GeoRDD manuscript

Maxime Rischard

February 4, 2017

Contents

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 $Y_T(\mathbf{s})$, and units in control region $Y_C(\mathbf{s})$.

- 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 $s \in \mathscr{S}_T$, 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 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.

$$
\begin{aligned}
Y_i(\mathbf{s}) &= \mu + \mathbf{s}^\intercal \beta + f(\mathbf{s}) + \tau\, \mathbb{I}\{\mathbf{s} \in \mathscr{S}_T\} + \epsilon_i && (1)\\
f(S) &\sim \mathcal{GP}\left(0, k(\mathbf{s}, \mathbf{s}')\right) && (2)\\
k(\mathbf{s}, \mathbf{s}') &= \sigma^2_{\mathrm{GP}} \exp\left(-\frac{(\mathbf{s}-\mathbf{s}')^\intercal (\mathbf{s}-\mathbf{s}')}{2\ell^2}\right) && (3)\\
\epsilon_i &\overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2_\epsilon\right) && (4)
\end{aligned}
$$

$f(S)$ is a smooth surface covering all of $\mathscr{S}$, 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.

$$
\begin{aligned}
Y_{T,i}(\mathbf{s}) &= \underbrace{\mu_T + \mathbf{s}^\intercal \beta_T + f_T(\mathbf{s})}_{g_T(\mathbf{s})} + \epsilon_i \\
Y_{C,i}(\mathbf{s}) &= \underbrace{\mu_C + \mathbf{s}^\intercal \beta_C + f_C(\mathbf{s})}_{g_C(\mathbf{s})} + \epsilon_i \\
f_T(S), f_C(S) &\overset{\perp\!\!\!\perp}{\sim} \mathcal{GP}\left(0, k(\mathbf{s}, \mathbf{s}')\right) \\
k(\mathbf{s}, \mathbf{s}') &= \sigma^2_{\mathrm{GP}} \exp\left(-\frac{(\mathbf{s}-\mathbf{s}')^\intercal (\mathbf{s}-\mathbf{s}')}{2\ell^2}\right)
\end{aligned}
\qquad (5)
$$

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.

$$
\tau(\mathbf{s}) = [\mu_T + \mathbf{s}^\intercal \beta_T + f_T(\mathbf{s})] - [\mu_C + \mathbf{s}^\intercal \beta_C + f_C(\mathbf{s})] \qquad (6)
$$

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 multivariate normal with variance-covariance given by

$$\begin{aligned}
\tau &\sim \mathcal{N}\left(0, \sigma_\tau^2\right) \\
\mu &\sim \mathcal{N}\left(0, \sigma_\mu^2\right) \\
\beta &\sim \mathcal{N}\left(0, \sigma_\beta^2\right) \\
\operatorname{cov}(Y_i(\mathbf{s}), \tau) &= \sigma_\tau^2 \mathbb{I}\left\{\mathbf{s} \in \mathscr{S}_T\right\} \\
\operatorname{cov}(Y_i(\mathbf{s}), \mu) &= \sigma_\mu^2 \\
\operatorname{cov}(Y_i(\mathbf{s}), \beta) &= \sigma_\beta^2 \mathbf{s}^\intercal \mathbf{s} \\
\operatorname{cov}(Y_i(\mathbf{s}), Y_i(\mathbf{s}')) &= \sigma_\mu^2 + \sigma_\tau^2 \mathbb{I}\left\{\mathbf{s} \in \mathscr{S}_T\right\} \mathbb{I}\left\{\mathbf{s}' \in \mathscr{S}_T\right\} + \sigma_\beta^2 \mathbf{s}^\intercal \mathbf{s}' + k(\mathbf{s}, \mathbf{s}') + \delta_{ij} \sigma_\epsilon^2 \\
\operatorname{cov}(Y(\mathbf{s}), f(\mathbf{s}')) = \operatorname{cov}(f(\mathbf{s}), f(\mathbf{s}')) &= k(\mathbf{s}, \mathbf{s}')
\end{aligned} \tag{7}$$

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

$$\tau \mid Y, \ell, \sigma_{\text{GP}} \sim \mathcal{N}\left(\operatorname{cov}(Y, \tau)^\intercal \operatorname{cov}(Y)^{-1} Y, \sigma_\tau^2 - \operatorname{cov}(Y, \tau)^\intercal \operatorname{cov}(Y)^{-1} \operatorname{cov}(Y, \tau)\right) \tag{8}$$

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.

$$\mathbb{E}(\tau \mid Y, \ell, \sigma_{\text{GP}}, \sigma_\epsilon) = \sigma_\tau^2 \mathbb{I}_T^\intercal \left\{\sigma_\mu^2 + \sigma_\tau^2 \mathbb{I}_T \mathbb{I}_T^\intercal + \sigma_\beta^2 S S^\intercal + \mathbf{K} + \sigma_\epsilon^2 \mathbf{I}\right\}^{-1} Y \tag{9}$$

$$\operatorname{var}(\tau \mid Y, \ell, \sigma_{\text{GP}}, \sigma_\epsilon) = \sigma_\tau^2 - \sigma_\tau^2 \mathbb{I}_T^\intercal \left\{\sigma_\mu^2 + \sigma_\tau^2 \mathbb{I}_T \mathbb{I}_T^\intercal + \sigma_\beta^2 S S^\intercal + \mathbf{K} + \sigma_\epsilon^2 \mathbf{I}\right\}^{-1} \mathbb{I}_T \tag{10}$$

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 $x = b$:

$$\tau = \lim_{x \downarrow b} \mathbb{E}[y \mid X = s] - \lim_{x \uparrow b} \mathbb{E}[y \mid X = x] = \mathbb{E}[Y_T \mid X = b] - \mathbb{E}[Y_C \mid X = b] \tag{11}$$

Analogously, we focus on the treatment effect at the boundary ∂ between the treatment and control regions. ∂ is therefore a one-dimensional subset of $\mathscr{S}$. 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 $\boldsymbol{\partial} = \{\partial_1, \ldots, \partial_k\}$, $\partial_i \in \partial$. The extrapolation step then proceeds mechanically through multivariate-normal theory.

$$g_T(\boldsymbol{\partial}) \mid Y_T, S_T, \ell, \sigma_{\text{GP}}, \sigma_\epsilon \sim \mathcal{N}\left(\mu_{\boldsymbol{\partial}|T}, \Sigma_{\boldsymbol{\partial}|T}\right) \tag{12}$$

$$\mu_{\boldsymbol{\partial}|T} \equiv \text{cov}\left(g_T(\boldsymbol{\partial}), Y_T\right) \text{cov}\left(Y_T\right)^{-1} Y_T \tag{13}$$

$$\Sigma_{\boldsymbol{\partial}|T} \equiv \text{cov}\left(g_T(\boldsymbol{\partial})\right) - \text{cov}\left(g_T(\boldsymbol{\partial}), Y_T\right) \text{cov}\left(Y_T\right)^{-1} \text{cov}\left(Y_T, g_T(\boldsymbol{\partial})\right) \tag{14}$$

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 $g_C(\boldsymbol{\partial})$ 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

$$\mu_{\partial|Y} = \mathbb{E}\left(\tau(\boldsymbol{\partial}) \mid Y_T, Y_C\right) = \mu_{\boldsymbol{\partial}|T} - \mu_{\boldsymbol{\partial}|C} \tag{15}$$

$$\Sigma_{\partial|Y} = \text{cov}\left(\tau(\boldsymbol{\partial}) \mid Y_T, Y_C\right) = \Sigma_{\boldsymbol{\partial}|T} + \Sigma_{\boldsymbol{\partial}|C}\,. \tag{16}$$

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 $\sigma_\gamma^2 D^\intercal D$ 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.

5 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. We wish to set up a hypothesis test where the null hypothesis is that $\tau(\partial) = 0$ everywhere along the boundary. We start by creating a null model $\mathscr{M}_0$ that satisfies the sharp null hypothesis, specified as a single Gaussian process spanning the control and treatment regions, with the same kernel and hyperparameters obtained in the 2GP procedure. $\mathscr{M}_0$ 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 this model should be lower under $\mathscr{M}_0$ than under $\mathscr{M}_1$, the 2GP model as specified in equation (5). We therefore choose the difference in log-likelihoods as our test statistic

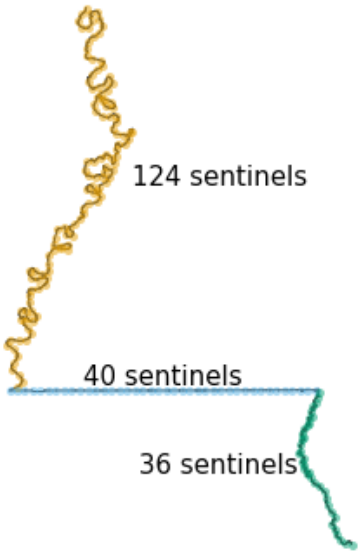
$$t = \log \mathbb{P}\left(Y_T, Y_C \mid \mathscr{M}_1\right) - \log \mathbb{P}\left(Y_T, Y_C \mid \mathscr{M}_0\right) \tag{17}$$

and wish to reject the sharp null hypothesis when t is high.

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

$$t^* = \log \mathbb{P}\left(Y_T^*, Y_C^* \mid \mathscr{M}_1\right) - \log \mathbb{P}\left(Y_T^*, Y_C^* \mid \mathscr{M}_0\right) \tag{18}$$

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 as a p-value.



mississippi counts

6 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.

$$\bar{\tau} \equiv \frac{\oint_{\partial} \tau(x)dx}{\oint_{\partial} dx} \tag{19}$$

$$\bar{\tau} \mid Y_T, Y_C, \sigma_{\text{GP}}, \sigma_\epsilon, \ell \sim \mathcal{N}\left(\mu_{\bar{\tau}|Y}, \Sigma_{\bar{\tau}|Y}\right) \tag{20}$$

$$\mu_{\bar{\tau}|Y} \approx \left(\mathbf{1}^\intercal \mu_{\boldsymbol{\partial}|Y}\right)/n_{\boldsymbol{\partial}} \tag{21}$$

$$\Sigma_{\bar{\tau}|Y} \approx \left(\mathbf{1}^\intercal \Sigma_{\boldsymbol{\partial}|Y}\, \mathbf{1}\right)/n_{\boldsymbol{\partial}}^2 \tag{22}$$

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, $\tau(\mathbf{s})$ 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 **??**). 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.

$$\tau^{IV} \mid Y_T, Y_C, \sigma_{\text{GP}}, \sigma_\epsilon, \ell \sim \mathcal{N}\left(\mu_{\tau^{IV}|Y}, \Sigma_{\tau^{IV}|Y}\right) \tag{23}$$

$$\mu_{\tau^{IV}|Y} \approx \left(\mathbf{1}^\intercal \Sigma^{-1}_{\boldsymbol{\partial}|Y} \mu_{\boldsymbol{\partial}|Y}\right) / \left(\mathbf{1}^\intercal \Sigma^{-1}_{\boldsymbol{\partial}|Y} \mathbf{1}\right) \tag{24}$$

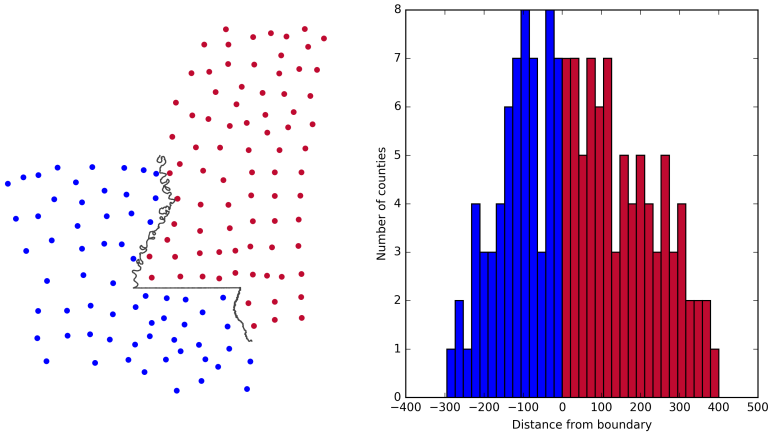
$$\Sigma_{\tau^{IV}|Y} \approx 1 / \left(\mathbf{1}^\intercal \Sigma^{-1}_{\boldsymbol{\partial}|Y} \mathbf{1}\right) \tag{25}$$

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.

7 Spatial advantage

- spreading units along a boundary doesn't necessarily reduce power
- multiple experiments interpretation

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 with an example. 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 7(a). For simplicity, we fix the hyperparameters to arbitrary values: $\sigma_\epsilon = \sigma_{\text{GP}} = 1.0$ and $\ell = 50\,\text{km}$. 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 (16) into the inverse-variance estimator (25). 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 7(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

9 Conclusion