A Bayesian Nonparametric Approach to Geographic Regression Discontinuity Designs: Estimating the Cost of a School District

Luke Miratrix, Maxime Rischard, Zach Branson, and Luke Bornn

October 1, 2019

1 Background/Context

It is a common belief that school districts have an impact on real estate price, as parents are willing to pay more to live in districts with better schools. In this project we ask: can we measure the discontinuous jump in house prices across a border separating school districts?

This type of question is perhaps best answered with a *Geographic Regression Discontinuity Design* (GeoRDDs) (e.g. Keele and Titiunik, 2015; Keele et al., 2015) a type of Regression Discontinuity Design (see, e.g., Thistlethwaite and Campbell (1960); Hahn et al. (2001); Imbens and Lemieux (2008) for core and overview papers or Matsudaira (2008); Ludwig and Miller (2007); Li et al. (2015) for examples in education) tailored to the multidimensional nature of a spatial setting such as this one. GeoRDDs arise when a treatment is assigned to one region, but not to another adjacent region. For outcomes that vary spatially, simple direct comparison units on either side of the boundary is invalid due to spatial confounding. However, under smoothness assumptions, we can account for this confounding and extract a natural experiment.

2 Purpose/Objective/Research Question

Previous research has focused on extending methods developed for 1D RDDs to GeoRDDs. E.g., some have used the signed distance from the border as the forcing variable in a 1D RDD (Martorell, 2004; Robinson, 2011; Cohodes and Goodman, 2012), but the resulting estimator is spatially confounded. In this work, we emphasize the importance of the geographical aspect of the problem, and therefore draw from the spatial statistics literature, which brings a rich set of tools designed to model spatial correlations. In particular, we use Gaussian process regression (GPR, or kriging in the spatial statistics literature) to fit the smooth surfaces to the outcomes. This models the entire response surface (Papay et al., 2011; Dee, 2012; Papay et al., 2014), but hopefully with less structure to avoid misspecification.

Overall, we aim to generate an inferential approach that adheres to the core principle of an RDD by extrapolating, with local estimates of trend, units on either side of a boundary to the boundary itself without strong overall modeling assumptions. Even with this piece in place, several difficult questions remain: what is the estimand of interest? How should different points on the boundary be weighted appropriately? As part of this project we carefully think through the implied estimands of different approaches, and argue for careful attention to possible treatment variation along the border of interest.

3 Data

We use house sales data from New York City. The dataset includes sale price, building class, and property address. We also obtained files that delineated different sub-districts so we could locate all houses by what sub-district they were a member of. After cleaning our data and removing non-family units, we had a resulting dataset of 19,578 sales. See Figure 1

4 Methodology

In a Geographic Regression Discontinuity context, we have impacts along entire boundaries. These boundaries are the black lines in Figure 1; we might imagine a different differential price at each point due to spatial trends. Our primary estimand is the local average impact at point x, where x lies on any point separating our regions of interest:

$$\tau(\mathbf{x}) = \mathbb{E}[Y_i(1) - Y_i(0) | X_{i1} = x_1, X_{i2} = x_2], \quad \text{where } \mathbf{x} \in \mathcal{B} \tag{1}$$

Our estimand, in other words, is a function.

Our framework for the estimation of $\tau(\mathbf{x})$ proceeds in three steps: (1) fit a smooth surface on either side of the border, (2) extrapolate the surfaces to the border, and (3) take the difference of the two extrapolations to estimate the treatment effect along the border. To do this we use Gaussian process regression, extending the use of Bayesian approaches in RDDs of Branson et al. (2019). Define $\mu_T(x) \equiv \mathbb{E}[Y_i(1) | X = x]$ and $\mu_C(x) \equiv \mathbb{E}[Y_i(0) | X = x]$ as the mean response functions for treatment and control located at some point $X = x$, respectively. These are the unknown functions we would like to estimate. We then assume that the treatment and control responses are generated as

$$\begin{aligned} &Y_i(1) = \mu_T(X_i) + \epsilon_{iT}, \quad \text{and} \quad Y_i(0) = \mu_C(X_i) + \epsilon_{iC}, \quad \text{where} \\ &\epsilon_{iT} \overset{iid}{\sim} N(0, \sigma^2_{yT}), \quad \text{and} \quad \epsilon_{iC} \overset{iid}{\sim} N(0, \sigma^2_{yC}) \end{aligned} \tag{2}$$

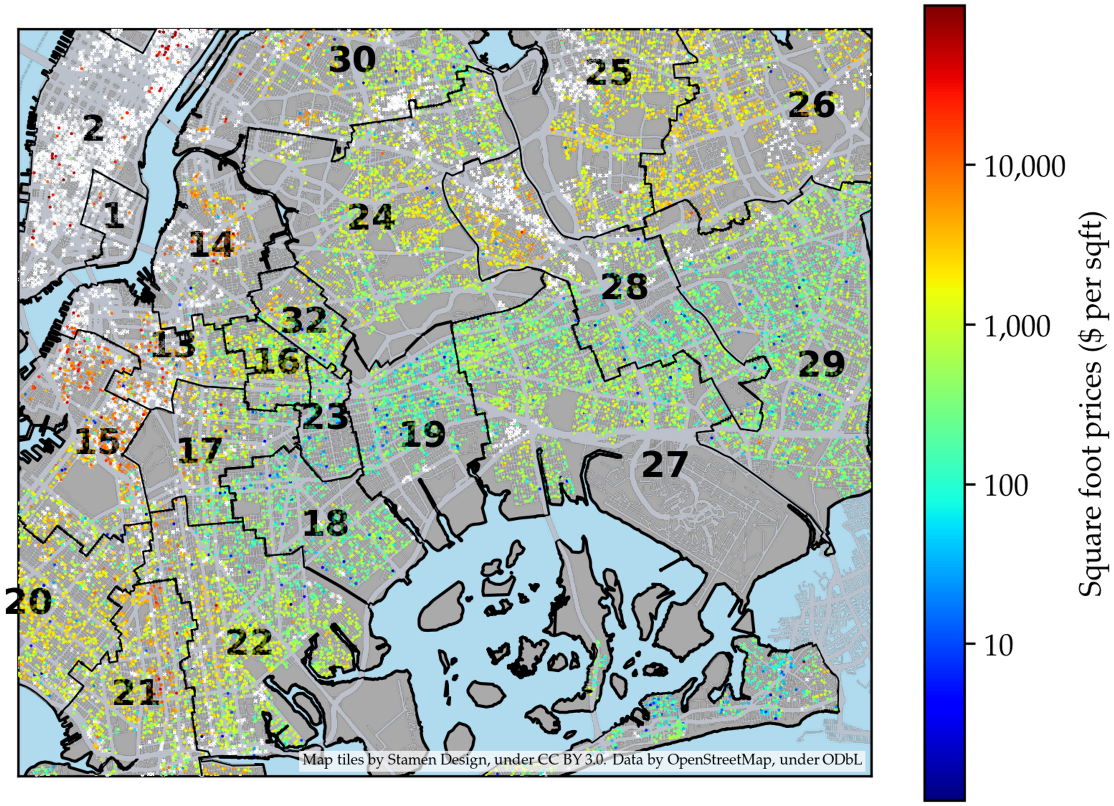
30
25
26
2
1
14
24
28
32
13
16
29
23
19
15
17
27
18
20
22
21
Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL
10,000
1,000
100
10
Square foot prices ($ per sqft)

Figure 1: Map of property sales in New York City. Each dot is a sale, and its color indicates the price per square foot. White crosses indicate sales of properties with missing square footage, which are therefore excluded from the analysis. School district boundaries are shown, and each district is labeled by its number.

Local linear regression methods (classic RDD) make the same above assumption, in addition to modeling $\mu_T(x)$ and $\mu_T(x)$ via weighted least squares. Instead of specifying a functional form we place a Gaussian process prior on both of these functions:

$$\begin{aligned} \mu_T(x) &\sim \mathrm{GP}(m_T(\mathbf{X}), K_T(\mathbf{X}, \mathbf{X}')) \\ \mu_C(x) &\sim \mathrm{GP}(m_C(\mathbf{X}), K_C(\mathbf{X}, \mathbf{X}')) \end{aligned} \tag{3}$$

where we treat the two Gaussian process priors in (3) as independent.[1]

Once this is done, we aggregate the border-specific impacts to obtain overall averages. We write this averaging as an intergral along the border, using a weighting function to appropriately account for varying population density of units and for the curvature of the

[1] Letting the treatment and control mean response functions be *a priori* independent is analogous to fitting two separate local linear regressions—one in the treatment group, one in the control group—which is by far the common practice in RDDs (Imbens and Lemieux, 2008).

border itself:

$$\tau^{pop} = \frac{\int_{\mathbf{x}\in\mathcal{B}} \rho(\mathbf{x})\tau(\mathbf{x})\partial\mathbf{x}}{\int_{\mathbf{x}\in\mathcal{B}} \rho(\mathbf{x})\partial\mathbf{x}} \tag{4}$$

Selecting the weighting function $\rho(x)$ has surprising pitfalls. Simply averaging the treatment effect uniformly along the border yields an estimand that is inefficient and undesirably sensitive to the topology of the border. For example, regions with lots of wiggles will have higher weight not due to more units in that region, but simply due to the border itself. We discuss several options, of which we advocate selecting $\rho(x)$ to be either population density or overall precision.

To test against the null hypothesis of zero treatment effect along the border, we also develop a test based on the posterior distribution of the LATE. To ensure good frequentist properties we calibrate the test, obtaining its distribution under the null model, with a parametric bootstrap.

5 Findings/Results

Figure 2 gives our results for a single border between two districts. We estimate that the same house located near the border will on average fetch a significantly and substantively 20% higher price in district 27 than in district 19. However, this effect cannot be attributed solely to the school district reputation: it also separates Brooklyn and Queens.

We found such significant effects between many of the 26 other pairs of adjacent school districts examined. However, frequently physical barriers such as parks, commercial zones, and major roads can separate neighborhoods, keep data away from the borders, and break the stationarity assumption of the spatial model, which casts doubt on the legitimacy of the estimated treatment effects.

6 Conclusions

The use of GPR to analyze GeoRDDs gives flexibility and extensibility to the general discontinuity approach, and also naturally incorporates spatial correlation which other methods frequently do not take into account. That being said, one must be careful using these methods due to the nature of actual geographic data.

Our GPR can be extended in various ways. For example, if the outcomes are binary, proportions, or counts, then binomial or Poisson likelihoods could be substituted for the i.i.d. normal likelihood used above. This approach also has important connections to the two-dimensional RDD designs used in some testing contexts, such as when high stakes consequences happen when either of two different tests is failed (see, e.g., Ou (2010); Papay et al. (2010); Reardon et al. (2010); Reardon and Robinson (2012); Papay et al. (2014)); this second stage of work is ongoing. Also see Porter et al. (2017) for a good summary of existing work on this latter problem, along with a good comparison of the methodology.

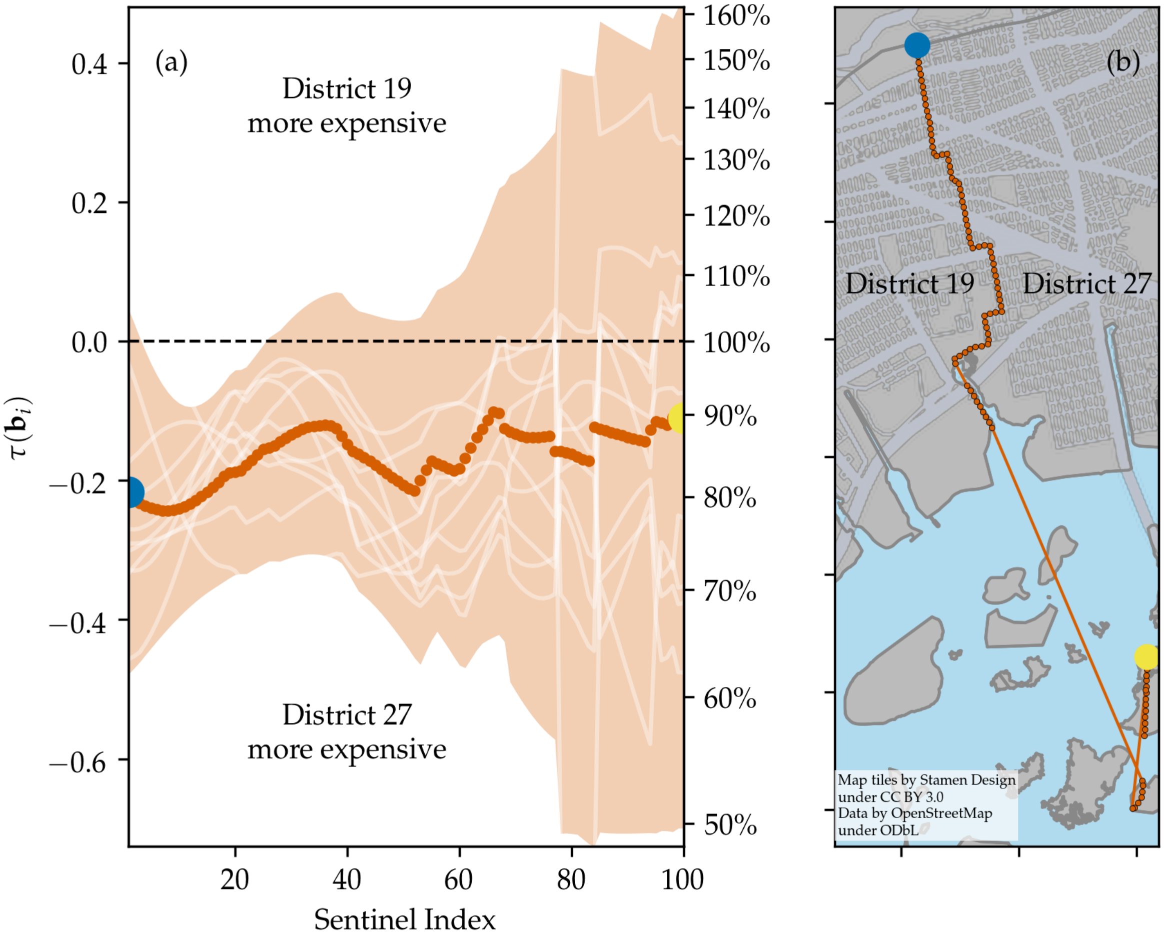

Figure 2: Cliff face estimator for the school district effect on house prices per square foot between district 27 and district 19, with 95% credible envelope. The left axis is in the scale of log prices per square foot; positive values correspond to houses near the border being more expensive in district 19 than 27. The right axis shows the corresponding ratio of the price. (b) The map of sentinels used to estimate impacts along the boundary, evenly spaced. The northernmost sentinel (shown as a blue circle in both plots) has index 1, while the last sentinels (shown in yellow) is on Rulers Bar Hassock.

References

Branson, Z., M. Rischard, L. Bornn, and L. W. Miratrix (2019). A nonparametric bayesian methodology for regression discontinuity designs. *Journal of Statistical Planning and Inference*.

Cohodes, S. R. and J. S. Goodman (2012). First degree earns: The impact of college quality on college completion rates. *HKS Faculty Research Working Paper Series*.

Dee, T. (2012). School turnarounds: Evidence from the 2009 stimulus. Technical report, National Bureau of Economic Research.

Hahn, J., P. Todd, and W. Van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica 69*(1), 201–209.

Imbens, G. W. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics 142*(2), 615–635.

Keele, L., R. Titiunik, and J. R. Zubizarreta (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 178*(1), 223–239.

Keele, L. J. and R. Titiunik (2015). Geographic boundaries as regression discontinuities. *Political Analysis 23*(1), 127–155.

Li, F., A. Mattei, and F. Mealli (2015). Evaluating the causal effect of university grants on student dropout: evidence from a regression discontinuity design using principal stratification. *The Annals of Applied Statistics*, 1906–1931.

Ludwig, J. and D. L. Miller (2007). Does head start improve children's life chances? evidence from a regression discontinuity design. *The Quarterly journal of economics 122*(1), 159–208.

Martorell, F. (2004). Do high school graduation exams matter? a regression discontinuity approach. *Unpublished manuscript. University of California Berkeley. Retrieved May 28*, 2012.

Matsudaira, J. D. (2008). Mandatory summer school and student achievement. *Journal of Econometrics 142*(2), 829–850.

Ou, D. (2010). To leave or not to leave? a regression discontinuity analysis of the impact of failing the high school exit exam. *Economics of Education Review 29*(2), 171–186.

Papay, J. P., R. J. Murnane, and J. B. Willett (2010). The consequences of high school exit examinations for low-performing urban students: Evidence from massachusetts. *Educational Evaluation and Policy Analysis 32*(1), 5–23.

Papay, J. P., R. J. Murnane, and J. B. Willett (2014). High-school exit examinations and the schooling decisions of teenagers: Evidence from regression-discontinuity approaches. *Journal of Research on Educational Effectiveness 7*(1), 1–27.

Papay, J. P., J. B. Willett, and R. J. Murnane (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics 161*(2), 203–207.

Porter, K. E., S. F. Reardon, F. Unlu, H. S. Bloom, and J. R. Cimpian (2017). Estimating causal effects of education interventions using a two-rating regression discontinuity design: Lessons from a simulation study and an application. *Journal of Research on Educational Effectiveness 10*(1), 138–167.

Reardon, S. F., N. Arshan, A. Atteberry, and M. Kurlaender (2010). Effects of failing a high school exit exam on course taking, achievement, persistence, and graduation. *Educational Evaluation and Policy Analysis 32*(4), 498–520.

Reardon, S. F. and J. P. Robinson (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness 5*(1), 83–104.

Robinson, J. P. (2011). Evaluating criteria for english learner reclassification: A causal-effects approach using a binding-score regression discontinuity design with instrumental variables. *Educational Evaluation and Policy Analysis 33*(3), 267–292.

Thistlethwaite, D. L. and D. T. Campbell (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology 51*(6), 309.