A Bayesian Nonparametric Approach to Geographic Regression Discontinuity Designs

Maxime Rischard[1], Zach Branson[1], Luke Miratrix[2], and Luke Bornn[3]
[1] Department of Statistics, Harvard University
[2] Graduate School of Education, Harvard University
[3] Simon Fraser University

What is a Geographical Regression Discontinuity Design?
A GeoRDD is a natural experiment where units one one side of a geographical border are given a treatment while units on the other side are not. It's the spatial analog of univariate RDDs, which are an increasingly popular tool in econometrics and other social sciences.

What is your estimand?
The difference in expected potential outcomes from the control side of the border to the treatment side. It is defined along the border. Interestingly, the estimand is *functional*.

treatment effect τ: $\mathcal{B} \to \mathbb{R}$ defined as $\tau(\boldsymbol{b}) = \mathbb{E}[Y_{i\mathrm{T}} - Y_{i\mathrm{C}} \mid \boldsymbol{s}_i = \boldsymbol{b}]$

$\mathcal{B}$: border; point on the border; potential outcome for unit i under treatment; potential outcome for unit i under control; $\boldsymbol{s}_i$: spatial location of unit i

treatment indicator: $Z_i = \mathbb{I}\{\boldsymbol{s}_i \in \mathcal{A}_\mathrm{T}\}$; $\mathcal{A}_\mathrm{T}$: treatment area

How do you estimate the treatment effect?
Our GeoRDD framework proceeds by analogy to the univariate RDD:

Univariate RDD	GeoRDD
① fit a smooth **function** to the outcomes against the forcing variable on each side of the discontinuity,	① fit a smooth **surface** to the outcomes against the geographical covariates in each region,
② extrapolate the functions to the **discontinuity point**, and	② extrapolate the surfaces to the **border curve**, and
③ take the difference between the two extrapolations to estimate the treatment effect at the threshold point.	③ take the pointwise difference between the two extrapolations to estimate the treatment effect along the border.

To implement this strategy, we use Gaussian process regression:

potential outcome for unit i under treatment; smooth surface; spatial location of unit i; iid normal noise

$$Y_{i\mathrm{T}} = m_\mathrm{T} + f_\mathrm{T}(\boldsymbol{s}_i) + \epsilon_i \quad \text{and} \quad Y_{i\mathrm{C}} = m_\mathrm{C} + f_\mathrm{C}(\boldsymbol{s}_i) + \epsilon_i\,;$$

Gaussian process; mean shift

$$f_\mathrm{T}, f_\mathrm{C} \overset{\perp\!\!\!\perp}{\sim} \mathcal{GP}(0, k(\boldsymbol{s}, \boldsymbol{s}')) \quad \text{with} \quad k(\boldsymbol{s}, \boldsymbol{s}') = \sigma^2_{\mathrm{GP}} \exp\left(-\frac{(\boldsymbol{s}-\boldsymbol{s}')^\top(\boldsymbol{s}-\boldsymbol{s}')}{2\ell^2}\right).$$

covariance between two locations; e.g. Squared Exponential kernel

Can we estimate the average treatment effect (ATE)?
We have to define which average. The geometry of the GeoRDD makes the choice non-trivial. We consider the class of weighted means over the border:

ATE; weight function; discretized at a set of R "sentinel" points

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

e.g. τ^{UNIF} : $w_\mathcal{B}(\boldsymbol{b}_R) = 1$

population density

τ^ρ : $w_\mathcal{B}(\boldsymbol{b}_R) = \rho(\boldsymbol{b}_R)$

posterior covariance of treatment effect

τ^{INV} : $w_\mathcal{B}(\boldsymbol{b}_{1:R}) = \Sigma^{-1}_{\boldsymbol{b}_{1:R}|Y}\,\mathbf{1}_R$

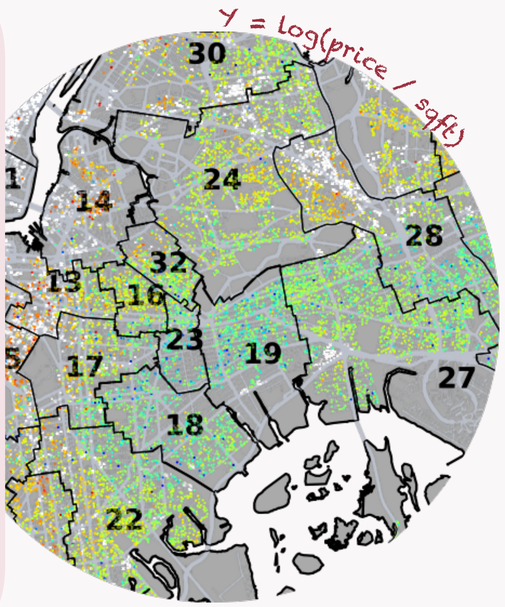
We prefer the inverse-variance weighted average treatment effect. **Ask me why!**

Illustration: Estimating the Cost of a NYC School District

The Data. A year of house sales in NYC, shown on the adjacent map. Each circle is a sale, coloured according to the price per square foot. Also shown are the outlines of the NYC school districts, numbered from 1 to 32. We also have the building class at time of sale.

The Question. It is a common belief that school districts impact real estate prices, as parents are willing to pay more to live in better districts.
Can we measure a discontinuous jump in house prices across the borders separating school districts?

Exploratory Analysis. Look closely at the border between districts 19 and 27. It looks like the dots in district 19 are bluer than those in district 27. But can we detect and quantify this difference at the border, and is it significant?

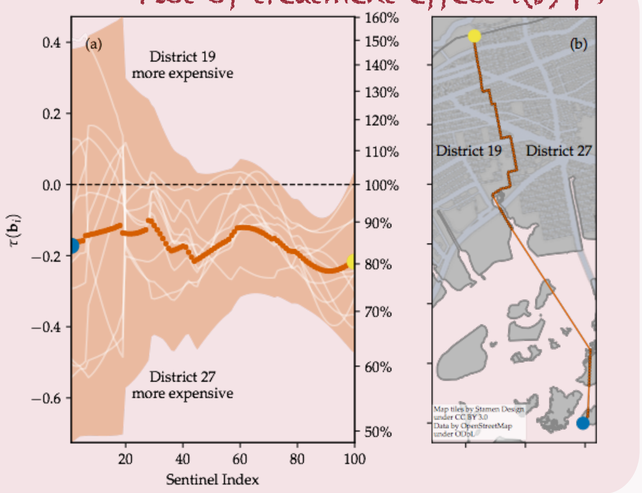


After fitting the hyperparameters by maximizing the marginal likelihood, and accounting for covariates (ask me how!), we extract the posterior mean and covariance of the treatment effect at 100 sentinel locations on the border between districts 19 and 27, shown on the right and in 3D below.

The inverse-variance weighted ATE is -0.19, which corresponds to an almost 20% increase in houses prices going from district 19 to district 27.

The difference is significant with p=0.002. So school districts cause differences in house prices? Not quite… the causal story is murkier. Ask me for details.

Plot of treatment effect $\tau(b) \mid Y$



The posterior treatment effect is analytically tractable and easily computed.

$$\tau(\boldsymbol{b}_{1:R}) \mid Y, \boldsymbol{\theta} \sim \mathcal{N}\left(\mu_{\boldsymbol{b}_{1:R}|Y}, \Sigma_{\boldsymbol{b}_{1:R}|Y}\right)$$

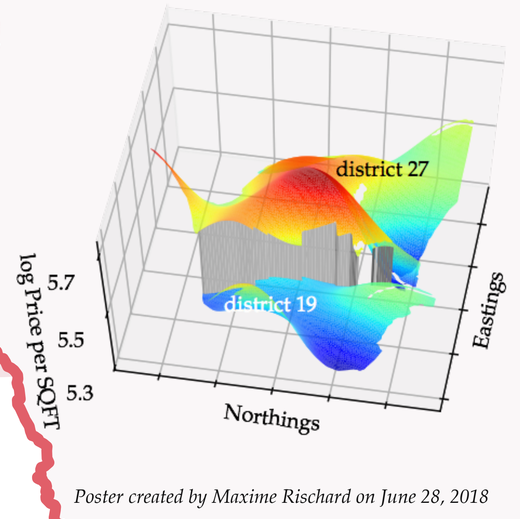
discretized at a set of R "sentinel" points; data; GP hyperparams; mean vector and covariance matrix

Is the treatment effect significant?
How do we know we're not just staring at noise? We can look at the tail probability of the ATE posterior distribution and derive a pseudo-p-value

$$\tilde{p}^w = 2\Phi\left(-\left|\mu_{\tau^w|Y}\right| \Big/ \sqrt{\Sigma_{\tau^w|Y}}\right)$$

normal CDF; posterior mean and variance of avg. treatment effect

But it's not a valid p-value in the frequentist sense. We do better by treating the posterior mean ATE as a test statistic and deriving (analytically or using a bootstrap) its distribution under a parametric null distribution. This gives us a valid frequentist test derived from a Bayesian posterior!



Poster created by Maxime Rischard on June 28, 2018