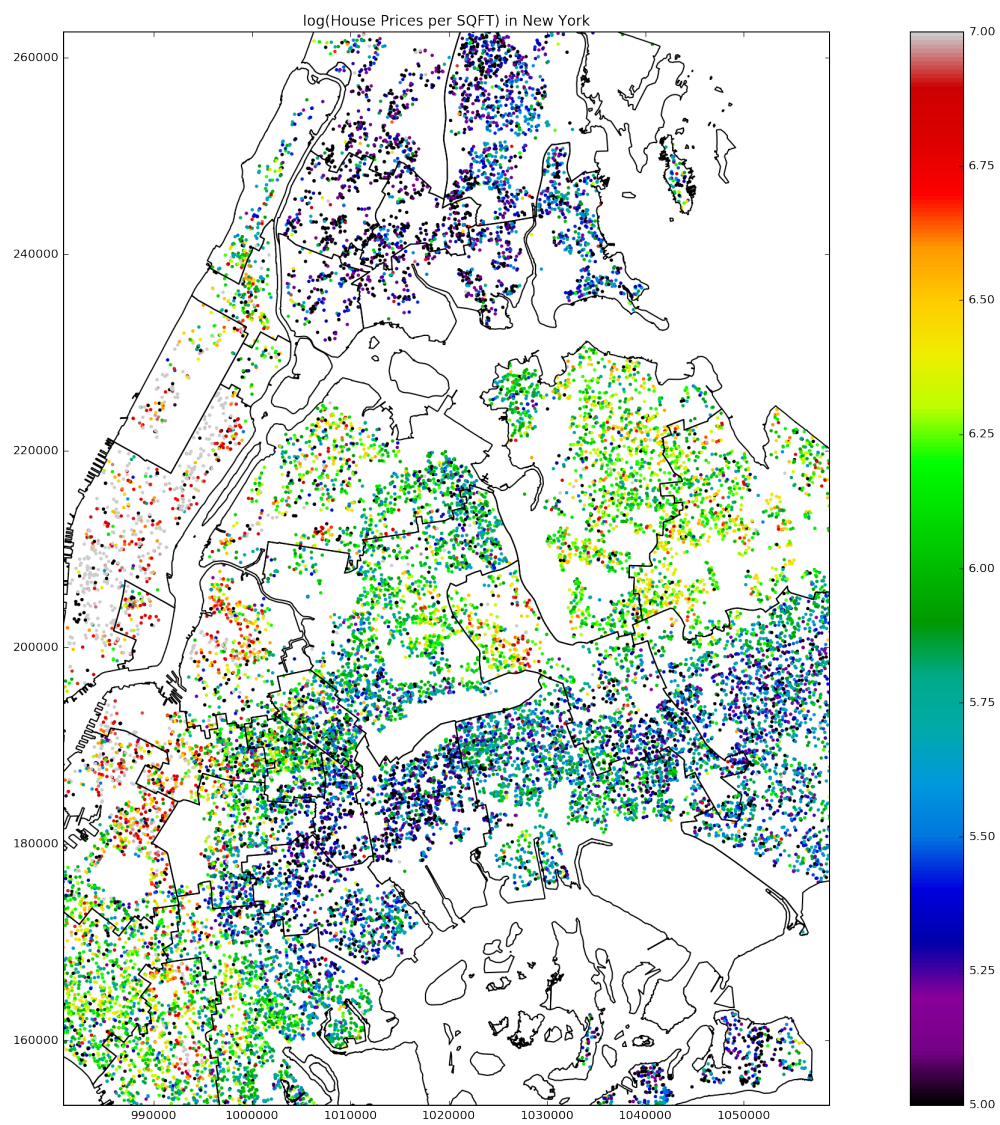
We then filter the sales data as follows, by removing

1. sales of properties without a reported sale price
1. sales of properties outside of the residential building class categories ("one family dwellings", "two family dwellings", "three family dwellings", "tax class 1 condos", "coops - walkup apartments", "coops - elevator apartments", "condos - walkup apartments", "condos - elevator apartments", "condos - 2-10 unit residential", "condo coops"),
2. any sale with missing data in the sale price, square footage, property covariates, geographical coordinates (due to failed geocoding),
3. sales outside of any NYC school district,
4. properties smaller than 100 sq ft, and
5. outliers in the price per square foot.

8.2 Exploratory analysis

8.3 Model for property prices

The outcome of interest is price per square foot. As is often done in the real estate literature, we take its logarithm to reduce the skew of the outcome. The complete model is then a Gaussian Process over the geographical covariates s super-imposed with a linear regression on the property covariates (building and tax class). Within a school district we could write the model as [suggestions for clearer notation welcome]:



sales map

$$Y_i = \log \left(\frac{\text{SalePrice}_i}{\text{SQFT}_i} \right) = \beta_0 + \beta_{1,\text{TaxClass}[i]} + \beta_{2,\text{BuildingClass}[i]} + f(s_i) + \epsilon_i \quad (16)$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma_y^2) \quad (17)$$

$$\beta_{1j}, \beta_{2j} \sim \mathcal{N}(0, \sigma_\beta^2) \quad (18)$$

$$f(s_i) \sim \mathcal{GP}(0, k(s, s')) \quad (19)$$

$$k(s, s') = \sigma_{\text{GP}}^2 \exp \left\{ -\frac{(s - s')^\top (s - s')}{2\ell^2} \right\} \quad (20)$$

A visual inspection of the house sales map above led me to focus on the boundary between districts 19 and 27. I found a map online of average maths performance in each school district, which shows that districts 19 and 27 are quite different. It's important to note that the boundary between the two districts is also part of the boundary between Brooklyn and Queens, so we won't be able to attribute a causal effect solely to the difference in school districts. A histogram of Y in both districts also shows that marginally the house prices are very different.

8.4 parameter optimization

8.5 cliff face

8.5.1 surface plots

8.6 average effect

8.7 significance tests

8.7.1 placebo tests

8.8 pairwise treatment effect (all districts)

9 Conclusion