

Robust inference for geographic regression discontinuity designs: assessing the impact of police precincts

Emmett B. Kendall¹ , Brenden Beck² and Joseph Antonelli³ 

¹Department of Statistics, NC State University, Raleigh, NC, USA

²School of Criminal Justice, Rutgers University Newark, Newark, NJ, USA

³Department of Statistics, University of Florida, Gainesville, FL, USA

Address for correspondence: Emmett B. Kendall, Department of Statistics, NC State University, 2311 Stinson Dr, Raleigh, NC 27607, USA. Email: ebkendal@ncsu.edu

Abstract

We study variation in policing outcomes attributable to differential policing practices in New York City (NYC) using geographic regression discontinuity designs (GeoRDDs). By focusing on small geographic windows near police precinct boundaries, we can estimate local average treatment effects of police precinct practices on arrest rates. We propose estimands and develop estimators for the GeoRDD when the data come from a spatial point process. Standard GeoRDDs rely on continuity assumptions of the potential outcome surface or a local randomization assumption within a window around the boundary. These assumptions, however, can easily be violated in real applications. We develop a novel and robust approach to testing whether there are differences in policing outcomes that are caused by differences in police precinct policies across NYC. Importantly, this approach is applicable to standard regression discontinuity designs with both numeric and point process data. This approach is robust to violations of traditional assumptions made and is valid under weaker assumptions. We use a unique form of resampling to provide a valid estimate of our test statistic's null distribution even under violations of standard assumptions. This procedure gives substantially different results in the analysis of NYC arrest rates than those that rely on standard assumptions.

Keywords: causal inference, criminology, point process, spatial data

1 Introduction

Policing varies across political boundaries, such as state or city borders. Such differences are expected, but we know very little about whether smaller, sub-municipal boundaries like police districts, precincts, and service areas also influence police outcomes (Klinger, 1997). This lack of research persists despite police officers reporting that their behaviour and perception are influenced by precinct boundaries (Hassell, 2007). Police have wide discretion when choosing to make an arrest, so arrest rates could be particularly susceptible to spatial variation (Herbert, 1996). Understanding whether precincts police differently has important implications for equity and policy. Variation in policing between cities is tolerated because it results, in part, from the electoral choices of residents. Variation within cities, however, generates a more troubling kind of inequality. Residents do not vote for their police precinct commander and they expect treatment equal to that of people in other neighbourhoods. Policing variation generated by differences in police precinct policies or practices might

Received: February 3, 2024. Revised: June 28, 2025. Accepted: July 22, 2025

© The Royal Statistical Society 2025.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

also compound other forms of spatial inequity like residential segregation, racial bias, or high crime areas (Bell, 2020).

Another potential consequence of between-precinct variation in policing is diminished policy efficacy. Some recent police reform efforts have attempted to reduce the number of pedestrian stops and frisks, reduce the use of deadly force, and improve police-community relations. Most such reforms are implemented at the city scale, but if significant variation exists between police precincts, such a one-size-fits-all approach might fail. Even place-based interventions like hot-spots policing target high-crime areas and ignore precinct boundaries. Research has long understood the salience of micro places in shaping crime, but less is known about how local characteristics shape policing. It is likely that police behaviour, like criminal behaviour, varies greatly by place. In this study, we examine the 77 precincts of the New York City Police Department (NYPD) to determine if arrest rates differ across precincts, thus providing insight into potential arresting practice differences among officers in different precincts.

This is a difficult question to answer because the regions each police precinct covers are different from one another with respect to important demographic and criminological variables. One precinct might have different low-level arrest rates than another because it has higher crime rates, more targets for theft, or more transient populations. Therefore, we have to isolate the effect of the police precinct practices or policies themselves. Randomization is the gold standard for drawing causal conclusions, but while these are occasionally available in the criminology literature to evaluate policies like hot spots policing (Puelz et al., 2022), in many scenarios, they are not available or feasible. When evaluating the impact of police precinct practices, we cannot randomize individuals to a police precinct by forcing them to live or work in certain areas of a city. The ubiquity of observational studies has led to a wide range of approaches to estimate causal effects under as weak of assumptions as possible. Common approaches are difference-in-difference estimators (Ashenfelter & Card, 1984; Lechner, 2011), the regression discontinuity design (Cattaneo et al., 2019; G. W. Imbens & Lemieux, 2008; Thistlethwaite & Campbell, 1960), interrupted time series analysis (Bernal et al., 2017; Cook et al., 1979), and synthetic control analysis (Abadie et al., 2010), among others. In the context of policing and criminology, these ideas have been used to address important issues, such as whether increased oversight of police leads to increases in crime and decreased effectiveness of the police force (Ba & Rivera, 2019), quantifying the impact of a penalty system for drivers in Italy on traffic incidents and traffic-related fatalities (De Paola et al., 2013), or estimating the heterogeneous effects of neighbourhood policing (Antonelli & Beck, 2023; Beck et al., 2022).

In this study, we focus on the regression discontinuity design and its extensions to geographic settings and point process data. For an in-depth review of standard regression discontinuity designs and implementation details, see G. W. Imbens and Lemieux (2008) and Cattaneo et al. (2019). The traditional regression discontinuity design occurs when treatment assignment is either partially or completely determined by a pretreatment covariate, typically referred to as the running or score variable. There exists a cut-off value of this running variable, above which units receive treatment, and below which units receive the control. The fundamental idea is that units within a small distance around the cut-off value form a locally randomized experiment (Mattei & Mealli, 2017). The estimand of interest is a local treatment effect at the cut-off value, and nearby observations are used to extrapolate what would happen both under treatment and control at this boundary value. This approach has been extended to multivariate running variables such as the results of two types of educational tests (Matsudaira, 2008). A specific example of a bivariate running variable is found in the geographic regression discontinuity design (GeoRDD), where latitude and longitude are used to determine treatment assignment. Important aspects specific to the geographic design are highlighted in L. J. Keele and Titiunik (2015). This design has been used to estimate the effect of private police departments on crime (MacDonald et al., 2016), the impact of voter initiatives on voter turnout (L. Keele et al., 2015), the effect of the Civil Rights Act of 1875 (Harvey, 2020), and whether school districts impact housing prices (Rischard et al., 2021).

Regression discontinuity designs rely on assumptions that the potential outcomes are smooth at the cut-off value or that treatment behaves as if it were randomized within a window around the cut-off value. To assess the validity of these assumptions, a number of falsification tests have been proposed. A negative control approach is to treat an observed covariate as an outcome, where we know the treatment should not affect this covariate, and estimate the treatment effect on this

covariate to see if the approach correctly estimates a null association (Lee et al., 2004). Another issue is that the running variable can be manipulated by subjects if they are aware of the cut-off value, and this can be evaluated by checking if the distribution of the running variable is continuous at the cut-off (Cattaneo et al., 2017; McCrary, 2008). Other approaches examine the sensitivity of results to bandwidth selection (Lemieux & Milligan, 2008), as robustness of results to this choice provides increased belief in the resulting findings.

In this work, we develop two key contributions to the literature on GeoRDDs. For one, we extend the GeoRDD literature to settings where the outcome is a point process, and causal estimands and assumptions are defined in terms of intensity surfaces. Standard GeoRDD approaches do not apply here because they are designed for settings with a numeric outcome (such as voting behaviour or test scores) measured at spatial locations. In contrast, we do not have a numeric outcome; instead, our outcome is the location of events, and we are interested in studying the expected number of events in specific subsets of the spatial domain. This necessitates a modification of key assumptions and estimands, and relies on distinct estimation strategies that we develop. Our second contribution, which is applicable to general GeoRDD settings, not just the point process one seen here, is that we provide valid hypothesis tests for causal estimands under certain violations of existing assumptions typically utilized in the GeoRDD. By using a novel resampling scheme, our approach allows for violations of the assumptions that treatment is as if randomized within a window around the cut-off point of the running variable, or that the potential outcomes are smooth at the cut-off. We exploit a large spatio-temporal data set of crime and arrest data in New York City (NYC) to find streets that behave similarly to precinct boundaries, but by definition have no effect of police department practices as they are fully contained within a single police precinct. We use these streets to construct a null distribution that accounts for violations of local randomization or continuity assumptions and provides a valid hypothesis test of individual precinct effects, as well as a test for the overall degree of variation in policing across NYC.

The relevant R code and data for reproducing all numerical results presented in this article are available at [ebkendall.github.io/research.html](https://github.com/ebkendall/research.html).

2 Policing data in NYC and preliminary analyses

Our analyses draw on two data sources made public by the NYPD: NYPD Arrest Data and NYPD Complaint Data. Both provide information at the incident level with geolocated, address data for all arrests and crimes reported to the police in the years 2010–2018. The NYPD is divided into 77 police precincts, each patrolling a particular geographic area of the city. We exclude the precinct corresponding to Central Park, which does not have any residents, leading to 76 precincts in our analysis. Our goal is to use these data to understand whether there is variability in arresting practices across police precincts in NYC, and whether individuals are more or less likely to be arrested depending on which precinct's police force they are exposed to. Before describing our problem in more detail, it is important to emphasize that police precincts can refer to both geographic areas, as well as the police organizational unit that polices that geographic area. Whenever we refer to effects of police precincts, we are referring to effects of decisions, policies, or practices of the police department within that precinct, not the effects of the geographic area itself. This is a well-defined treatment variable of interest to study, because police departments have different police commanders, different policies, and other features that may affect arrest rates.

Using these data, we can visualize both when and where arrests occur as well as the precinct from which the arresting officer originates. Figure 1 highlights the arrest data for Precinct 77 in NYC during the year 2014, both with and without the roadmap of the city overlaid on the figure using the R package *ggmap* (Kahle & Wickham, 2013). This figure reveals features of the data we will leverage in our approach to inference. Both arrests and crimes fall directly on streets, and are not spread out across the entire spatial domain. Note that because the observed spatial locations are restricted to a grid, a single spatial location can correspond to multiple arrests or crimes. Additionally, precinct boundaries are streets themselves, and there are arrests that fall directly on these boundaries.

Suppose that we are interested in estimating whether there is a difference in arresting practices between two neighbouring police precincts. Figure 2 plots the number of arrests divided by the number of crimes within each precinct, and we can see significant variation across police



Figure 1. Arrest locations from officers of Precinct 77 in New York City (NYC) during the year 2014.

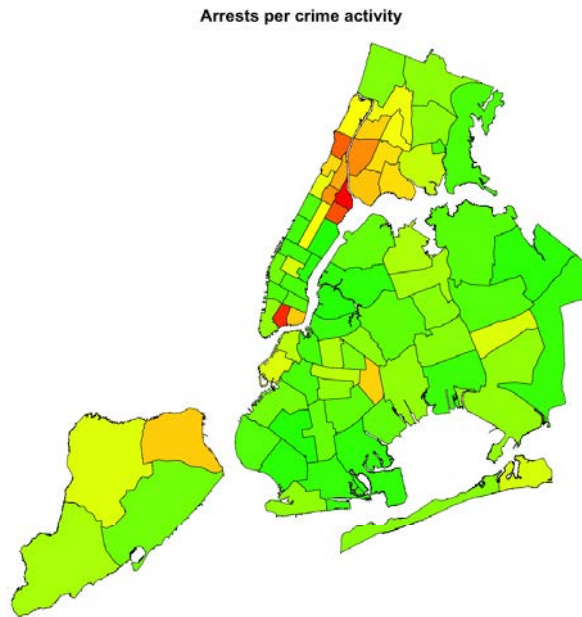


Figure 2. Number of arrests divided by the number of crimes over the entire study period, separated by precinct.

precincts. This could be driven by many factors, such as the types of crime committed within each precinct. Alternatively, it could be that separate police precincts have distinct policies or practices that lead to such variation, and our goal is to study whether it is the latter. Formally, we want to assess whether there is a causal effect of police precinct on arrest rates. Given the lack of rich covariate information, but a high degree of spatial resolution, one might choose to use a GeoRDD to estimate the effect of police precincts between two neighbouring regions. The GeoRDD leverages the fact that nearby areas should be similar with respect to important unobserved characteristics that are associated with arrest rates. In this context, this assumption would be satisfied if areas on either side of the border between the two precincts are similar to each other. If individuals generally do not choose where to live based on which police precinct their home falls in, which is a reasonable assumption for many police precinct boundaries, then differences in arrest rates would be attributable to differences in police precinct practices.

Existing GeoRDD methodology does not immediately apply to the point process data seen here. We formalize estimands and methodology for estimation in this setting in Section 3, but as a preliminary analysis we can draw a buffer around the border of two precincts and study the difference in arrest rates made by each precinct within the buffer over time. An illustration of the setup can be found in Figure 3, where arrests are colour-coded based on the arresting officer's precinct. If there is no variation in policing practices across precincts, then under certain causal assumptions

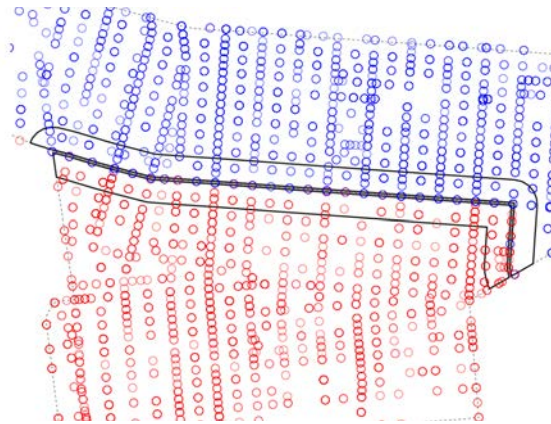


Figure 3. Arrest locations near the border of Precincts 71 and 77.

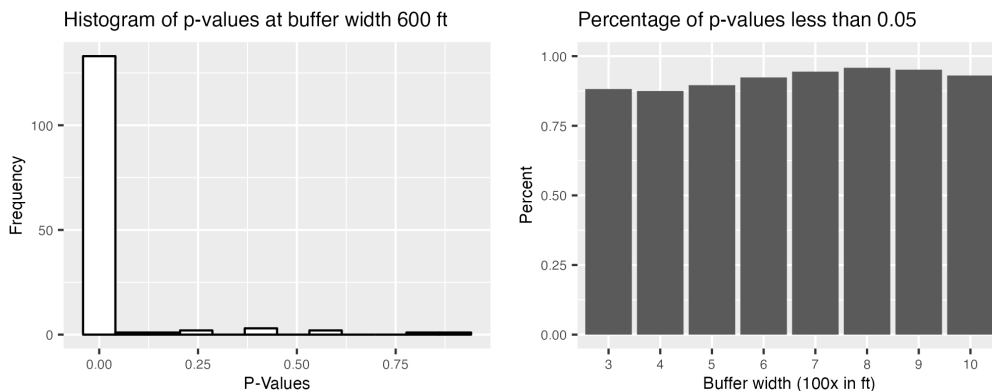


Figure 4. Distribution of p -values across all 144 borders at a buffer width of 600 feet (left) along with the per cent of rejections as a function of buffer size around the border (right) for counts of arrests.

detailed in Section 3, we would expect to find a significant difference in arresting practices between two neighbouring precincts with probability α , where α is the prespecified type I error. Given that there are 144 precinct–precinct borders of interest, we would expect to see roughly $\alpha \times 144$ significant associations and the distribution of p -values across these tests to be roughly uniformly distributed if police precincts do not affect arrest rates.

Suppose buffers with radius lengths of 300–1,000 feet are drawn around each border. Then, the number of arrests made by each of the two precincts within the bounds of the buffer zone is counted so that we can test if there is a significant difference in arrest counts. Specifically, if we let (Y_0, Y_1) be the number of arrests on the two sides of the border, we expect that $Y_0 \sim \text{binomial}(Y_0 + Y_1, 0.5)$ under the null hypothesis of no precinct effect. We can use this result to test the hypothesis that the expected number of arrests on either side of the boundary is the same. Figure 4 shows the distribution of p -values from this test for a 600-foot buffer, as well as the percentage of rejections out of 144 borders as a function of the buffer width. We see that 92.4% of the p -values are $<.05$ when we use a buffer radius of 600 feet, with similar percentages for other buffer widths, and the p -value histogram is far from uniform. This may occur because: (1) there truly are large differences in arresting practices across NYC police precincts, (2) either the causal assumptions do not hold or the statistical test is invalid, or (3) some combination of these two.

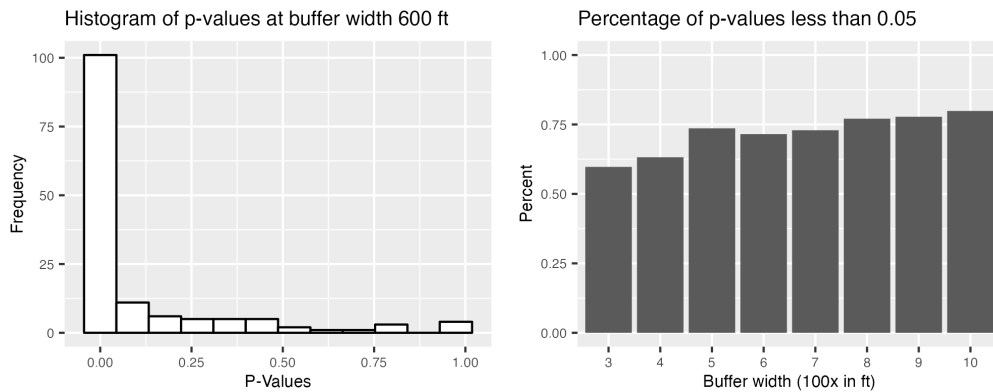


Figure 5. Distribution of p -values across all 144 borders at a buffer width of 600 feet (left) along with the per cent of rejections as a function of buffer size around the border (right) for the negative control outcome looking at counts of trees.

2.1 Negative control analyses

While we cannot know whether the significant effects on arrest rates are due to a true causal effect of police precincts or violations of key assumptions, we can examine similar data that is known to have no police precinct effect. As a negative control outcome, we consider the 2015 Street Tree Census data set provided by *NYC Open Data*. This data set provides the location of all trees in NYC, and has a similar structure to the crime/arrest data, as all locations are recorded at street locations. Clearly, police departments do not decide where to place trees in a precinct, which might lead one to assume that the difference in tree counts on either side of any two bordering precincts is expected to be small. This intuition is not correct in this instance, as Figure 5 shows very similar results to that of the arrest data. The p -value distribution across precinct boundaries is highly skewed, and the percentage of significant findings, even for small buffer widths of 300 feet, is well above the desired type I error rate. This motivates the need to develop a methodology that is robust to violations of assumptions on the similarity of the populations on both sides of precinct boundaries, and that is able to provide valid inference in this setting.

3 Causal inference for spatial point processes with GeoRDDs

The observed data on our outcome of interest consists of a set of geographic locations corresponding to each arrest in NYC between 2010 and 2018. Our data can be thought of as being generated from a point process, in that an arrest will randomly occur at some time point, and at a given location. Since we are interested in studying variability in arrest rates across space, rather than time, we aggregate across the temporal component to focus solely on space. We denote geographical coordinates by $S = (S_1, S_2)$, which correspond to latitude and longitude. Therefore, the observed outcomes are given by $\mathcal{S} = \{S_1, S_2, \dots, S_N\}$ where S_1, S_2, \dots, S_N represent each location at which an outcome is observed (e.g. where each arrest occurs). We use $|\mathcal{S}| = N$ to denote the cardinality of this set, which corresponds to the number of events in the entire spatial domain studied. To reiterate, standard GeoRDD estimands and inferential strategies do not apply here because we do not observe a numeric outcome value at these locations. Instead, the outcome is the location of arrests in NYC, and its distribution is governed by an unknown point process.

At times, we focus on specific subregions of the entire spatial domain and therefore we can define corresponding region-specific quantities. Denote R as a subregion under study, such as the region within 300 feet of the boundary between two precincts. Let $Y(R) = \sum_{i=1}^N \mathbf{1}(S_i \in R)$ represent the number of outcomes that occurred in region R . We assume that our data follow an inhomogeneous point process (Daley & Vere-Jones, 2003) with intensity function given by $\lambda(s)$, where s denotes a spatial location. Specifically, this implies that

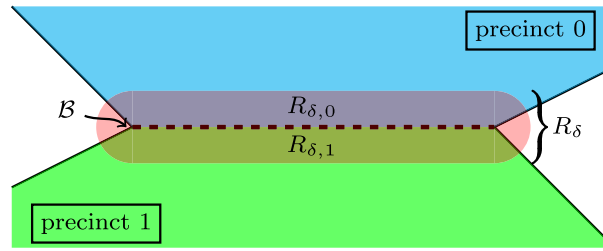


Figure 6. Illustration of the components surrounding the border.

$$E(Y(R)) = \int_R \lambda(s) ds, \quad (1)$$

for any region R in the domain of interest.

3.1 Single border estimands

Now that we have introduced notation relevant to our problem, we can formally define potential outcomes and causal estimands in the context of spatial point process data. Our interest is in whether arrest rates differ across police precincts throughout NYC, and we can first answer this question by focusing on two adjacent precincts at a time, which we refer to as *precinct 0* and *precinct 1*. Note that most precincts have more than one neighbouring precinct; therefore, in the event that a precinct has, say, four neighbours, then there exists four distinct estimands for each adjacent precinct pair. We can then extend these ideas to all adjacent precinct pairs in NYC in Section 3.8. First, let \mathcal{B} denote the spatial boundary separating precinct 0 and precinct 1 (i.e. the border between the two precincts). Further, define a distance function $d(s, \mathcal{B})$, which is the shortest distance between a point in space, s , and the boundary, \mathcal{B} . Next, define the treatment variable $T(s)$ to be an indicator of whether a location is policed by police precinct 0 or 1. Clearly, this is a deterministic function of s as $T(s) = 1(s \in \text{precinct 1})$. This is referred to as a sharp regression discontinuity design as the forcing variable (s in this setting) completely determines the treatment assignment (Trochim, 1990). With both the distance metric and treatment defined, we can more formally characterize the regions relevant to our study. Define

$$\begin{aligned} R_{\delta,0} &= \{s : d(s, \mathcal{B}) < \delta, T(s) = 0\}, \\ R_{\delta,1} &= \{s : d(s, \mathcal{B}) < \delta, T(s) = 1\}. \end{aligned}$$

Intuitively, $R_{\delta,0}$ is the region within distance δ of the boundary, \mathcal{B} , on the side of precinct 0, with an analogous interpretation for $R_{\delta,1}$ and precinct 1. Then, denote R_δ to simply be the combined area of $R_{\delta,0}$ and $R_{\delta,1}$ (i.e. the total region within distance δ of boundary \mathcal{B}). Figure 6 provides an illustration of these terms.

We frame our problem and the regression discontinuity design within the potential outcomes framework (Rubin, 1974). We define \mathcal{S}^1 to be the set of locations with an arrest had every area been exposed to policing by police precinct 1 and \mathcal{S}^0 be the corresponding quantity for precinct 0. Accordingly, we can define $Y^1(R_\delta)$ to be the number of outcomes we would observe in region R_δ if exposed to policing by precinct 1 and $Y^0(R_\delta)$ be the same quantity for precinct 0. We assume that these potential outcome point patterns come from inhomogeneous point processes with intensity functions $\lambda^1(s)$ and $\lambda^0(s)$, respectively. Precincts 0 and 1 could have different observed arrest rates for a number of reasons, many of which are *not* due to a causal effect of police precincts. One area could have higher crime rates, different types of crimes committed with different clearance rates, or a different demographic of individuals in the population that live there. However, since precinct boundaries are defined on streets and many individuals are unaware of which precinct they reside in, we can focus our analyses on regions close to the boundary between the two precincts (i.e. R_δ for sufficiently small δ) because the individuals on either side of the street are more

likely to be similar and neighbourhood characteristics should be more comparable. In particular, we examine a local average treatment effect defined by

$$\theta(R_\delta) = E(Y^1(R_\delta) - Y^0(R_\delta)) = \int_{R_\delta} \{\lambda^1(s) - \lambda^0(s)\} ds. \quad (2)$$

This treatment effect is local in the sense that it is the effect of being exposed to policing practices of precinct 1 versus precinct 0, but only in the region near the boundary, defined by R_δ . This is standard in the regression discontinuity literature where treatment effects are typically identified at, or near, the cut-off of the forcing variable.

Alternative estimands that provide more detailed information about the nature of the treatment effect are also of interest. One such estimand, which acknowledges that the treatment effect may vary spatially across the boundary of interest, can be defined as

$$\tau(b) = \lambda^1(b) - \lambda^0(b) \quad \forall b \in \mathcal{B}. \quad (3)$$

This represents the difference in the underlying point process intensity surfaces at all locations on the boundary of interest, and is an extension of estimands seen in prior spatial regression discontinuity designs (L. J. Keele & Titiunik, 2015) to the point process setting. This estimand allows for heterogeneity of the treatment effect across the boundary of interest. We can also aggregate this effect across the boundary using similar ideas as in L. J. Keele and Titiunik (2015) and Rischard et al. (2021) by defining the following:

$$\tau = \frac{\int_{s \in \mathcal{B}} w(s) \tau(s) ds}{\int_{s \in \mathcal{B}} w(s) ds}, \quad (4)$$

where $w(s)$ is a weight function that assigns weight to each point on the boundary. Throughout, we assign equal weights, $w(b) = 1$ for all b , but the ideas used hold for any choice of weights. Other weights such as those that assign weight proportional to population size, or weights that minimize the variance of the estimated treatment effect, may also be of interest, though we refer readers to Rischard et al. (2021) for a broader discussion around this choice. Similar to (2) this provides a local average treatment effect at the boundary, but we will see in subsequent sections that the identification assumptions and estimation strategies are slightly different between the two estimands. In general, we recommend using the point process estimand in equation (3) if one is interested in spatial heterogeneity of the effect across the border, and the average estimand in (4) otherwise. In subsequent sections, we will see reasons for choosing these estimands over $\theta(R_\delta)$. In particular, the identification assumptions are arguably more plausible and easier to justify (see Section 3.2), and the selection of tuning parameters is more straightforward for estimating these estimands (see Section 3.5). Despite this, we still present results and discussion around the estimand in (2) as it is easier to study and provide intuition for mathematically, in addition to being computationally less demanding to estimate as it does not require estimating intensity surfaces or implementing cross-validation.

3.2 Identifying assumptions

The main idea behind the regression discontinuity design is that by looking in a close window around the boundary, the two regions on either side of the boundary are very similar with respect to all important features except for which precinct's police department they are being policed by. The impact of important confounding variables associated with arrest rates that may differ between the two precincts should be mitigated when looking within small geographic areas, as long as the confounding factors are continuous at the boundary between the two precincts. We can therefore compare outcomes on either side of the boundary and attribute differences to the effect of the police departments in each precinct. Here, we formalize this notion by explicitly writing down the assumptions by which the regression discontinuity design is able to identify the aforementioned local average treatment effects from the observed data. First, for ease of exposition, we show these assumptions in the absence of additional covariate information. We discuss the role of

covariates in detail in Section 3.6, and weaken the following identification assumptions to incorporate covariates in [online supplementary material, Section 2](#). Additionally, in what follows, we aim to weaken these assumptions and provide valid hypothesis tests even when some of these assumptions are violated.

The two most commonly used assumptions unique to the regression discontinuity design are local randomization assumptions or assumptions on continuity of potential outcomes at the cut-off of the running variable. Local randomization states that treatment assignment is independent of the potential outcomes when looking only within a small window around the cut-off of the running variable ([Mattei & Mealli, 2017](#)). Continuity of potential outcomes is the most commonly used assumption in the regression discontinuity literature and specifies that the conditional mean of the potential outcomes under both treatment and control are continuous functions in the running variable ([G. W. Imbens & Lemieux, 2008](#)). Here, we extend these assumptions to the spatial point process setting and show how they can be used to identify $\theta(R_\delta)$ and $\tau(\mathbf{b})$, respectively. First, we detail the assumptions needed to identify $\theta(R_\delta)$, which are given in Assumptions 1 and 2.

Assumption 1 Consistency of potential outcomes.

$$Y^t(R_{\delta,t}) = Y(R_{\delta,t}) \quad \text{for } t = 0, 1.$$

Assumption 2a Constant integrated intensity functions.

$$E(Y^t(R_{\delta,1})) = E(Y^t(R_{\delta,0})) \quad \text{for } t = 0, 1.$$

Assumption 1 is a standard assumption required to link our observed data to the potential outcomes, and ensures that there exists only one version of treatment and that the potential outcomes for one region do not depend on treatment values of other regions (no interference). We believe this assumption is plausible in our study as we do not expect arrest rates in one region to depend on the police department practices of other regions. The second assumption is arguably stronger and states that for both the control and treated potential outcome point processes, the expected number of events that fall on one side of the boundary within a distance of δ is the same regardless of which side of the boundary is being looked at. This is a point process extension of local randomization assumptions used previously, and is needed because we need to use what happened under the precinct 0 side of the border to infer what would happen on the precinct 1 side of the border had they both been exposed to policing by the police department in precinct 0. As shown in [online supplementary material, Section 1](#), under these assumptions we can identify the effect of interest as

$$\theta(R_\delta) = 2E[Y(R_{\delta,1}) - Y(R_{\delta,0})],$$

which is a fully observable quantity. Also of interest is $\tau(\mathbf{b})$, which represents the local treatment effect at location $\mathbf{b} \in \mathcal{B}$. To identify this effect, we need a point process extension of the conditional continuity assumptions used in regression discontinuity designs, which is given in Assumption 2b.

Assumption 2b Continuity of potential outcome intensity surfaces.

$$\lim_{s \rightarrow \mathbf{b}} \lambda^t(s) = \lambda^t(\mathbf{b}) \quad \text{for } t = 0, 1.$$

This assumption guarantees that if we see a discontinuity at \mathbf{b} in the observed intensity surfaces, then it can be attributed to an effect of the treatment, which also has a discontinuity there. [Online supplementary material, Section 1](#) shows that $\tau(\mathbf{b})$ is identified under Assumptions 1 and 2b as

$$\tau(\mathbf{b}) = \lambda^1(\mathbf{b}) - \lambda^0(\mathbf{b}) = \lim_{s \rightarrow \mathbf{b}^1} \lambda(s) - \lim_{s \rightarrow \mathbf{b}^0} \lambda(s).$$

Note here that we use the notation $\lim_{s \rightarrow \mathbf{b}^1}$ to denote a limit that approaches the boundary location \mathbf{b} from the precinct 1 side of the boundary, with analogous notation for precinct 0.

Assumptions 2a or 2b, depending on which estimand is being studied, is arguably the most crucial assumption needed for identification and the most likely one to not hold in practice. It could fail, for instance, if people decided where to live based on the geographic boundary, and therefore the two sides of the border would not be comparable with respect to important characteristics that drive arrest rates, such as socioeconomic or criminological variables. Due to this, in Section 3.4, we detail a procedure to test the null hypothesis of no treatment effect that is robust to certain violations of these assumptions. Additionally, in Section 3.6 and [online supplementary material, Section 2](#), we detail how additional covariate adjustment can be incorporated to weaken these assumptions.

3.3 Estimation of treatment effects

Before detailing our strategy for hypothesis testing that is robust to certain violations of Assumptions 2a or 2b, we must first discuss estimation strategies for the spatial regression discontinuity design with point process data. Estimation of $\theta(R_\delta)$ is straightforward for a given choice of δ , as a natural estimator is simply a plug-in estimator given by $\hat{\theta}(R_\delta) = 2(Y(R_{\delta,1}) - Y(R_{\delta,0}))$, where $Y(R_{\delta,t})$ is the number of events on the precinct t side of the boundary. In order to estimate $\tau(\mathbf{b}) = \lim_{s \rightarrow b^1} \lambda(s) - \lim_{s \rightarrow b^0} \lambda(s)$, we fit two separate models, one for each of the two limits of interest. For $\lim_{s \rightarrow b^1} \lambda(s)$, we utilize only the data on the precinct 1 side of the boundary, and fit a model to estimate the intensity surface of the point process on that side of the boundary. We can then extrapolate this intensity surface to estimate the intensity surface at \mathbf{b} . Intensity surface estimation for this analysis utilizes the methodology implemented in the R package `spatstat` ([Baddeley & Turner, 2005](#); [Baddeley et al., 2015](#)). Specifically, let W be a two-dimensional spatial window with which we define an intensity surface over. Further, let l_j represent a distinct spatial location of an observation within W for $j \in \{1, \dots, J\}$, and m_j represents the weight associated with l_j (e.g. if there exists four observations located at l_j , then $m_j = 4$). Then, at any given point $s \in W$, the intensity surface value, $\lambda(s)$, is estimated using a fixed-bandwidth kernel estimate given by

$$\hat{\lambda}(s) = \frac{\sum_{j=1}^J \kappa(l_j - s) \cdot m_j}{e(s)},$$

where $e(s) = \int_W \kappa(\mathbf{v} - s) d\mathbf{v}$ and $\kappa(\cdot)$ is the kernel of an isotropic Gaussian density function. The variance (i.e. bandwidth) for this density, denoted by σ^2 , is what defines the spatial smoothness of the intensity surface. The choice of σ^2 is one that empirically minimizes the mean squared error (MSE) of the estimator for the local intensity surface about s . The MSE for the estimator is derived from [Diggle \(1985\)](#) and [Berman and Diggle \(1989\)](#), where it is assumed that the point process is a stationary, isotropic Cox process, allowing the MSE to be written as a function of only the smoothness parameter σ^2 . Given their importance for all of the estimands considered, we discuss smoothness parameters such as σ^2 in more detail in Section 3.5.

3.4 Resampling to obtain robust test of null treatment effect

Of major concern when using the regression discontinuity design is that the aforementioned assumptions do not hold. While Assumption 1 is reasonable in many applications, Assumptions 2a and 2b are relatively strong and can fail in certain scenarios. For simplicity, in this section, we focus attention on $\theta(R_\delta)$ and Assumption 2a, but identical ideas hold for $\tau(\mathbf{b})$ and Assumption 2b. Violations of Assumption 2a are problematic as they can lead to bias in the estimated treatment effects and inflation of type I error rates, which could potentially be leading to the results seen in [Figure 4](#). For NYC, in particular, this assumption would be violated if the communities on either side of the boundary are systematically different with respect to unmeasured variables that affect the potential outcome distributions. While we can reduce this by forcing δ to be as small as possible, neighbourhoods in NYC can change drastically over short geographic distances. In other applications of GeoRDDs, units may choose to live on one side of the boundary due to the boundary itself, which can also violate this assumption. Our interest will be in testing the null

hypothesis $H_0 : \theta(R_\delta) = 0$, and our goal will be to create a hypothesis test that has valid type I error, even in the presence of certain violations of Assumption 2a.

We develop a two-step procedure to test this null hypothesis. The first step is to define a test statistic for this hypothesis, while the second step involves resampling new boundaries in NYC to estimate the distribution of this statistic under the null hypothesis of no precinct effect. For step one, we define

$$Z = |Y(R_{\delta,1}) - Y(R_{\delta,0})|, \quad (5)$$

which is the difference in the number of events between regions $R_{\delta,1}$ and $R_{\delta,0}$. Clearly, this test statistic will be large when there are differences in policing practices by police departments in precincts 1 and 0. One difficulty we must overcome is that this test statistic does not have a known distribution that can be used for inference. A larger problem, however, is what happens when Assumption 2a does not hold. Even if the distribution of this test statistic is known under the null hypothesis, violations of Assumption 2a will lead to larger values of Z , and we must account for this to obtain valid inference.

Our goal is to estimate the null distribution of our test statistic, and we refer to the cumulative distribution function (CDF) of this distribution by F_0 . To estimate this null distribution, we can sample new precinct boundaries that behave similarly to the original precinct boundary of interest. The key difference is that these new boundaries, which we call *null streets*, are fully contained within a single precinct and therefore have no precinct effect, i.e. $\theta(R_\delta) = 0$ by design. Fortunately, we have a very rich data set that includes information on all of NYC, not just at the boundaries of the precincts, and we can leverage this data set to find a large number of null streets. An illustration of this for one precinct can be found in Figure 7, and a map showing streets across all of NYC can be found in [online supplementary material, Section 7](#).

Assuming we can find a large number, B , of streets that are not near precinct boundaries, we can estimate the test statistic at each null street and use the distribution of these statistics as an estimate of F_0 . Note that our procedure will be valid for any test statistic, though we will proceed with Z from (5). We denote these test statistics by Z^b for $b = 1, \dots, B$. We can then estimate the null distribution via $\hat{F}_0(a) = \frac{1}{B} \sum_{b=1}^B \mathbf{1}(Z^b < a)$, which allows us to perform hypothesis testing. The intuition behind using this test to provide more robust hypothesis testing is that if Assumption 2a does not hold at the boundary of interest, then it likely does not hold in other areas of NYC as

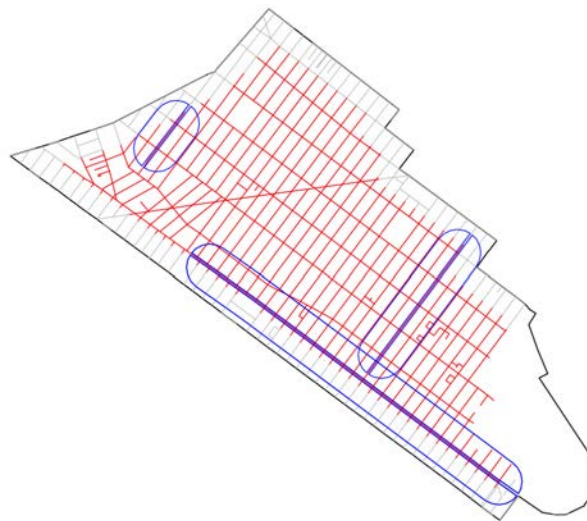


Figure 7. The red streets in Precinct 83 are ones that could potentially be used as null streets. The buffers (blue) are drawn around three potential null streets to illustrate how they meet the qualification for being completely contained in one precinct.

well. For instance, there could be substantial spatial variability in individuals across NYC that changes far more locally than distances of δ . We would not be able to account for this with observed covariates that are available only at the census-tract level, which is not sufficiently spatially resolved. However, it is likely that this variability is not unique to precinct boundaries, and that this variation also affects estimates at our resampled locations as well. By using these resampled locations, we are no longer relying on Assumption 2a holding to obtain a valid hypothesis test, but rather a modified assumption at the resampled locations. To provide intuition for this, let us first simplify the test statistic to be $Z = Y(R_{\delta,1}) - Y(R_{\delta,0})$. Under Assumption 1, the mean of this difference can be written as

$$\begin{aligned} E(Z) &= E[Y(R_{\delta,1}) - Y(R_{\delta,0})] \\ &= \theta(R_\delta) + E[Y^0(R_{\delta,1}) - Y^1(R_{\delta,0})]. \end{aligned}$$

If we further adopt Assumption 2a, we have that $E(Z) = \theta(R_\delta)/2$ and it is therefore zero under the null hypothesis of no precinct effect. If, however, Assumption 2a is violated in the sense that $E[Y^1(R_{\delta,1}) - Y^1(R_{\delta,0})] = E[Y^0(R_{\delta,1}) - Y^0(R_{\delta,0})] = \mu$, we have that

$$E(Z) = \theta(R_\delta)/2 + \mu.$$

This shows that if there are structural differences on the two sides of the boundary with respect to the potential outcomes, this will inflate our test statistic. In particular, under the null hypothesis of no treatment effect, this expected difference becomes μ instead of zero, which can lead to inflated type I error if this is not taken into account. To describe when our procedure can lead to valid inference, we first define

$$U^t(\delta, \mathcal{B}) = Y^t(R_{\delta,1}) - Y^t(R_{\delta,0}) \quad \text{for } t = 0, 1,$$

where again \mathcal{B} is used to denote the fact that these are at the boundary of interest. We also let $U^t(\delta, \mathcal{B}) \sim G$ for some distribution G under the null hypothesis of no treatment effect. If we let \mathcal{B}^* represent a null street boundary, then we make the following assumption:

Assumption 3 The null streets' test statistics match those of the precinct boundary locations under the null hypothesis, in the sense that $U^t(\delta, \mathcal{B}^*) \sim G$ for $t = 0, 1$.

If this assumption holds and the null streets have similar levels of violations of Assumption 2a, then the test statistics found at the null streets should approximate the true null distribution F_0 and we can construct rejection regions for our test using the relevant quantiles of the estimated null distribution. Assumption 2a makes the restrictive assumption that $E(U^t(\delta, \mathcal{B})) = 0$, while Assumption 3 allows for violations in the sense that $E(U^t(\delta, \mathcal{B})) = \mu$ as long as we can find null streets with similar violations. We see in Section 3.7 that this is actually a stronger assumption than what is required for valid inference. If the null streets have larger violations of Assumption 2a than at the precinct boundaries, then we should obtain valid, albeit conservative, inference. Note that while we argued for this procedure for testing $\theta(R_\delta) = 0$, the same ideas hold for testing whether $\tau(\mathbf{b}) = 0$. We would simply need to change our test statistic to be $Z = |\lim_{s \rightarrow b^1} \hat{\lambda}(\mathbf{b}) - \lim_{s \rightarrow b^0} \hat{\lambda}(\mathbf{b})|$, and all other ideas remain unchanged.

3.5 Choosing smoothing parameters

Note that for both estimands $\theta(R_\delta)$ and $\tau(\mathbf{b})$, there exist a parameter that governs how localized estimation is over space. For $\theta(R_\delta)$, the distance δ dictates a bias-variance trade-off for our estimation strategy because decreasing δ makes Assumption 2a more plausible and reduces bias in estimation of the causal effect, however, it reduces the amount of data we have to estimate the treatment effect, thereby increasing variability. For $\tau(\mathbf{b})$, the spatial smoothing parameter σ^2 dictates a similar bias-variance trade-off, where a smoother intensity surface, $\lambda(\mathbf{b})$, will have lower variability, but may induce bias by using information too far from the boundary of interest.

Related issues arise in standard regression discontinuity designs or GeoRDDs that utilize local linear regression for estimation, which contains a crucial bandwidth parameter. Selection of the bandwidth parameter has seen significant attention in the regression discontinuity design literature with a focus on finding optimal (in terms of MSE) bandwidth parameters (G. Imbens & Kalyanaraman, 2012) or performing inference in a way that accounts for bias in treatment effect estimation (Calonico et al., 2014a, 2014b). One key issue in this literature is that constructing valid confidence intervals when using cross-validation or MSE-optimal choices of bandwidth parameters is difficult because of the asymptotic bias in these estimators caused by oversmoothing (L. J. Keele & Titiunik, 2015). One can attempt to undersmooth by choosing a smaller bandwidth than what is chosen by cross-validation, but the degree of undersmoothing is generally not known. Bias-corrected confidence intervals have been developed in standard regression discontinuity designs, but this theory has not been developed for the point process setting seen here. Fortunately, our resampling procedure described in Section 3.4 helps resolve some of these issues. For estimation of $\tau(b)$, we recommend choosing the smoothing parameter σ^2 that minimizes the MSE of the intensity surface estimator. This should lead to good estimates of the intensity surfaces that balance competing interests of using enough data, while also focusing in a close window around the boundary. In our case, the MSE-optimal σ^2 is determined by a numerical minimization of the MSE estimate defined in Diggle (1985) and Berman and Diggle (1989). Although traditional cross-validation through resampling (e.g. leave-one-out cross-validation) could be used as another model-selection strategy, we find that for the size of our data it is computationally prohibitive. Additionally, both the spatial dependencies in the data as well as the fact that our data consist of only *one* realization of the point process, makes the independence assumption for cross-validation unreasonable. Hence, an area to explore in the future is applying novel cross-validation strategies that account for spatial dependencies, such as the work of Cronie and Van Lieshout (2018) and Cronie et al. (2024), to improve the choice of σ^2 thus improving the overall intensity surface estimate. In any case, while the possibility of undersmoothing or oversmoothing exists, it is expected that these issues occur at both the boundaries of interest and the null streets in a similar manner, thus leading to valid hypothesis tests.

Finding an optimal choice for δ to be used when studying $\theta(R_\delta)$ is less clear, however, because there is no natural way to perform cross-validation with respect to δ . Nonetheless, we can provide general guidance for the choice of this parameter. A general rule of thumb is to set δ to be a small value, which has two benefits in our framework. For one, it makes Assumption 2a more plausible than larger values of δ . It also has an advantage with respect to the resampling procedure described in the previous section, which is that smaller values of δ will have more available null streets to select from, which can lead to null streets being more similar to the precinct boundaries of interest. This is because null streets, and their corresponding buffer regions, must be fully contained within a single precinct, but this becomes less likely as the size of the buffer region grows. One approach we recommend to determine what a ‘small value’ for δ means in any application is to first consider the spatial surface over which the point process lies. In our case, we know all observations fall along street lines. Hence, we can look at the distances between neighbouring streets (as quantified by the distance between street midpoints) and compute the minimum distance between each street and its neighbours within each precinct. We can then set δ to be the 95%-quantile of the distribution of the minimum distance from streets to their nearest neighbouring street. In our application in NYC, the 95%-quantile is 273.25 feet, and therefore we recommend $\delta = 300$ as it is greater than most minimum distances between neighbouring streets, while still small enough to make Assumption 2a more plausible. Another way in which data can be used to select δ is if additional information on a negative control variable is available, such as in our study of tree locations in NYC. One can apply the resampling procedure for all possible values of δ and choose the value of δ that leads to the desired type I error rate.

In general, however, we recommend finding a small value of δ as described above and then performing inference for a range of small to moderate values of δ (we consider $\delta \in \{300, 400, \dots, 1,000\}$). This assesses whether results are consistent across these values, which would increase belief in the overall findings. Additionally, our testing procedure should be fairly robust to the choice of δ (or σ^2). In the NYC policing analysis in Section 5, we apply our procedure for a range of smoothness parameters and find relatively consistent results across all values explored.

3.6 The role of covariates

As with nearly any observational study aiming to study causal effects, we must discuss the different ways in which covariates are accounted for. This is particularly important in the present setting, as there are multiple manners in which covariates can be included in our analysis, and it is important to distinguish among these. There are two distinct places that covariates can be incorporated: (1) the identification assumptions and corresponding estimation strategy described in Sections 3.2 and 3.3, and (2) utilizing covariates to find the best null streets for estimating the null distribution in Section 3.4. While commonalities exist across these two aspects of our proposed procedure, key distinctions remain which are worth spelling out.

3.6.1 Effect on identification assumptions

Before discussing how covariates can be explicitly incorporated into the proposed procedure, we must also emphasize that regression discontinuity designs are useful, and so widely used, because they implicitly adjust for important confounding variables by design. In the context of spatial regression discontinuity designs, if important confounding variables are expected to vary smoothly across space, then the GeoRDD eliminates issues stemming from these variables by estimating treatment effects at the boundary. If the confounding factors are continuous at the boundary of interest, then the potential intensity surface will be continuous as well, and the GeoRDD can identify causal effects even without explicit adjustment for these variables. This logic has led to the regression discontinuity design being used in a variety of settings without the additional adjustment of covariates. This can be violated, however, in certain settings, such as at county or state lines, where important variables might change drastically at the boundary, as different counties have better schools, childcare options, or other factors influencing who ultimately decides to live there. In the current context of police precinct boundaries, this is expected to be less of an issue as these boundaries do not typically coincide with other important government boundaries that influence the type of people living in each area. Notably, a recent sociological study titled *Upsold* investigates ‘consumers’ preferences and decision-making in the context of purchasing homes’ (Besbris, 2020), and among the NYC home-buyers, police precinct was *not* a factor considered. Similar conclusions were found in a separate, recent book titled *Race Brokers* (Korver-Glenn, 2021). It is still possible, however, that police precincts could align with other sub-municipal boundaries that do affect where people live. We study this in [online supplementary material, Section 9](#) where we find that police precincts do not generally align with other important boundaries in NYC, which helps justify the assumption that precincts do not typically influence people’s residential choices in NYC. Despite our justification above, we acknowledge that this is still an unverifiable assumption, which could affect the validity of our results if violated.

If there are still concerns about differences in the distribution of important confounding factors on the two sides of the boundary of interest, then additional covariate adjustment can be incorporated to remove these differences, which increases the plausibility of the GeoRDD. In this setting, when referring to covariates, we are referring to spatial covariates that describe features of the geographic areas examined. For instance, one might be interested in adjusting for socioeconomic status if it is thought that socioeconomic status differs drastically on one side of the boundary compared with the other. The identifying assumptions for the point process GeoRDD can be relaxed to hold conditionally on observed covariates. We describe these extensions, the identification of causal effects incorporating covariates, and the corresponding estimators in detail in [online supplementary material, Section 2](#). We do not implement this explicit covariate adjustment in our NYC policing analysis in Section 5, because the only covariates available to us are United States Census variables, which are constant within Census blocks, and therefore are not spatially varying enough to assist in our analysis.

3.6.2 Finding null streets

The second way in which covariates influence our testing strategy, which we do implement in Sections 4 and 5, is in the selection of null streets. Assumption 3 states that the distribution of test statistics at the null streets and the precinct boundaries of interest should be equal under the null hypothesis of no police precinct effect. The distribution of the test statistics can depend on a number of covariates, however, and we need to incorporate these into the selection of which

Table 1. The choice of covariate used to find null streets for each application

Data application	Covariate/feature
Simulation	Buffer region area
Negative control	Street length in buffer region
Arrest data	Crime locations in buffer region

Note. These covariates are numeric summaries and are distinct for *each side* of a given boundary.

streets to use when estimating the null distribution for our test statistics. These covariates may be potential confounding factors, such as socioeconomic status, but they need not be. For instance, the size of the null street is a potentially important factor to use when selecting null streets as small streets will have far fewer data points than large streets, and will subsequently have more variability in their corresponding test statistics. Given the importance of null streets, we now detail the steps taken to find adequate null streets to be used in the estimation of the null distribution for a given precinct boundary's test statistic.

The first step in determining which null streets are 'similar' to the boundary of interest is to choose features in the data that potentially have an affect on the distribution of the test statistic being computed. This step is largely application-dependent and Table 1 shows a list of the different covariates used to find null streets for each of the three analyses we perform. Note that for a given boundary, there exist two values for each covariate; one on each side of the boundary. As a general guideline, the choice of covariate is based on both expert knowledge as well as specific features of the data application. For the negative control analysis in Section 5.2, we know the data are observed only on streets, so including relevant street information as a covariate is justifiable. For the simulation in Section 4, we randomly place data across the entire geographic domain, and the data are not restricted to street locations. Therefore, the total area/size of the buffer region is now a more relevant covariate to include. Finally, in the NYC policing analysis in Section 5, we use the amount of crime as a covariate because it is expected that areas with higher amounts of crime will have more arrests, irrespective of which police precinct governs that area. While each covariate we included in order to find null streets has a practical justification based on subject matter expertise, [online supplementary material, Section 4](#) presents empirical justifications for their use by showing the association of these covariates with the magnitude and variability of test statistics under the null hypothesis. Additionally, while we use a single numeric feature to find null streets in each analysis, this can be extended to using multiple covariates, if it is expected that these additional features impact the distribution of the test statistic.

Now, for a given choice of covariate, we can formally describe how it is used to find adequate null streets. Let \mathcal{M}_i denote the set of null streets for precinct boundary \mathcal{B}_i , $i \in \{1, 2, \dots, 144\}$. Further, let $t^{\text{sum}}(\cdot)$ denote the sum of the covariate values from both sides of the boundary (e.g. the total amount of crime), and let $t^{\text{ratio}}(\cdot) \geq 1$ denote the ratio of the covariate values from each side of the boundary (e.g. the ratio of the amount of crime on the two sides of the boundary). Then, a street m is considered a null street for \mathcal{B}_i (i.e. $m \in \mathcal{M}_i$) if both $c \cdot t^{\text{sum}}(\mathcal{B}_i) < t^{\text{sum}}(m) < (1/c) \cdot t^{\text{sum}}(\mathcal{B}_i)$ and $c \cdot t^{\text{ratio}}(\mathcal{B}_i) < t^{\text{ratio}}(m) < (1/c) \cdot t^{\text{ratio}}(\mathcal{B}_i)$. Here, $c \in (0, 1)$ measures the degree of similarity that a street must have to the boundary of interest to be chosen as a null street. Values of c near 1 ensure null streets are very similar to boundary \mathcal{B}_i with respect to the variable being considered, though potentially at a cost of reducing the number of available null streets. Alternatively, a value of c close to 0 means null streets are less similar to the boundary \mathcal{B}_i , but there will exist a large number of available null streets. Section 5.1 offers further commentary on choosing the value of c to use in our motivating application. Once a value of c is chosen, the null distribution for the test statistic at \mathcal{B}_i is constructed by computing the test statistic for every $m \in \mathcal{M}_i$. Lastly, the reason for using $t^{\text{sum}}(\cdot)$ and $t^{\text{ratio}}(\cdot)$ can best be explained in the context of the real data analysis of Section 5. We choose null streets based on the total amount of crime because areas with more people and more events are likely to have lower variability in their corresponding test statistics. We also use the ratio of crimes on the two sides of the boundary, because crime likely correlates with many important confounding factors that we do not measure, and finding null

streets based on this ratio helps ensure that the null streets have similar violations of Assumptions 2a or 2b, which makes Assumption 3 more plausible. Given the importance of this choice of covariate to our resampling strategy, we study it in further detail in the following section.

3.7 Theoretical insights for resampling procedure

In this section, we examine when the resampling procedure will provide valid inference, and provide guidance for choosing the number of null streets B , which is a critical choice for both statistical validity and power of the proposed hypothesis tests. Larger values of B should lead to more efficient estimates of the unknown CDF F_0 . However, increasing B also may lead to using null streets that do not closely match the boundary of interest and Assumption 3 will not be satisfied. For the rest of this section, assume a fixed buffer width δ , though all results will hold regardless of the chosen buffer width if Assumption 3 is satisfied. Let Z_i represent the test statistic at precinct boundary i . Further, let $Z_i^{(b)}$ represent a resampled test statistic for precinct boundary i where $b \in \{1, 2, \dots, B\}$. Next, define X_i to be characteristics associated with precinct boundary i and similarly, $X_i^{(b)}$ denotes characteristics for null street b . For simplicity, we let X_i and $X_i^{(b)}$ be univariate here, but the same ideas hold for a vector of covariates. Assume under the null hypothesis that $P(Z \leq z \mid X = x) = F(x, z)$ is the CDF of the test statistic of interest, and $f(x, z)$ represents the corresponding density function. Note that we are assuming that the distribution of the test statistic depends on characteristics X . Potential characteristics in NYC are the length of the border \mathcal{B} or the ratio of crime rates on the two sides of the border. As noted in Section 3.6.2, the distribution of the test statistic in NYC indeed depends heavily on characteristics such as these. For this reason, we want to find null streets with similar values of these characteristics, i.e. $X_i^{(b)} \approx X_i$ for $b = 1, \dots, B$ so that Assumption 3 holds, though we quantify this idea more rigorously in what follows.

Our first goal of the resampling procedure is to obtain statistical validity and maintain type I error control at rate α for our hypothesis test. We can define the true $1 - \alpha$ quantile of Z_i under the null hypothesis as $Q_{1-\alpha}$. Our corresponding estimate of this quantile is given by

$$\hat{Q}_{1-\alpha} = \min \left\{ q : \hat{F}(X_i, q) \geq 1 - \alpha \right\},$$

where our estimate of the CDF is given by

$$\hat{F}(X_i, q) = \frac{1}{B} \sum_{b=1}^B \mathbf{1}(Z_i^{(b)} \leq q).$$

We use this estimate of the CDF throughout this section, though kernel smoothing can be used to improve estimation of the CDF when B is small. In [online supplementary material, Section 3](#), we show that the type I error of our procedure is given by

$$P(\text{reject } H_0 \mid H_0) = 1 - E_{\hat{Q}}[F(X_i, \hat{Q}_{1-\alpha})].$$

This result implies that we can obtain type I error control at level α if

$$E_{\hat{Q}}[F(X_i, \hat{Q}_{1-\alpha})] \geq 1 - \alpha = F(X, Q_{1-\alpha}).$$

This shows that validity does not rely on an unbiased, or even conservative, estimate of $Q_{1-\alpha}$. Rather we need our estimated quantiles $\hat{Q}_{1-\alpha}$ to be such that on average, the true CDF at our estimated quantiles is above $1 - \alpha$. Even if we have an unbiased estimator of $Q_{1-\alpha}$, if it has excessive variance, then it might not satisfy the condition above and will lead to anti-conservative inference. To gain further intuition into this, we study the properties of our estimator of the CDF, denoted by

$\widehat{F}(X_i, q)$. First, we can look at the mean of this estimator, which we show in [online supplementary material, Section 3](#) can be approximated as follows:

$$\begin{aligned} E[\widehat{F}(X_i, q)] &\approx F(X_i, q) + \frac{d}{dX_i} F(X_i, q) \cdot E(X_i^{(b)} - X_i) \\ &\quad + \frac{d^2}{dX_i^2} F(X_i, q) \cdot E[(X_i^{(b)} - X_i)^2]. \end{aligned}$$

This shows that the bias of the estimator is a function of how close the null street covariates $X_i^{(b)}$ are to X_i . Therefore, finding null streets that have similar characteristics as the precinct boundary of interest is of crucial importance. Of course, if X_i does not affect the distribution of the test statistic, then $\frac{d}{dX_i} F(X_i, q) = \frac{d^2}{dX_i^2} F(X_i, q) = 0$ and we have no bias regardless of how similar the null streets are. Also of importance is the variance of this estimator, which we show in [online supplementary material, Section 3](#) is approximated by

$$\begin{aligned} \text{Var}[\widehat{F}(X_i, q)] &\approx \frac{1}{B} \left\{ E[F(X_i^{(b)}, q) \cdot (1 - F(X_i^{(b)}, q))] \right. \\ &\quad \left. + \left(\frac{d}{dX_i} F(X_i, q) \right)^2 \cdot \text{Var}(X_i^{(b)}) \right\}. \end{aligned}$$

One would expect that the variance generally decreases as we increase the number of null streets, B . However, this shows an important feature of the resampling procedure, which is that the variance need not necessarily go down as we increase B . This is because increasing B can also increase $\text{Var}(X_i^{(b)})$ by including null streets with very different values of $X_i^{(b)}$, which leads to an increase in the overall variance. These two results show that there is a trade-off involved when choosing B . On the one hand, we want to increase B to decrease variability in $\widehat{F}(X_i, q)$. On the other hand, we want to keep B small enough so that the null streets are closely aligned with the precinct boundary of interest in the sense that $X_i^{(b)} \approx X_i$, which reduces bias, and potentially reduces the variance of the estimate of the CDF. We discuss this trade-off in the context of the NYC data in Section 5.1.

3.8 Global test of variation by precinct

So far, we have focused on performing a hypothesis test at a single border (namely between precincts 0 and 1), but there exists many such precinct borders in NYC. While there is interest in knowing whether any two bordering precincts have differential arresting practices, also of interest is whether there is any variation across all NYC police precincts. In this setting, we might wish to test whether the arrest rates differ by any precinct in NYC. It is difficult to compare any two precincts that are not bordering each other as we would not be able to focus on the border between these two precincts and therefore cannot utilize the GeoRDD. For this reason, we restrict our attention to assessing whether any bordering precincts have differential policing practices. Focusing first on the local average treatment effect within δ of the boundary, let $\theta(R_\delta^{(i)}) = E(Y^1(R_\delta^{(i)}) - Y^0(R_\delta^{(i)}))$ be the local treatment effect at boundary i , thus leading to the following hypothesis:

$$\begin{aligned} H_0: \theta(R_\delta^{(i)}) &= 0 \quad \text{for } i = 1, \dots, M, \\ H_a: \theta(R_\delta^{(i)}) &\neq 0 \quad \text{for at least one } i. \end{aligned}$$

Note that $R_\delta^{(i)}$ is defined precisely as R_δ is defined in Section 3.1 except $i = 1, 2, \dots, M$ specifies the exact boundary of interest. To perform this hypothesis test, we use a test statistic given by

$$\bar{Z} = \frac{1}{M} \sum_{i=1}^M Z_i = \frac{1}{M} \sum_{i=1}^M |Y(R_{\delta,1}^{(i)}) - Y(R_{\delta,0}^{(i)})|.$$

Larger values of this test statistic provide additional evidence against the null hypothesis. We can use the same resampling procedure described in Sections 3.4 and 3.7 in order to perform inference using this test statistic. We approximate the distribution of \bar{Z} under the null hypothesis using the empirical distribution of \bar{Z}^b for $b = 1, \dots, B$. Note that while we focused on the local average treatment effect within δ of the precinct boundaries, the same ideas would apply for τ defined in (4). Specifically, we could use the resampling procedure to test

$$\begin{aligned} H_0 : \tau_i &= 0 & \text{for } i = 1, \dots, M, \\ H_a : \tau_i &\neq 0 & \text{for at least one } i. \end{aligned}$$

where τ_i is the estimand defined in (4) applied to precinct boundary i . Similarly, the test statistics Z_i would be updated to be estimators of τ_i , which are defined in Section 3.3. Note that while the test statistic is an average across all precinct boundaries, another justifiable test statistic would be to use $\max_i Z_i$, which is analogous to using a minimum p -value over all hypothesis tests (Tippett, 1931). This statistic may have more power if only a small subset of the precinct boundaries have an effect of police precincts.

4 Simulation study

Here, we assess the performance of the proposed approach to testing in the GeoRDD using simulated outcome data across NYC. We generate data from four scenarios to evaluate performance in a wide range of plausible settings. In each scenario, we generate 1,000 data sets. In each data set, we first generate the intensity surface $\lambda(\cdot)$ of the point process across the surface of NYC. Counts of outcomes within any particular region R , such as the area around a precinct boundary, are then drawn from a Poisson distribution with mean given by $\Lambda(R) = \int_R \lambda(s) ds$. Similar to the negative control analysis, simply doing a binomial test to compare the number of simulated counts on either side of a boundary leads to invalid results and inflated type I error rates. With this in mind, we run our proposed procedure on each of the simulated data sets and evaluate the probability that the null hypothesis of no precinct effect is rejected. For the individual precinct boundary tests, results are averaged over all 1,000 data sets and all 144 precinct boundaries in NYC. For the global test, only one test is run for each simulated data set, and results are averaged over 1,000 simulations.

4.1 Surface construction

We refer to the four different underlying intensity surfaces of the observed outcomes for the NYC landscape as (1) *Constant*, (2) *Random*, (3) *Spatial*, and (4) *Precinct Effect*. Heat maps of one realization for each of the surfaces are shown in Figure 8 where the dark red areas represent higher values of the outcome. The Constant, Random, and Spatial surfaces represent situations with no precinct effect, and therefore, the null hypothesis of no precinct effect is true. The Random and Spatial surfaces, however, represent situations where standard regression discontinuity designs might fail because there will likely be more counts on one side of the precinct boundary than the other due to randomness or spatial variation in the surfaces not driven by any precinct effect. These are intended to represent realistic situations in NYC such as crime hotspots or spatial correlation in crime levels that can lead to differential counts of outcomes in one precinct than another that is not attributable to the precincts themselves. In these situations, our goal is to maintain type I error control at level α despite these differential counts. The Precinct Effect surface, however, has clear precinct effects and we want to assess the power to detect these differences.

4.2 Type I error control and power

The results from all simulations can be found in Table 2, which shows the percentage of rejected hypothesis tests for both the individual tests and global tests, respectively. Note that these results are for hypotheses in terms of $\theta(R_\delta)$, not $\tau(b)$. Estimation of $\tau(b)$ requires finding the MSE-optimal smoothing parameter for the intensity surface estimation at all boundaries of interest, including null streets, and is therefore computationally prohibitive to run on such a large scale over 1,000 simulations. We see that for the Constant, Random, and Spatial surfaces, we are able to recover $\alpha = 0.05$ type I error rates. The Spatial surface is somewhat more challenging and leads to slightly

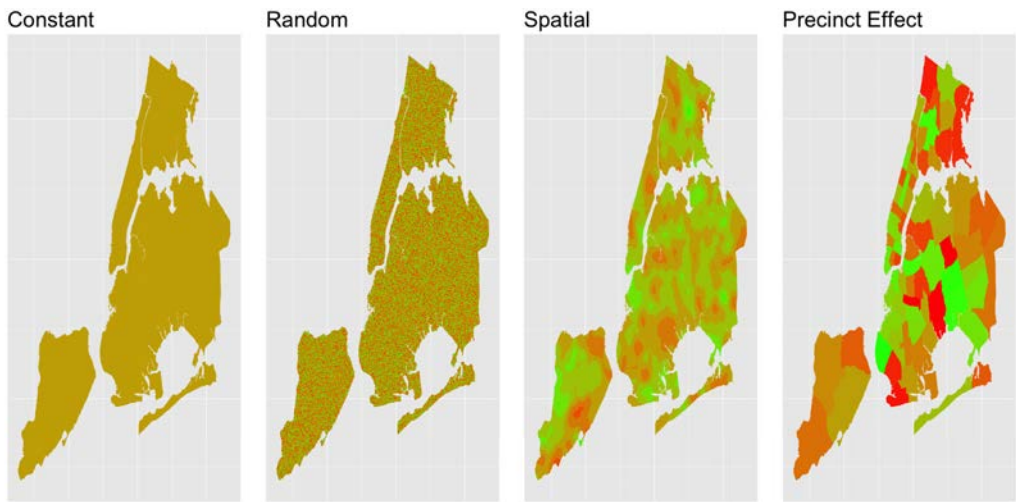


Figure 8. Heat maps representing one realization of an intensity surface for the outcomes.

Table 2. Probability of rejecting the null hypothesis across the four simulation scenarios and differing buffer widths

δ	Constant	Random	Spatial	Precinct
Individual				
300	0.049	0.054	0.051	0.920
400	0.048	0.054	0.053	0.935
500	0.048	0.051	0.054	0.942
600	0.050	0.053	0.063	0.953
700	0.051	0.058	0.074	0.963
800	0.050	0.057	0.080	0.967
900	0.051	0.057	0.082	0.970
1,000	0.052	0.059	0.089	0.973
Global ($\max_i Z_i; \bar{Z}$)				
300	0.038; 0.040	0.032; 0.041	0.017; 0.024	1.000; 1.000
400	0.048; 0.024	0.028; 0.034	0.019; 0.030	1.000; 1.000
500	0.040; 0.028	0.031; 0.022	0.034; 0.045	1.000; 1.000
600	0.048; 0.032	0.041; 0.034	0.043; 0.108	1.000; 1.000
700	0.044; 0.035	0.040; 0.056	0.061; 0.231	1.000; 1.000
800	0.045; 0.038	0.031; 0.045	0.077; 0.297	1.000; 1.000
900	0.035; 0.045	0.048; 0.049	0.091; 0.348	1.000; 1.000
1,000	0.056; 0.053	0.038; 0.054	0.111; 0.408	1.000; 1.000

Note. The top half of the table corresponds to hypothesis tests at individual precinct boundaries, while the bottom half corresponds to the global test of variation across NYC using two different test statistics.

inflated type I error rates for larger buffer widths, while maintaining type I error control at smaller buffer widths. In terms of the global test, both statistics perform relatively well, though the maximum statistic is somewhat more robust with smaller type I error rates in the Spatial surface setting. Overall, this shows that the proposed approach is indeed able to provide valid inference even in settings with differential outcome levels on either side of the boundary, i.e. when Assumption 2a is violated. In this setting, violations of Assumption 2a occur in a similar manner across the city,

and therefore Assumption 3 holds and we obtain valid inference. In the Precinct Effect scenario, the null hypothesis is not true, and we see that our approach has high power to detect these differences across precincts. The power is slightly below 1 for the individual tests and this is because some precincts have very small differences in counts from their neighbouring precincts. The global test does not suffer from this issue as it uses either the average or maximum of the test statistics over all precinct boundaries, leading to a power of 1. Additionally, in [online supplementary material, Section 8](#) we run additional simulation studies with reduced sample sizes and obtain similar results.

5 Analysis of precinct by precinct arrest rates

Here, we analyse the NYC arrest data to estimate the degree of variation in policing across the city as well as whether there are significant differences between individual precincts with regards to their arresting practices. We first discuss our strategy for finding null streets in NYC and use our procedure to test the null hypothesis of no precinct effects in the negative control data. Then, we use our procedure to test for precinct-specific effects and global variation in policing with respect to arrest rates. Given that our outcome of interest is arrest rates and not raw totals of arrests, we scale all results by the number of crimes in the corresponding area. For $\theta(R_\delta)$, this is done by dividing all arrest counts by the number of crimes in the same area, and for τ , this is done by dividing intensity surfaces for arrests by the intensity surfaces for crimes.

For null streets and the negative control analysis, if an event occurs directly on the boundary between the two sides of interest, that point is randomly assigned to one side of the boundary. For the main analysis of police precincts, we have information on the precinct of the arresting officer, and we use this to assign a precinct to observations that fall directly on the border. Additionally, throughout this section, we present results for differing spatial smoothness values. The spatial smoothness level is determined by first using the MSE-optimal estimate of σ^2 as in Section 3.5, then scaling $\hat{\sigma}^2$ by some factor ('smoothing multiplier') to vary the smoothness. Also, note that for constructing intensity surfaces, we use observations that fall within a specified radius ('region size') from the boundary of interest. We focus on a region size of 600 feet here, but ran the same analyses for a range of region sizes and found very similar results. Lastly, all results presented in this section are for hypotheses with respect to τ . To see analyses targeting $\theta(R_\delta)$, as well as additional results from the analyses for τ , see [online supplementary material, Sections 5 and 6](#).

5.1 Quantifying the similarity of null streets

As discussed in Sections 3.6 and 3.7, it is important to consider both which covariates to use when finding null streets and the number of null streets to use, as both have implications for type I error control. One feature of streets that can alter the distribution of the test statistics under the null hypothesis is the size of the street being considered. We expect larger streets with more crime to have less variability in their test statistics, while small streets with small counts on either side of the boundary to have much more variability. In other words, the null streets can be used to learn which covariates impact the test statistics and therefore should be incorporated in the final analysis. Recall from Section 3.6.2 that $c \in (0, 1)$ measures the degree of similarity that a street must have to the boundary of interest. [Figure 9](#) illustrates the importance of c as it displays the type I error for the negative control analysis of Section 5.2 as a function of c . We see that large values of c lead to type I error rates close to the desired level $\alpha = 0.05$. However, as c approaches 0 ($c = 0$ means all streets are considered null streets and covariates are not incorporated) the type I error approaches zero. Smaller values of c lead to less similarity between the null streets and the precinct boundary of interest, which in this case leads to overly conservative inference. For this reason, we proceed with $c = 0.9$ moving forward to ensure our procedure has well-calibrated type I error and there are sufficient numbers of null streets to estimate the null distribution.

5.2 Negative control analysis

[Figure 10](#) shows the per cent of significant associations out of the 144 borders using the proposed resampling approach for the negative control analysis as a function of the smoothness parameter. The percentage of rejected tests for the naive test in [Figure 5](#) is far above the desired 0.05 level as we

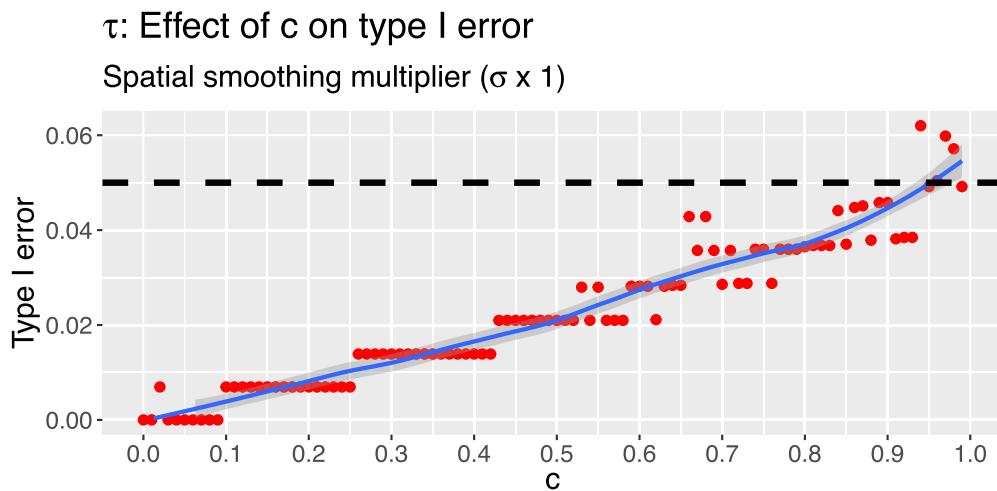


Figure 9. The type I error as a function of c , a measure of how similar the null streets are to the precinct boundaries of interest.

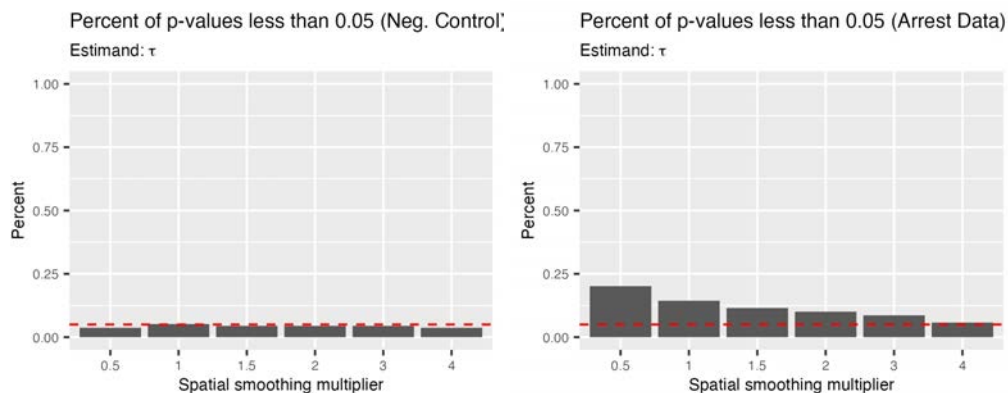


Figure 10. The percentage of p -values across the 144 boundaries that are < 0.05 using test statistics calculated from the constructed intensity surfaces and the proposed resampling procedure. The *left* panel provides the results for the negative control analysis, and the *right* panel corresponds to the arrest data analysis. We explored various smoothing parameter values with larger values corresponding to smoother intensity surfaces.

see roughly anywhere between 60% and 80% rejection rates, with an increasing trend as a function of the buffer width. Given that this outcome should not be affected by police precincts, these results point to a lack of validity of the statistical test being run or the assumptions underlying the regression discontinuity design. With our resampling approach to inference, however, the results drop to a far more reasonable level with rejection rates close to 0.05 for each smoothness level. Additionally, in [online supplementary material, Section 5](#), we show that the p -value histogram for each spatial smoothness value appears to be approximately uniformly distributed, as we would expect. Overall, the negative control analysis provides further justification for using our proposed resampling procedure, and gives increased belief in our findings on arrest rates in the following sections.

5.3 Individual boundary estimates of precinct effects

Now that we have constructed null streets for each of the 144 precinct boundaries, we can perform hypothesis tests for each boundary to assess whether there is a causal effect of police precincts near

Table 3. Results from the global test of variation in policing across NYC illustrating the p -value as a function of spatial smoothness

	Spatial smoothing multiplier					
	0.5	1.0	1.5	2.0	3.0	4.0
$\max_i Z_i$	0.340	0.670	0.662	0.838	0.878	0.891
\bar{Z}	0.345	0.429	0.271	0.309	0.569	0.734

the boundary between any two precincts. Figure 10 shows the results of these analyses by presenting the percentage of the 144 precinct boundaries for which the hypothesis test of no precinct effect was rejected using our proposed procedure. The naive hypothesis tests in Figure 4 show a large proportion of significant differences with more than 88% of the tests being rejected. Using the proposed approach, this number is far smaller. The large difference in results between the two approaches highlights that either Assumption 2b does not hold in this data set, or our statistical test is invalid. Nonetheless, the proportion of significant differences is still greater than 0.05 for most levels of spatial smoothness. For instance, at the MSE-optimal smoothing parameter (i.e. multiplier =1), 14.4% of the individual tests are rejected at the $\alpha = 0.05$ level, which is suggestive of a small amount of precinct-level differences in arrest rates. Note also that the percentage of rejections decreases as a function of the spatial smoothing multiplier, with a multiplier value of 4 leading to rejections in only 5.8% of the precinct boundaries. Smoothing parameters that are four times the MSE-optimal choice lead to extreme oversmoothing, which tends to remove effects of finer-level spatial variability. While our tests should still be valid (in terms of type I error) in this setting, this may impact the power to detect effects. Although we include these results to illustrate our approach across a wide range of scenarios, we recommend focusing on smoothing parameter values near the MSE-optimal choice of σ^2 , especially since we acknowledge that there exists some sensitivity in type I error to the smoothness level of the surface.

5.4 Global variation in arrest rates

In this section, we apply the approach of Section 3.8 to assess whether there is an overall effect of police precincts on arrest rates across NYC. For comparison, we will calculate both \bar{Z} and $\max_i Z_i$ as test statistics, defined in Section 3.8, to assess the magnitude of the overall precinct effect across the entire city. To understand the distribution of each test statistic under the global null hypothesis of no precinct effects, we also calculate each test statistic using null streets to obtain $\bar{Z}^{(b)}$ and $\max_i Z_i^{(b)}$ for $b = 1, \dots, B$, using the null streets discussed in Section 3.6.2. We perform this procedure B times for each of the distinct levels of spatial smoothness. The results of this procedure can be seen in Table 3 and Figure 11. We see that the p -value is quite large and above the $\alpha = 0.05$ cut-off for all smoothness values and for both test statistics. As in Section 5.3, the results tend to get more conservative as the smoothness parameter is increased. Figure 11 shows the estimated null distribution of each test statistic for the MSE-optimal spatial smoothing value, and we see that the observed statistic is well contained within the estimated null distribution in each case. This indicates that there is not a large degree of variation in policing practices across different precincts across NYC. While there may be differences at a small number of precincts, as indicated by the results in Section 5.3, these effects appear to be relatively small and not widespread across the city.

6 Discussion

In this manuscript, we first formalized estimands and developed estimation procedures for the GeoRDD when the data follow a point process. Additionally, we proposed an approach to hypothesis testing for GeoRDDs that weakens the local randomization or continuity assumptions that are typically made in such studies. By leveraging the rich spatio-temporal information in our data on crime and arrests in NYC, we showed that valid hypothesis tests can be constructed

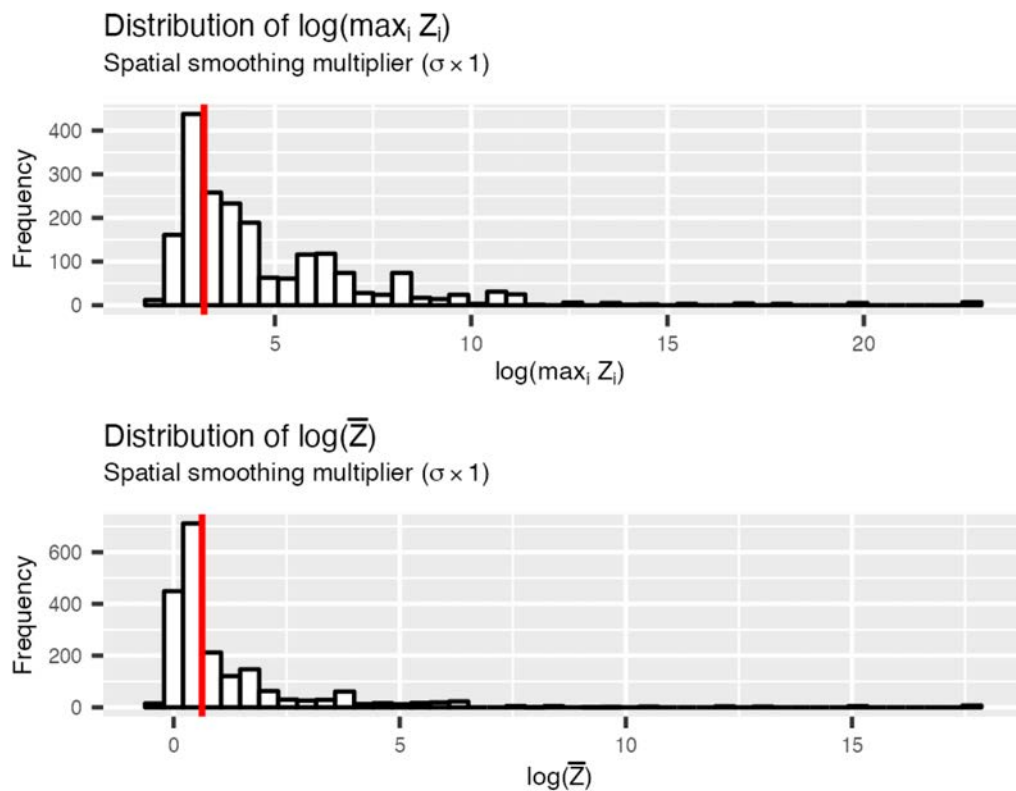


Figure 11. Histograms of resampled test statistics from our proposed procedure using either $\max_i Z_i$ (top) or \bar{Z} (bottom). The vertical red lines are the observed test statistics computed at the actual precinct borders. All results are displayed on the log scale to improve visualization in the presence of outlying values.

even in the presence of certain violations of local randomization or continuity assumptions around the boundary of interest. The main idea is to find new boundaries that behave similarly to the boundary of interest, but are not near the border of two police precincts and therefore necessarily have no precinct effect. In the analysis of NYC arrest data, we found that analyses relying on a local randomization assumption lead to very strong conclusions that police precincts greatly impact arrest rates, while our approach based on resampling new streets leads to the conclusion that there is, at most, a small effect of police precincts on arresting practices.

Our procedure was shown to work in a geographic regression discontinuity setting, though it is potentially applicable to other regression discontinuity settings as well. The only requirement is that new cut-offs of the running or forcing variable must be used where no treatment effect exists, and that the data set is rich enough to provide a large number of these new locations that are sufficiently independent of each other. While our procedure is able to provide statistical validity (type I error control) to tests of the hypothesis of no treatment effect, it is not able to correct for biases in estimation of treatment effects themselves. Further research is required to reduce the assumptions needed for estimation of treatment effects in regression discontinuity designs. Additionally, future applications of our approach to areas outside of NYC with potentially different characteristics and population densities would offer an interesting new case study as well as insight into the robustness of our approach to different contexts.

Acknowledgments

The authors would like to thank Brigid Dunn and Max Besbris for insightful conversations that led to an improved understanding of the underlying application in NYC.

Conflict of interests: None declared.

Funding

No funding support is declared.

Data availability

Most data are publicly available via *NYC Open Data*; for the data on crime and arrest rates in NYC please email the corresponding author for access.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

Author contributions statement

E.B.K. and J.A. developed the statistical methodology and analysed the results. B.B. created all data sets in New York City and served as the expert criminologist for interpreting results. E.B.K. conducted all statistical analyses and programming. E.B.K., J.A., and B.B. wrote and reviewed the manuscript.

References

- Abadie A., Diamond A., & Hainmueller J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490), 493–505. <https://doi.org/10.1198/jasa.2009.ap08746>
- Antonelli J., & Beck B. (2023). Heterogeneous causal effects of neighbourhood policing in New York city with staggered adoption of the policy. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(4), 772–787. <https://doi.org/10.1093/rjsssa/qnad058>
- Ashenfelter O. C., & Card D. (1984). Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, 67, 648–660. <https://doi.org/10.2307/1924810>
- Ba B. A., & Rivera R. (2019). The effect of police oversight on crime and allegations of misconduct: Evidence from Chicago. *U of Penn, Inst for Law & Econ Research Paper* (19–42).
- Baddeley A., Rubak E., & Turner R. (2015). *Spatial point patterns: Methodology and applications with R*. CRC Press.
- Baddeley A., & Turner R. (2005). Spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6), 1–42. <https://doi.org/10.18637/jss.v012.i06>
- Beck B., Antonelli J., & Piñeros G. (2022). Effects of New York city's neighborhood policing policy. *Police Quarterly*, 25(4), 470–496. <https://doi.org/10.1177/10986111211046991>
- Bell M. C. (2020). Anti-segregation policing. *NYUL Review*, 95, 650.
- Berman M., & Diggle P. (1989). Estimating weighted integrals of the second-order intensity of a spatial point process. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, 51(1), 81–92. <https://doi.org/10.1111/j.2517-6161.1989.tb01750.x>
- Bernal J. L., Cummins S., & Gasparrini A. (2017). Interrupted time series regression for the evaluation of public health interventions: A tutorial. *International Journal of Epidemiology*, 46(1), 348–355. <https://doi.org/10.1093/ije/dyw098>
- Besbris M. (2020). Upsold: Real estate agents, prices, and neighborhood inequality. In *Upsold*. University of Chicago Press.
- Calonico S., Cattaneo M. D., & Titiunik R. (2014a). Robust data-driven inference in the regression-discontinuity design. *The Stata Journal*, 14(4), 909–946. <https://doi.org/10.1177/1536867X1401400413>
- Calonico S., Cattaneo M. D., & Titiunik R. (2014b). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica: Journal of the Econometric Society*, 82(6), 2295–2326. <https://doi.org/10.3982/ECTA11757>
- Cattaneo M. D., Titiunik R., & Vazquez-Bare G. (2017). Comparing inference approaches for RD designs: A re-examination of the effect of head start on child mortality. *Journal of Policy Analysis and Management*, 36(3), 643–681. <https://doi.org/10.1002/pam.2017.36.issue-3>
- Cattaneo M. D., Titiunik R., & Vazquez-Bare G. (2019). The regression discontinuity design. In L. Curini and R. J. Franzese (Eds.), *Handbook of research methods in political science and international relations*. Sage Publications.
- Cook T. D., Campbell D. T., & Day A. (1979). *Quasi-experimentation: Design & analysis issues for field settings (Vol. 351)*. Houghton Mifflin Boston.

