GeoRDD manuscript

Maxime Rischard

May 2, 2017

Contents

1 Introduction

- Problem we're trying to solve
  - treatment applied to one region and not a neighboring region
  - with no overlap
  - how to estimate the causal effect of the treatment?
  - if the outcome is not spatially varying, there's no problem
  - otherwise, the treatment is confounded with location
- In 1D, this is recognied as a regression discontinuity design
  - which is now a well-established methodology [citations]
  - and comes with a causal inference story [Imbens]
- In spatial settings, practicioners therefore attempt to use these tools
  - but they don't generalize easily to 2D
  - so often end up projecting onto distance from boundary
- We think this is a bad idea
  - ignores spatial structure / correlation
  - low power and could get the wrong answer
- Mention some of the more sophisticated approaches to this same problem [Keele]
  - maybe briefly mention why we don't like them
- Our approach: framework analogous to 1D RDD
  - 1D:
    1. fit the outcome **function** on both sides
    2. extrapolate to the **discontinuity point** $x*$
    3. take difference to obtain $\tau(x*) \in \mathbb{R}$
  - 2D:
    1. fit the outcome **surface** on both sides
    2. extrapolate to **boundary curve** ∂
    3. take difference to obtain $\tau(\partial) : \mathbb{R} \to \mathbb{R}$ [help with notation? or is this too mathematical anyway?]
- Challenge 1: functional estimand is unusual
- Challenge 2: how to fit surface on both sides
  - in this framework there is not restriction on how surface fitting and extrapolation are performed
  - in 1D RDD, local linear regression has become standard
    - though other options have been explored (like splines?)
  - but this isn't suitable in 2D
- Challenge 3: summarizing functional estimand
  - how to take an average?
  - 1D manifold embedded in 2D space poses problems
  - pitfalls detailed in Section X
- Challenge 4: hypothesis testing on functional estimand
- We use Gaussian processes (kriging), which are widespread in spatial statistics
  - many advantages
    - flexible
    - known to perform well in spatial settings
    - analytic solutions

- we explore and address the 4 challenges using GPs
- by the way, GP's can also be used in 1D [cite Zach]

If units in a region receive a treatment, while units in a neighboring region do not, we cannot simply compare mean outcomes between the two regions, as any underlying spatial variation in the outcome would confound the difference in means.

Regression discontinuity design are now a well-established methodology to recover an estimate of a causal effect in settings where a threshold in a covariate determines assignment to treatment or control group.

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 are labeled $Y_i = Y_T(\mathbf{s}_i)$, and units in control region $Y_i = Y_C(\mathbf{s}_i)$.
- Potential outcomes framework: Each unit has a potential outcome under treatment $Y_T(\mathbf{s})$ and a potential outcome under control $Y_C(\mathbf{s})$. If $\mathbf{s} \in \mathcal{S}_T$, then $Y_T(\mathbf{s})$ is observed, otherwise $Y_C(\mathbf{s})$ is observed.

2.2 1GP solution

Most straightforwardly, we model the observed outcomes Y at locations S (an $n \times 2$ matrix) 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.

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

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

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

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

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

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

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, $(Y, f(S), \tau, \mu, \beta) \mid \ell, \sigma_{GP}$ is multi-variate normal with variance-covariance given by

$$
\begin{aligned}
\tau &\sim \mathcal{N}\left(0, \sigma^2_\tau\right) \\
\mu &\sim \mathcal{N}\left(0, \sigma^2_\mu\right) \\
\beta &\sim \mathcal{N}\left(0, \sigma^2_\beta\right) \\
\operatorname{cov}(Y_i(\mathbf{s}), \tau) &= \sigma^2_\tau \mathbb{I}\{\mathbf{s} \in \mathcal{S}_T\} \\
\operatorname{cov}(Y_i(\mathbf{s}), \mu) &= \sigma^2_\mu \\
\operatorname{cov}(Y_i(\mathbf{s}), \beta) &= \sigma^2_\beta \mathbf{s}^\intercal \mathbf{s} \\
\operatorname{cov}(Y_i(\mathbf{s}), Y_i(\mathbf{s}')) &= \sigma^2_\mu + \sigma^2_\tau \mathbb{I}\{\mathbf{s} \in \mathcal{S}_T\} \mathbb{I}\left\{\mathbf{s}' \in \mathcal{S}_T\right\} + \sigma^2_\beta \mathbf{s}^\intercal \mathbf{s}' + k(\mathbf{s}, \mathbf{s}') + \delta_{ij} \sigma^2_\epsilon \\
\operatorname{cov}(Y(\mathbf{s}), f(\mathbf{s}')) &= \operatorname{cov}(f(\mathbf{s}), f(\mathbf{s}')) = k(\mathbf{s}, \mathbf{s}')
\end{aligned}
\tag{4}
$$

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

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

To proceed computationally, we define the treatment indicator vector $\mathbb{I}_T$ with ith 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.

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

What remains is the inference on the hyperparameters $\sigma_\epsilon, \sigma_{GP}$ 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 $x = b$:

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

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

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

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

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

4 Handling covariates

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

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.

In the 2GP model, the model becomes

$$\begin{aligned}
Y_{T,i}(\mathbf{s}) &= \underbrace{\mu_T + \mathbf{d}_i^\intercal \gamma + \mathbf{s}^\intercal \beta_T + f_T(\mathbf{s})}_{g_T(\mathbf{s})} + \epsilon_i \\
Y_{C,i}(\mathbf{s}) &= \underbrace{\mu_C + \mathbf{d}_i^\intercal \gamma + \mathbf{s}^\intercal \beta_C + f_C(\mathbf{s})}_{g_C(\mathbf{s})} + \epsilon_i \\
f_T(S), f_C(S) &\overset{\perp\!\!\!\perp}{\sim} \mathcal{GP}\left(0, k(\mathbf{s}, \mathbf{s}')\right) \\
k(\mathbf{s}, \mathbf{s}') &= \sigma_{GP}^2 \exp\left(-\frac{(\mathbf{s}-\mathbf{s}')^\intercal (\mathbf{s}-\mathbf{s}')}{2\ell^2}\right) \\
\gamma_j &\overset{\perp\!\!\!\perp}{\sim} \mathcal{N}\left(0, \sigma_\gamma^2\right) \text{ for } j = 1, 2, \ldots, p
\end{aligned} \tag{10}$$

Unfortunately, the linear term induces a covariance between the treatment and control region, and so the computational advantage over the 1GP model is lost. When the two regions are independent, fitting the Gaussian processes required obtaining the Cholesky decomposition of a $n_T \times n_T$ matrix, and of an $n_C \times n_C$ matrix. Just like in the 1GP case, the linear regression component adds a term $\sigma_\gamma^2 D^\intercal D$ to the covariance of Y, which is therefore no longer block diagonal. Thus the Cholesky decomposition of an $(n_T + n_C) \times (n_T + n_C)$ is now required. Cholesky decomposition algorithms generally have computational complexity $O(n^3)$. Therefore, if the units are evenly split between the two regions, the overall complexity of the model fitting increases fourfold.

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 $\tau(\mathbf{s})$ along the boundary, by averaging the entries of the mean posterior at the 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.

$$\begin{aligned}
\bar{\tau} &\equiv \frac{\oint_\partial \tau(x) dx}{\oint_\partial dx} \\
\bar{\tau} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell &\sim \mathcal{N}\left(\mu_{\bar{\tau}|Y}, \Sigma_{\bar{\tau}|Y}\right) \\
\mu_{\bar{\tau}|Y} &\approx \left(\mathbf{1}^\intercal \mu_{\partial|Y}\right) / n_\partial \\
\Sigma_{\bar{\tau}|Y} &\approx \left(\mathbf{1}^\intercal \Sigma_{\partial|Y} \mathbf{1}\right) / n_\partial^2
\end{aligned} \tag{11}$$

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

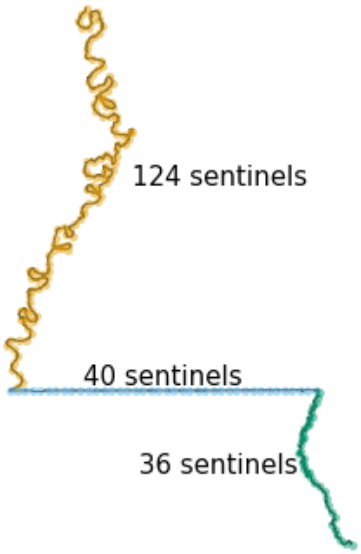

Figure 1: Evenly spaced sentinels along the border between Mississippi and Louisiana.

there has little meaning and importance. Furthermore, $\tau(\mathbf{s})$ 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 1). 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 avereage with minimum variance.

$$
\begin{aligned}
\tau^{IV} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell &\sim \mathcal{N}\left(\mu_{\tau^{IV}|Y}, \Sigma_{\tau^{IV}|Y}\right) \\
\mu_{\tau^{IV}|Y} &\approx \left(\mathbf{1}^\intercal \Sigma_{\partial|Y}^{-1} \mu_{\partial|Y}\right) / \left(\mathbf{1}^\intercal \Sigma_{\partial|Y}^{-1} \mathbf{1}\right) \\
\Sigma_{\tau^{IV}|Y} &\approx 1 / \left(\mathbf{1}^\intercal \Sigma_{\partial|Y}^{-1} \mathbf{1}\right)
\end{aligned}
\tag{12}
$$

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 rather 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 (12). While the posterior is a Bayesian object, we can use it heuristically to derive a pseudo-p-value

$$
\begin{aligned}
Z_0 &\sim \mathcal{N}\left(0, \Sigma_{\tau^{IV}|Y}\right) \\
p^{INV} &= \mathbb{P}\left(|Z_0| > \left|\mu_{\tau^{IV}|Y}\right|\right) \\
&= 2\Phi\left(-\frac{\left|\mu_{\tau^{IV}|Y}\right|}{\sqrt{\Sigma_{\tau^{IV}|Y}}}\right)
\end{aligned}
\tag{13}
$$

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 $2\mathcal{GP}$ model as specified in equation (2). We therefore choose the difference in log-likelihoods as our test statistic

$$
t = \log \mathbb{P}\left(Y_T, Y_C \mid \mathcal{M}_1\right) - \log \mathbb{P}\left(Y_T, Y_C \mid \mathcal{M}_0\right) \tag{14}
$$

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

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

$$t^* = \log \mathbb{P}\left(Y_T^*, Y_C^* \mid \mathscr{M}_1\right) - \log \mathbb{P}\left(Y_T^*, Y_C^* \mid \mathscr{M}_0\right) \tag{15}$$

Repeating this procedure, we obtain a distribution of t under $\mathscr{M}_0$, 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.

$$p^{\text{lik}} = \mathbb{P}\left(t^* > t_{obs} \mid \mathscr{M}_0\right) \tag{16}$$

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 (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 $y^\intercal \Sigma^{-1} y$ has distribution χ^2_k. This suggests that we could use $S = \mu^\intercal \Sigma^{-1} \mu$ as a test statistic, and obtain a p-value by comparing it to a χ^2_k distribution, where k is the number of sentinels. However, we face two problems. Firstly, this test derived heuristically from a Bayesian posterior is invalid from a frequentist perspective. Secondly, while 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, this time using S as the test statistic instead of the likelihood ratio. Because S involves inverting a matrix Σ that is mathematically of full rank, but numerically of low rank, we may worry about the numerical stability of computing S. We rely on julia's matrix division polyalgorithm to ensure numerical stability, and check in simulated examples that adding a small constant to the diagonal of Σ does not greatly affect the computed S. Furthermore, even if numerical stability was an issue, the parametric bootstrap ensures the frequentist validity of the test, though its power could be lowered.

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 2(a). We will simulate outcomes from a single Gaussian Process covering both states. For simplicity, we fix the hyperparameters to arbitrary values: $\sigma_\epsilon = \sigma_{GP} = 1.0$ and $\ell = 50\,\text{km}$. We then add a constant treatment effect τ to all the outcomes in Louisiana. The results for $\tau = 0$ (null hypothesis) and $\tau = 0.5$ are shown in Figure 3.

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

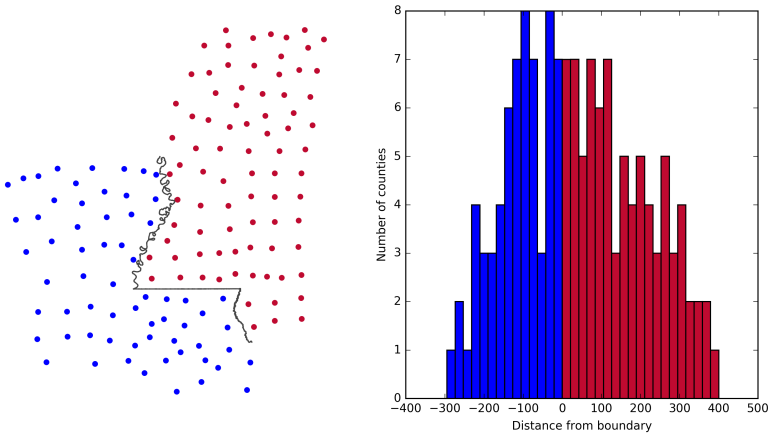

Figure 2: Position of units in an imaginary experiment in Louisiana and Mississippi

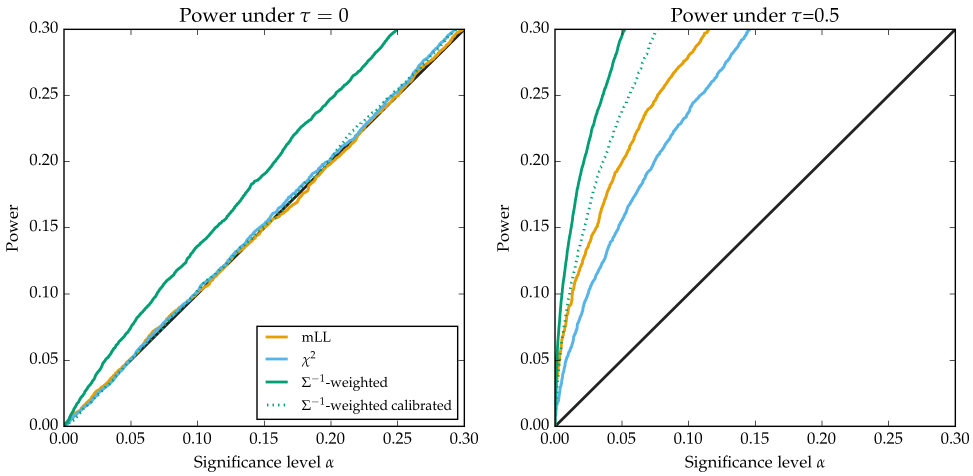

Figure 3: Power of hypothesis tests

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 practicioners to probe the validity of the above hypothesis tests by running a "placebo" test. A placebo test repeatedly applies the hypothesis test on data that are known to have zero treatment effect (a "placebo"), in order to verify that the returned p-values are uniformly distributed. In our spatial setting, we will use the treatment and control regions separately as placebo groups. Within each placebo group, we repeatedly draw an arbitrary geographical boundary, creating new treatment and control groups. Because the boundary was chosen arbitrarily by us, we should not expect there to be a discontinuous jump in outcomes at this boundary. We then apply the bootstrapped likelihood test procedure described above to this arbitrarily divided data, store the results, and hope to obtain a roughly uniform distribution of p-values. In our implementation, we drew lines that split the placebo units in half at a sequence of angles $1^\circ, 2^\circ, 3^\circ, \ldots, 180^\circ$. The resulting p-values will obviously be highly correlated, so we should only expect a very roughly uniform distribution (because of the small effective sample size), but at the very least, this procedure allows us to visually verify that the p-values are not blatantly biased.

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 (12). 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 2(b). Following the exact same 2GP procedure with the same hyperparameters as in the spatial setting, and with a discontinuity at $x = 0$, we again compute the posterior standard deviation of the treatment effect at the boundary (now a single number rather than a continuous function) , this time obtaining 0.58. This higher figure indicates that, perhaps counter-intuitively, the spatial experiment actually has more power than its one-dimensional analog.

To gain intuition about the higher power of the spatial RDD, we turn to the interpretation of regression discontinuity designs as natural experiments [need reference]. Near the discontinuity, we can reasonably claim that the side of the discontinuity that each unit fell into was largely dictated by random noise in the covariate. This in turn allows us to claim that a natural randomized experiment took place near the boundary, with treatment and control units coming from the same population. We can extend this interpretation to the spatial setting, by conceiving of multiple correlated experiments taking place all along the boundary. The average treatment effect estimator then pools the information supplied by all of these experiments. The question then becomes: do we get more powerful inference by grouping all the units into a single experiment, or by spreading them along a multitude of weaker experiments? There are two sources of uncertainty in our model: the observation noise 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, but the city-wide district is itself divided into 32 sub-districts. Within these districts, schools also have attendance zones, and children living within a zone are guaranteed attendance in their zone school unless the school is full [is this true? insideschools.com gives a more complete picture]. It is commonly held [could cite this article at cityrealty.com] that school districts therefore have an impact on real estate price, as parents are willing to pay more to live in districts with better schools. We therefore ask: can we measure a discontinuous jump in house prices across school district boundaries?

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 $\mathbf{s}$ super-imposed with a linear regression on the property covariates (building and tax class). Within a school district we could write the model as [suggestions for clearer notation welcome]:

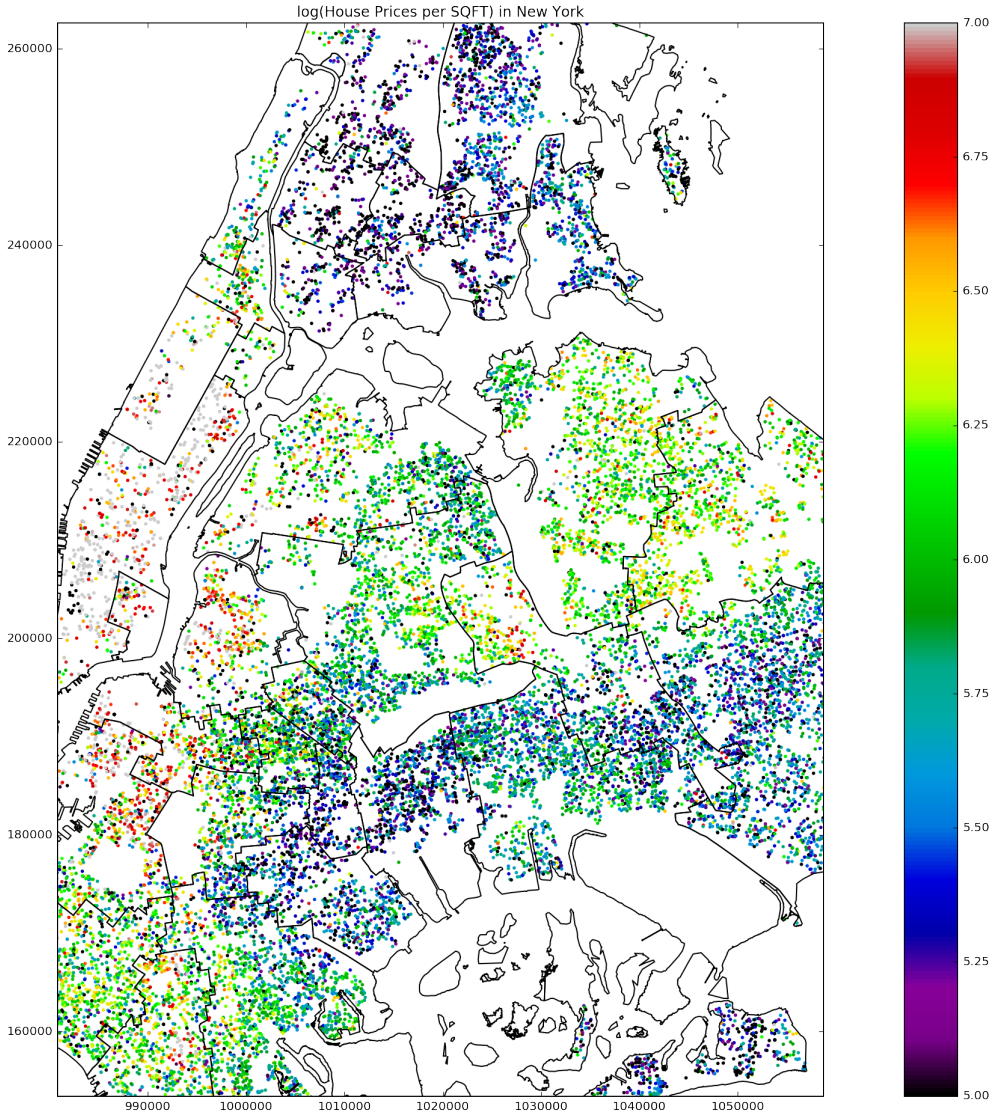

Figure 4: sales map

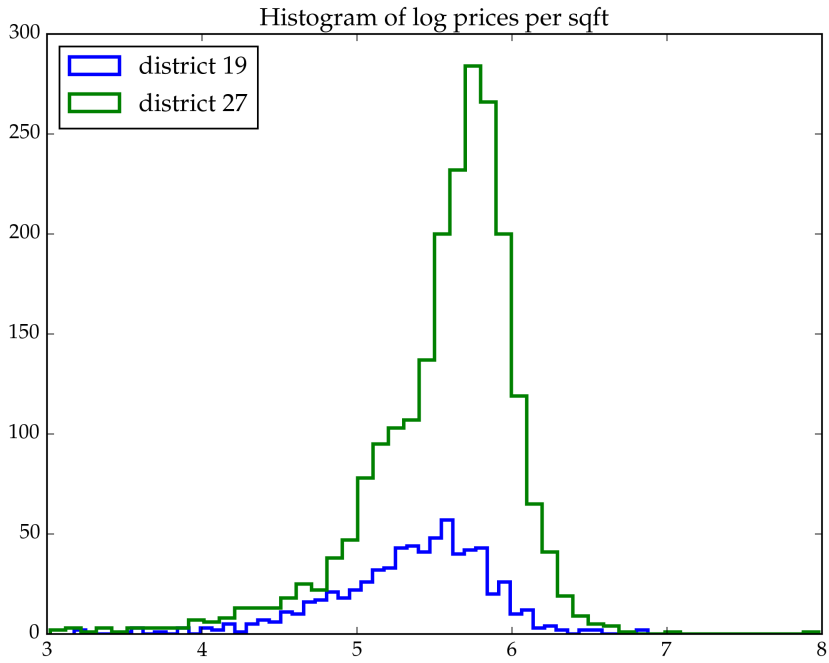

Figure 5:

$$
\begin{aligned}
Y_i = \log\left(\frac{\texttt{SalePrice}_i}{\texttt{SQFT}_i}\right) = {} & \mu_{\texttt{District[i]}} + \beta_{1,\texttt{TaxClass[i]}} + \beta_{2,\texttt{BuildingClass[i]}} \\
& + f_{\texttt{District[i]}}(\mathbf{s}_i) + \epsilon_i \\
\epsilon_i &\sim \mathcal{N}\left(0, \sigma_y^2\right) \\
\mu_j &\sim \mathcal{N}\left(\bar{Y}_j, \sigma_\mu^2\right) \\
\beta_{1j}, \beta_{2j} &\sim \mathcal{N}\left(0, \sigma_\beta^2\right) \\
f_j(\mathbf{s}_i) &\sim \mathcal{GP}\left(0, k(\mathbf{s}, \mathbf{s}')\right) \\
k(\mathbf{s}, \mathbf{s}') &= \sigma_{GP}^2 \exp\left\{-\frac{(\mathbf{s}-\mathbf{s}')^\intercal(\mathbf{s}-\mathbf{s}')}{2\ell^2}\right\}
\end{aligned}
\tag{17}
$$

A visual inspection of the house sales map above suggests examining the boundary between districts 19 and 27. Importantly, the boundary between the two districts is also part of the boundary between Brooklyn and Queens, so we won't be able to attribute a causal effect solely to the difference in school districts. We are first and foremost *measuring* a discontinuity in the house prices at the district. Attributing the discontinuity to a particular cause (school district or borough) is an interpretation that is not directly supported by the data. A histogram of Y in both districts also shows that marginally the house prices are very different. Our goal is to establish that this difference is measurable at the boundary, and not merely an underlying trend that spans both districts.

8.4 parameter optimization

We initially fit the hyperparameters σ_β, σ_{GP}, ℓ and σ_ϵ by optimizing the marginal log-likelihood of the data within a single district. We choose district 27 as it contains more sales. We hold σ_μ fixed to 10 to give the

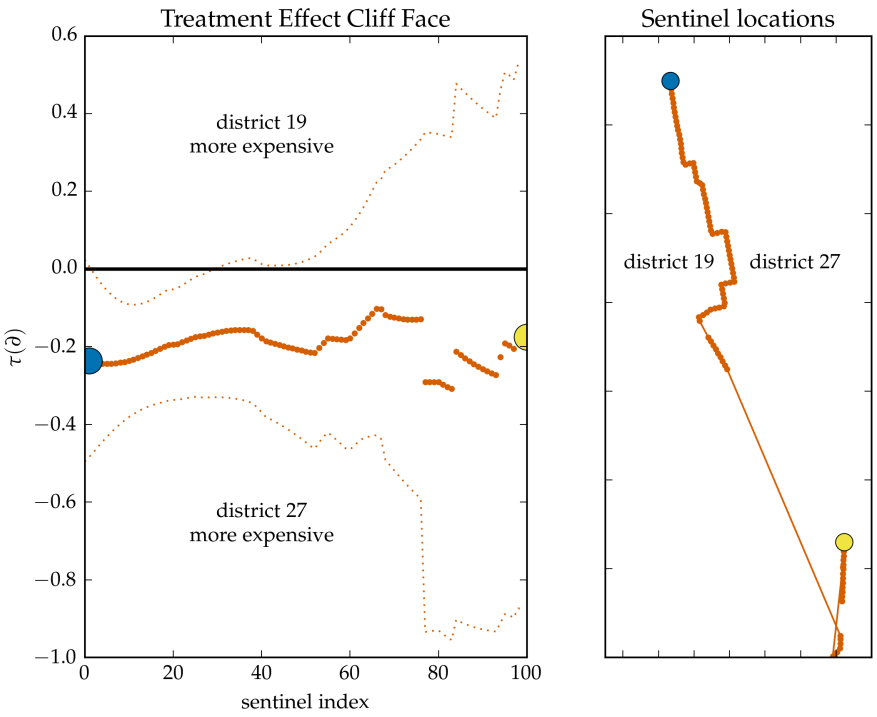

Figure 6: NYC cliff face

district means μ_j a fairly uninformative prior. The fitted hyperparameters were $\sigma_\epsilon = 0.4179$, $\sigma_{GP} = 0.2426$, $\sigma_\beta = 0.1306$, and $\ell = 3378.5800$ ft.

8.5 cliff face

We seek the treatment effect function $\tau(\partial)$ between the two districts. We could proceed by computing the joint predictive distributions $g_T(\boldsymbol{\partial}), g_C(\boldsymbol{\partial}) \mid Y_T, Y_C, \sigma_\beta, \sigma_{GP}, \ell, \sigma_\epsilon$, which is a $2n_{\boldsymbol{\partial}}$-dimensional multivariate normal distribution. Instead, we obtain the posterior means of the β_{1j} and β_{2j} coefficients, extract the residuals $Y_T - D_T\hat{\beta}$ and $Y_C - D_C\hat{\beta}$. This decorrelates $g_T(\boldsymbol{\partial})$ and $g_C(\boldsymbol{\partial})$ so they become independent multivariate normal distributions $g_T(\boldsymbol{\partial}) \mid Y_T, \hat{\beta}, \sigma_{GP}, \ell, \sigma_\epsilon$ and $g_C(\boldsymbol{\partial}) \mid Y_C, \hat{\beta}, \sigma_{GP}, \ell, \sigma_\epsilon$. In this example, we find that the posterior variance of β is low, and therefore the two approaches yield very similar results, but conditioning on the estimate of β is computationally convenient. We therefore proceed with this two-step approach.

Equipped with multivariate normal posteriors on $g_C(\boldsymbol{\partial})$ and $g_T(\boldsymbol{\partial})$, which are uncorrelated conditional on $\beta = \hat{\beta}$, we can now take their difference according to the procedure outline in section 2.3, to obtain the posterior distribution of the cliff-face $\tau(\boldsymbol{\partial})$ obtained at the sentinel locations. The cliff-face is shown in Figure XX, and shows that the estimated $\tau(\boldsymbol{\partial})$ is negative everywhere along the border, which corresponds to higher property prices in district 27. However, the credible envelope is wide, especially in the Southern section of the border, and therefore it isn't clear that this effect isn't due to random variation.

The treatment effect can also be visualized directly in Figure XX as the difference between the two log-price mean surfaces $g(\mathbf{s})$. This pictures also gives a better sense of the important spatial variation in prices captured by the model, which explains the wide credible envelope in the cliff face, despite the large number of sales in both districts.

8.6 Average Log-Price Increase

The cliff-face plot shows a negative treatment effect everywhere along the border, which can be averaged by the estimators we developed in section XX. The most obvious approach is to take an unweighted mean

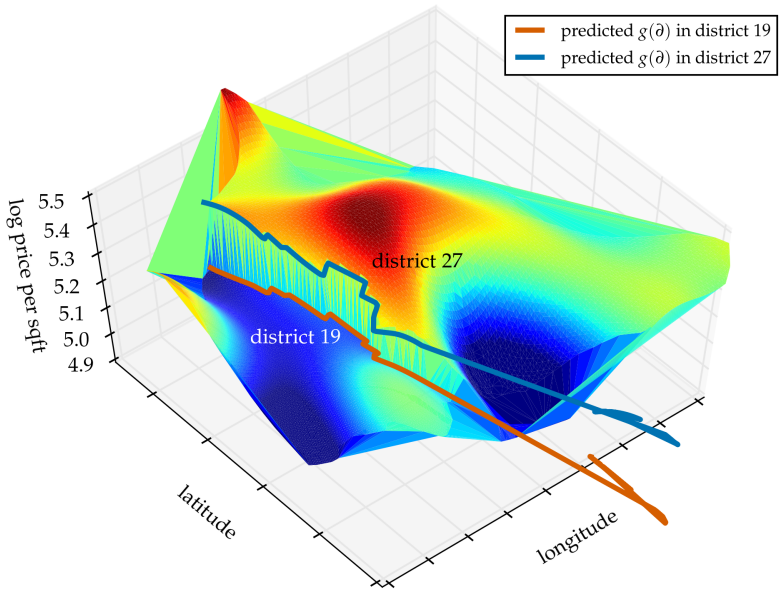

Figure 7: NYC surface plot

at each equispaced sentinel, which has posterior distribution

$$\begin{aligned} \bar{\tau} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \hat{\beta}, \ell \sim \mathcal{N}\left(-0.198, 0.091^2\right) \text{ and tail probability} \\ \mathbb{P}\left(\bar{\tau} > 0 \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \hat{\beta}, \ell\right) = 1.515\%\,. \end{aligned} \tag{18}$$

The inverse-variance weighted mean estimator is robust to changes in the border topology, and gives higher weight to sections of the border where the difference in house prices is easier to measure. It is guaranteed to minimize the posterior variance amongst weighted mean estimators, which is reflected here by the narrower posterior distribution

$$\begin{aligned} \tau^{IV} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \hat{\beta}, \ell \sim \mathcal{N}\left(-0.192, 0.059^2\right) \text{ and reduced tail probability} \\ \mathbb{P}\left(\tau^{IV} > 0 \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \hat{\beta}, \ell\right) = 0.017\%\,. \end{aligned} \tag{19}$$

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

8.7 Significant Difference in Price?

The inverse-variance weighted mean treatment effect hints at a significant treatment effect. But the posterior tail probability cannot be interpreted as a p-value. For this, we turn to the three tests developed in section XX. In applied settings, running multiple tests invalidates their results, but as we are proposing this new methodology, we apply all three tests in order to gain insight into their differences. Their results are found in Table XX.

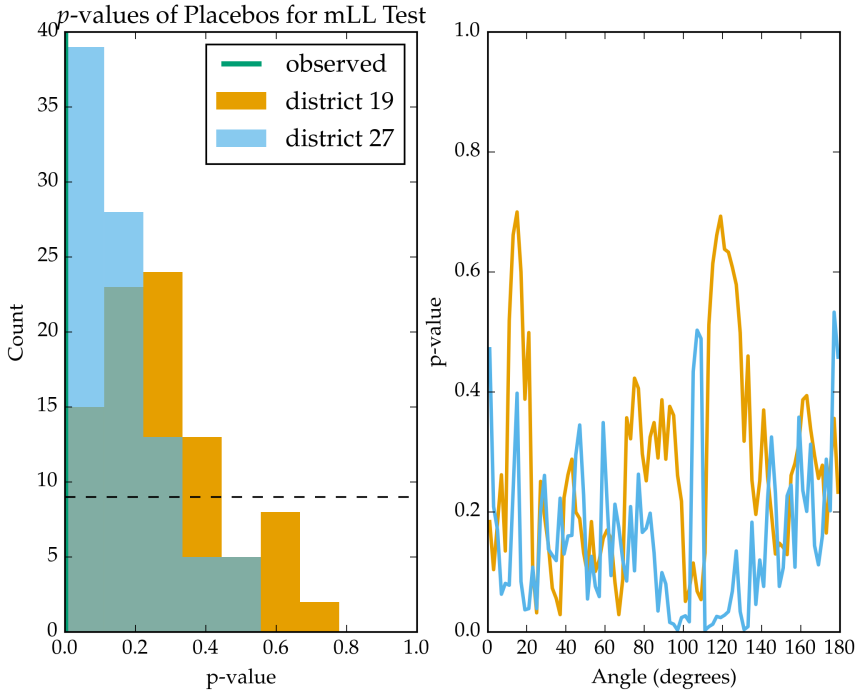

Figure 8: Placebo test for mLL test

Test	p-value
χ^2 bootstrap	0.145
mLL bootstrap	0.0015
τ^{IV} uncalibrated	0.0003
τ^{IV} calibrated	0.0005

The three tests tell very different stories. The χ^2 test fails to reject the null hypothesis even at the $\alpha = 0.1$ level. This strongly contradicts the inverse-variance test and its low p-value of 0.0005, backed by the likelihood-ratio test with $p = 0.0015$. The possibility of such a contradiction was anticipated in section XX, where we saw that the χ^2 has the lowest power of the three tests, and therefore could easily fail to reject an effect that is easily detected by the inverse-variance test. Because the inverse-variance test has the highest power in detecting constant treatment effects, we would recommend its use in applications — such as this one — where a very heterogenous treatment effect is not expected.

8.7.1 placebo tests

To assess the validity of the three tests, we apply the placebo tests that we devised in Section X. Within each district, we split the data in half by a line at angles $1°, 3°, 5°, 6°, \ldots, 179°$. Because these lines were drawn arbitrarily, we don't expect a discontinuous treatment effect between the two halves, and so we hope to see a uniform distribution of placebo p-values. However, these tests will be highly correlated, and so the low effective sample size could lead to some apparent departures from uniformity. There is in fact visible autocorrelation in the graphs of placebo p-values as a function of angle.

The mLL placebo p-values show a pronounced bias towards low values. This confirms our earlier concern that the marginal log-likelihood may be sensitive to features of the data other than the discontinuity at the boundary. In particular, model misspecification, which is a big concern in spatial models, makes the

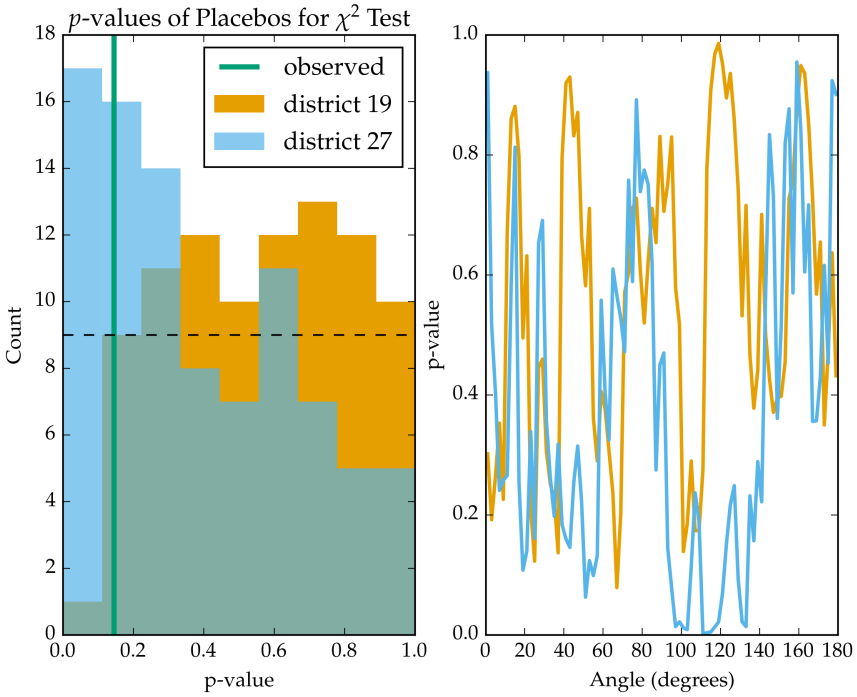

Figure 9: Placebo test for χ^2 test

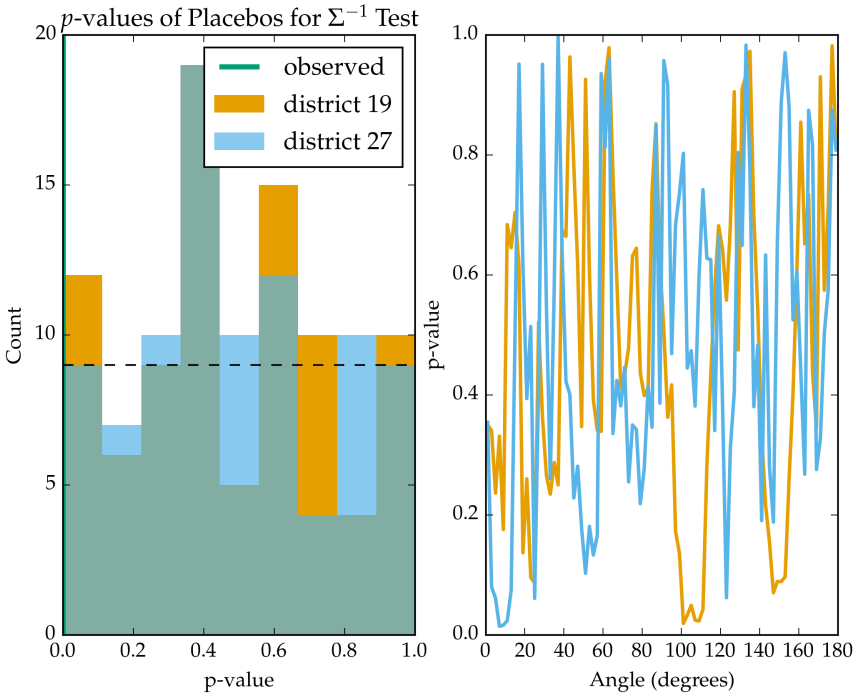

Figure 10: Placebo test for Σ^{-1} test

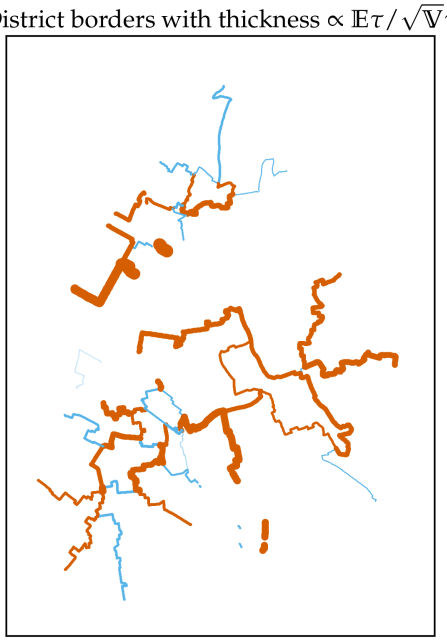

Figure 11: Pairwise effect size between adjacent districts.

interpretation of the mLL test unreliable. Based on this vulnerability, and its manifestation in this example, we do not recommend relying on the likelihood-ratio test.

The χ^2 test shows more robustness, with some negative bias in district 27, and some positive bias in district 19, which could simply be due to the low effective sample size. We therefore believe that the 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 inverse-variance placebo p-values display no obvious bias. Its high power and robustness to misspecification make a strong argument for its use in most applications.

8.8 pairwise treatment effect (all districts)

9 Conclusion

10 Appendices

10.1 Posterior mean of $\hat{\beta}$

[Derivation of $\hat{\beta}$ below: should it be in the covariates section? should it be in an appendix? is it too elementary to be in this paper?]

$$
\begin{aligned}
\Sigma_{Y|\beta} &\equiv \operatorname{cov}(Y \mid \beta) && \text{conditional variance of } Y \\
\operatorname{cov}\left(Y_i, Y_j \mid \beta\right) &= \sigma_\epsilon^2 \delta_{ij} + k\left(\mathbf{s}_i, \mathbf{s}_j\right) \delta_{\texttt{District[i]},\texttt{District[j]}} && \text{(block diagonal)} \\
\Sigma_\beta &\equiv \operatorname{cov}(\beta) = \sigma_\beta^2 I_p && \text{prior variance of } \beta \\
\Sigma_Y &\equiv \operatorname{cov}(Y) = \Sigma_{Y|\beta} + D^\intercal \Sigma_\beta D && \text{unconditional variance of } Y \\
T_\beta &= D^\intercal \Sigma_{Y|\beta}^{-1} D + \Sigma_\beta^{-1} && \text{precision matrix of } \beta \\
\hat{\beta} &= \left(T_\beta^{-1} D\right)\left(\Sigma_{Y|\beta}^{-1}\left(Y - \mu\right)\right) && \text{posterior mean of } \beta
\end{aligned}
\tag{20}
$$

10.2 Calibration of inverse-variance test

First, let's remind ourselves how the inverse-variance posterior mean estimate was obtained. We will then derive its distribution under the null hypothesis.

$$
\begin{aligned}
\tau^{IV} \mid Y_T, Y_C, \sigma_{GP}, \sigma_\epsilon, \ell &\sim \mathcal{N}\left(\mu_{\tau^{IV}|Y}, \Sigma_{\tau^{IV}|Y}\right) \\
\mu_{\tau^{IV}|Y} &\approx \left(\mathbf{1}^\intercal \Sigma_{\partial|Y}^{-1} \mu_{\partial|Y}\right) / \left(\mathbf{1}^\intercal \Sigma_{\partial|Y}^{-1} \mathbf{1}\right) \\
\mu_{\partial|T} &\equiv \operatorname{cov}\left(g_T(\partial), Y_T\right) \operatorname{cov}\left(Y_T\right)^{-1} Y_T \\
\mu_{\partial|C} &\equiv \operatorname{cov}\left(g_T(\partial), Y_C\right) \operatorname{cov}\left(Y_C\right)^{-1} Y_C \\
\mu_{\partial|Y} &= \mu_{\partial|T} - \mu_{\partial|C} \\
\mu_{\tau^{IV}|Y} &= \left(\mathbf{1}^\intercal \Sigma_{\partial|Y}^{-1} \mu_{\partial|Y}\right) / \left(\mathbf{1}^\intercal \Sigma_{\partial|Y}^{-1} \mathbf{1}\right)
\end{aligned}
\tag{21}
$$

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

$$
\begin{aligned}
\operatorname{cov}\left(\begin{pmatrix} Y_T \\ Y_C \end{pmatrix} \mid H_0\right) &= \begin{bmatrix} \Sigma_{TT} & \Sigma_{TC} \\ \Sigma_{TC}^\intercal & \Sigma_{CC} \end{bmatrix} \text{ where} \\
\Sigma_{TT} &\equiv K_{TT} + \sigma_\epsilon^2 I_{n_T} \\
\Sigma_{CC} &\equiv K_{CC} + \sigma_\epsilon^2 I_{n_C} \\
\Sigma_{TC} &\equiv K_{TC}
\end{aligned}
\tag{22}
$$

where the entries of K_{TT}, K_{CC} and K_{TC} are obtained simply be evaluating the Gaussian process kernel for each pair of points within and between the treatment and control regions. The predicted mean outcomes at the sentinels $\mu_{\partial|T}$ and $\mu_{\partial|T}$ are obtained by left-multiplying Y_T and Y_C by matrices that are deterministic functions of the unit locations and the hyperparameters

$$
\begin{aligned}
A_T &\equiv \operatorname{cov}\left(g_T(\partial), Y_T\right) \operatorname{cov}\left(Y_T\right)^{-1} = K_{\partial T} \Sigma_{TT}^{-1}, \text{ and} \\
A_C &\equiv \operatorname{cov}\left(g_C(\partial), Y_C\right) \operatorname{cov}\left(Y_C\right)^{-1} = K_{\partial C} \Sigma_{CC}^{-1}.
\end{aligned}
\tag{23}
$$

where we dropped the explicit conditioning on the null hypothesis for readability.

The joint distribution of $\mu_{\partial|T}$ and $\mu_{\partial|T}$ is consequently also multivariate normal with mean zero and covariance

$$
\operatorname{cov}\left(\begin{pmatrix} A_T Y_T \\ A_C Y_C \end{pmatrix} \mid H_0\right) = \begin{bmatrix} A_T \Sigma_{TT} A_T^\intercal & A_T \Sigma_{TC} A_C^\intercal \\ \left(A_T \Sigma_{TC} A_C^\intercal\right)^\intercal & A_C \Sigma_{CC} A_C^\intercal \end{bmatrix}
\tag{24}
$$

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

$$\begin{aligned}\operatorname{cov}\left(\mu_{\partial|Y} \mid H_0\right) &= \operatorname{cov} A_T Y_T - A_C Y_C \\ &= A_T \Sigma_{TT} A_T^\intercal + A_C \Sigma_{CC} A_C^\intercal - A_T \Sigma_{TC} A_C^\intercal - \left(A_T \Sigma_{TC} A_C^\intercal\right)^\intercal\end{aligned} \tag{25}$$

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

$$\begin{aligned}\operatorname{var}\left(\bar{\tau}^{\boldsymbol{\nu}} \mid H_0\right) &= \operatorname{cov}\left(\frac{\boldsymbol{\nu}^\intercal \mu_{\partial|Y}}{\mathbf{1}_{n_\partial}^\intercal \boldsymbol{\nu}}\right) \\ &= \frac{\boldsymbol{\nu}^\intercal \operatorname{cov}\left(\mu_{\partial|Y}\right) \boldsymbol{\nu}}{\left(\mathbf{1}_{n_\partial}^\intercal \boldsymbol{\nu}\right)^2}.\end{aligned} \tag{26}$$

From this null distribution the p-value follows:

$$\mathbb{P}\left(|\bar{\tau}^{\boldsymbol{\nu}}| > |\bar{\tau}^{\boldsymbol{\nu}}_{obs}| \mid H_0\right) = 2\Phi\left(-\frac{|\bar{\tau}^{\boldsymbol{\nu}}_{obs}|}{\sqrt{\operatorname{var}\left(\bar{\tau}^{\boldsymbol{\nu}} \mid H_0\right)}}\right). \tag{27}$$

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

References