Sampling for spatial discontinuity designs*

Florence Kondylis John Loeser Jeremy Magruder

April 20, 2026

Abstract

Spatial discontinuity designs are ubiquitous as an identification strategy in the evaluation of place-based polices and infrastructure developments. Yet little guidance exists for sampling and data collection in the context of spatial discontinuity designs. What identifying assumptions are required to estimate causal impacts, and what sampling strategies lend themselves to spatial discontinuity designs? In this chapter, we describe the identification, sampling, and data collection strategies employed by Jones et al. (2022) (JKLM), who leverage a discontinuity in access to irrigation at the boundaries of hillside irrigation schemes in Rwanda to evaluate their impacts. We discuss challenges unique to sampling and data collection for spatial discontinuity designs, relative to regression discontinuity designs, and how these were addressed in JKLM. Specifically, we discuss the identifying assumptions required to accommodate spatial units of observation that typically occupy discrete areas (e.g., agricultural plots in JKLM). We then cover the intricacies of matching spatial units to informed respondents (e.g., plots to farmers in JKLM). Finally, we address the need to account for spillovers across spatial units of observation (e.g., a single farmer managing multiple plots in the context offered by JKLM). We frame our discussion as a case study in the broader context of sampling and data collection for spatial discontinuity designs in development economics, and discuss alternative identification and sampling strategies used in the literature.

*Kondylis: World Bank, fkondylis@worldbank.org; Loeser: World Bank, jloeser@worldbank.org; Magruder: UC Berkeley, NBER, jmagruder@berkeley.edu. We thank the European Union, the Global Agriculture and Food Security Program (GAFSP), the World Bank Rwanda Country Management Unit, the World Bank i2i fund, 3ie, and IGC for generous research funding. The views expressed in this manuscript do not reflect the views of the World Bank. All errors are our own.

Learning objectives

- Define a spatial discontinuity design for causal inference
- Describe typical assumptions for causal identification used in spatial discontinuity designs in empirical practice
- Identify the steps to sampling respondents for a spatial discontinuity design when existing data is not available

1 Identification and estimation with a spatial discontinuity

1.1 What is a spatial discontinuity design?

What is a spatial discontinuity design? When a policy varies discontinuously across a geographic boundary, but other characteristics of individuals do not, average differences in outcomes between individuals on either side of the geographic boundary are attributable to the policy. In a spatial discontinuity design, the impact of the policy is identified as the difference in outcomes between individuals at the boundary who happen to lie on the sides exposed and not exposed to the policy.[1]

The spatial discontinuity design in JKLM In JKLM, we observed that access to irrigation (the "policy") varied discontinuously across gravity fed irrigation canals (the geographic boundary); a photograph of this discontinuity is presented in Figure 1. In Figure 1, plots directly above the canal are not irrigated, and are cultivated predominantly with maize (left of the figure). In contrast, plots directly below the canal, which have access to irrigation are either visibly irrigated and cultivated with horticulture (top right of the figure), or are under preparation, perhaps for a shorter cycle crop than maize (bottom right of the figure). If these plots would have been cultivated with the same crops absent access to irrigation, differences in crop choice between these plots are attributable to the impact of access to irrigation. This is likely the case if the placement of the canal is not affected by determinants of preferred

[1] Keele and Titiunik (2015), Keele et al. (2017), and Kondylis and Loeser (2019) review spatial discontinuity designs, and we follow their discussions closely in this Section 1.1.

cultivation choices on these plots, as JKLM argue is the case. Consequentially, the impact of access to irrigation on cultivation of horticulture and use of irrigation is simply the differences in these outcomes between plots below the canal, which have access to irrigation, and plots above the canal, which do not.

Figure 1: Discontinuity in access to irrigation across a gravity-fed canal

Notes: A photograph of plots adjacent to a canal in Karongi 12, one of the hillside irrigation schemes studied in JKLM, is presented in this figure.

Identifying assumptions The key assumption underpinning identification of policy impacts in a spatial discontinuity design is conditional exogeneity.[2] Under conditional exogeneity, units of observation near the boundary, on either side of the boundary, would have had the same average outcomes absent the policy of interest. In JKLM, conditional exogeneity imposes that, absent access to irrigation below the canal, the same crops would have been cultivated on plots close to the canal on either side. Under conditional exogeneity, differences in outcomes for units of observation near the boundary, on either side of the boundary, reflect the impact of the policy of

[2] Keele et al. (2017) discuss this assumption in depth as "geographic mean independence".

interest.

The conditional exogeneity assumption is distinct from the "continuity framework"-based identifying assumption typically used in regression discontinuity designs (Cattaneo and Titiunik, 2022). Under the continuity framework, it is assumed that individuals can be sampled arbitrarily close to the discontinuity. In spatial discontinuity designs, as is generally true of spatial economic data, units of observation occupy discrete space, and it is therefore often not possible to sample units arbitrarily close to the boundary (Keele et al., 2017).

Testing internal validity Standard tests of internal validity for causal inference apply to the conditional exogeneity assumption for spatial discontinuity designs. These include:

- **Balance** tests demonstrate that units on either side of the geographic boundary are observably similar before the introduction of the policy of interest, both in terms of fixed characteristics and pre-policy outcomes. Balance tests double as a test on inference: if standard errors are too small, balance tests will overreject. In JKLM, balance corresponds to the test that characteristics of plots, and their managers, and plot outcomes before the construction of the hillside irrigation scheme, are similar above and below the canal.

- Tests of **no differential attrition** demonstrate that outcomes are equally liked to be observed for units on either side of the boundary. In JKLM, tests of no differential attrition correspond to showing that rentals and purchases of plots, non-cultivation of plots, and household non-response, do not differentially induce plot-level attrition above relative to below the canal.

- **Robustness** tests demonstrate that results are unaffected by alternative specifications of the controls conditional on which exogeneity is expected to hold. Other complementary approaches to testing robustness include testing alternative sampling bandwidths and alternative approaches to inference (alternative levels of clustering for standard errors, or randomization inference). In JKLM, tests of robustness correspond to showing that results are unaffected by the choice of specifications discussed in Section 1.2.

1.2 Identification and estimation in spatial discontinuity designs in practice

The units of observation in spatial discontinuity designs, and the nature of the boundary, vary substantially across applications; this has led to a range of approaches to estimation, although all build on a version of the conditional exogeneity assumption. We describe estimation motivated by four versions of the conditional exogeneity assumption below, which we link to sampling strategies in our discussion in Section 2. In Section 1.2.1, we describe estimation when exogeneity holds within a short distance of the boundary. We then describe estimation when exogeneity holds conditional on additional controls for distance to the boundary (Section 1.2.2), or on spatial controls for spatial heterogeneity (Section 1.2.3).

1.2.1 Exogeneity conditional on proximity to discontinuity

The most basic version of conditional exogeneity imposes that, absent the policy of interest, units of observation within some fixed distance of the boundary would have had the same average outcomes. Keele et al. (2017) call this assumption "geographic mean independence", and note its equivalence to the "local randomization" framework for regression discontinuity designs with distance to the boundary as the running variable (Cattaneo and Titiunik, 2022).

Let Y_i denote the outcome Y for unit of observation i. Let D_i be a binary indicator that i is on the side of the boundary exposed to the policy of interest. We can estimate the effect of the policy with the estimating equation

$$Y_i = \alpha + \beta D_i + \epsilon_i \tag{1}$$

where ϵ_i is an error term, and we restrict to observations within a fixed distance of the boundary. If exogeneity conditional on proximity to discontinuity holds, β is the effect of the policy of interest.

Application Perhaps the known application of a spatial discontinuity design is Card and Krueger (1994), who compare fast food restaurants on either side of the Pennsylvania-New Jersey border, before and after an increase in the New Jersey minimum wage. After restricting to observations close to the border (that is, by excluding western Pennsylvania), they argue that these restaurants would have experienced the same changes in employment absent the increase in the New Jersey minimum wage.

1.2.2 Exogeneity conditional on distance to discontinuity

In some applications, one may be concerned that outcomes might systematically vary as one moves towards or away from the discontinuity; for example, in JKLM, we note that plots in the command area are systematically at lower elevations, as they lie below the irrigation canal, and therefore elevation is systematically decreasing as one moves towards the command area boundary. This concern motivates controlling for the running variable (along with additional corrections to inference) in the "continuity framework" for regression discontinuity designs; similarly, in spatial discontinuity designs, it is common to control for distance to the boundary, even when restricting to observations very close to the boundary.

Continue to let Y_i denote the outcome Y for unit of observation i, D_i a binary indicator that i is on the side of the boundary exposed to the policy of interest, and ϵ_i an error term. Let M_i denote the distance of observation i to the boundary, coded so distances are negative when $D_i = 0$ and positive when $D_i = 1$. Under exogeneity conditional on distance to discontinuity, we also include controls for a K-degree polynomial in M_i interacted with D_i to flexibly control for the running variable; Pei et al. (2022) provide tools to select k, while Gelman and Imbens (2019) present evidence that choosing $K = 1$ (that is, locally linear) is typically optimal in practice. We can estimate the effect of the policy with the estimating equation

$$Y_i = \alpha + \beta D_i + \sum_{k=1}^{K} \delta_k M_i^k + \sum_{k=1}^{K} \gamma_k D_i M_i^k + \epsilon_i \tag{2}$$

If exogeneity conditional on distance to discontinuity holds, β is the effect of the policy of interest.

Application: Linear control for distance An example application of this approach is in JKLM, in which we control linearly for distance to the boundary and its interaction with an indicator for lying within the boundary in our "RDD" specification. This control is motivated by the observation that, even within a small distance of the canal, plots above the canal will be systematically located higher on the hillside and this may be associated with differences in agronomic characteristics; as these differences are expected to be continuous with respect to distance to the canal, controlling linearly for distance can partly account for such differences.

Application: Robustness to flexible controls for distance In another application, Dell (2010) estimate the impacts of a historical system of forced labor, using a spatial discontinuity in its imposition, on long-run household welfare. They make use of part of the boundary which deviated from the Andean precipice, along which they demonstrate balance by showing that districts on either side of the boundary were observably similar before the imposition of the system of forced labor. To test robustness, they implement specifications that either control for polynomials ranging from degree one through four in distance to the boundary, either interacted or not interacted with an indicator for lying within the boundary, or control for polynomials ranging from degree one through four in latitude and longitude.

1.2.3 Exogeneity conditional on spatial fixed effects

In practice, when units of observation occupy discrete space and as distances become small, the notion of distance to the boundary as a running variable may lose some meaning; in the application of Dube et al. (2016), for instance, all units are located at the boundary, and in that sense have a distance to the boundary of 0. One common approach in this case is to apply "spatial fixed effects" (Magruder, 2012), that is to estimate impacts by taking differences in outcomes between spatially proximate units on either side of the boundary. Spatial fixed effects are motivated by a natural extension of the basic version of conditional exogeneity: absent the policy of interest, *nearby* units of observation within some fixed distance of the boundary would have had the same average outcomes. A number of closely related implementations of spatial fixed effects have emerged in practice, which may vary in appropriateness by the nature of the discontinuity.

Continue to let Y_i denote the outcome Y for unit of observation i, D_i a binary indicator that i is on the side of the boundary exposed to the policy of interest, and ϵ_i an error term. Let $\mathcal{S}_i$ denote the set of units locally proximate to i (including i); in our examples below, this will either be unit i and its pairwise match, i and all units with the same nearest boundary segment, all units within a fixed distance of i, or the k units nearest to i. We can estimate the effect of the policy with the estimating equation

$$Y_i - \frac{\sum_{j \in \mathcal{S}_i} Y_j}{|\mathcal{S}_i|} = \alpha + \beta \left(D_i - \frac{\sum_{j \in \mathcal{S}_i} D_j}{|\mathcal{S}_i|} \right) + \epsilon_i - \frac{\sum_{j \in \mathcal{S}_i} \epsilon_j}{|\mathcal{S}_i|} \tag{3}$$

If exogeneity conditional on spatial fixed effects holds, β is the effect of the policy of interest.

Application: Pairwise matching Dube et al. (2016) estimate the impacts of minimum wages using discontinuities in changes in minimum wages across state borders. They pair counties that share a border, but are in different states, and control for county-pair fixed effects. Estimated impacts with pairwise matching are differences in outcomes between units of observation (counties) on either side of the boundary (state borders), among paired observations (pairs of counties that share a border).

Application: Boundary segment fixed effects Dell (2010) divide their boundary into four equal-length segments, and control for boundary segment fixed effects. Estimated impacts with boundary segment fixed effects are differences in outcomes between units of observation (districts) on either side of the boundary, among units of observation closest to the same boundary segment.

Application: Spatial differencing Magruder (2013) estimate the impacts of minimum wages in Indonesia using discontinuities in changes in minimum wages across Province (and sometimes sub-Province) borders. Before running regressions of outcomes on treatment (minimum wages), they spatially demean outcomes and treatment by subtracting off their average across units of observation within some fixed distance. Estimated impacts with spatial differencing are differences in outcomes between units of observation on either side of the boundary; however, spatial differencing places a higher weight on units with many observations on the opposite side of the boundary within a given distance, in a manner similar to controlling linearly for the running variable in a one-dimensional regression discontinuity design.

Application: Spatial k-nearest neighbors differencing In JKLM, in our "SFE" specification, we first spatially demeaned outcomes and treatment (that a plot is located in the command area) by subtracting off their average across the five nearest plots. Estimated impacts with spatial k-nearest neighbors differencing are differences in outcomes between units of observation (plots) on either side of the boundary, with a higher weight placed on units i with many observations on the opposite side of the boundary either among i's k-nearest neighbors, or that have i as one of their k-nearest neighbors.

1.3 Unanswered questions on spatial discontinuity designs

This chapter, and the literature on spatial discontinuity designs, leaves many questions that warrant further discussion:

- Outcomes are often strongly spatially correlated, suggesting adjustments to inference beyond unit-level clustering are needed. This is complicated, potentially with bias, when remote sensed or proxy outcomes are used.
- Impacts are often heterogeneous across space, or may vary with respect to sampling probabilities. Alternatively, when the discontinuity is "fuzzy", impacts may be heterogeneous with respect to the strength of policy implementation across space. In all cases, observations are implicitly weighted, and this may shape the interpretation of results.
- Units of observation occupy discrete space, yet this fact is rarely incorporated formally into the econometrics of spatial discontinuity designs.

2 Sampling spatially around a spatial discontinuity

2.1 Where to sample?

Analogy to regression discontinuity designs In the canonical regression discontinuity design, it is imagined that with a larger-and-larger sample, the sample used to estimate the effect of the discontinuity will be progressively shrunk to an arbitrarily small bandwidth around the discontinuity. Consequentially, when sampling for a regression discontinuity design, it is optimal for internal validity to sample units as close as possible to the discontinuity (Cattaneo et al., 2019). Such sampling is possible, and desirable, when the location of units relative to the discontinuity is known, but key outcomes of interest require costly surveys to observe (Cattaneo and Titiunik, 2022).[3]

[3] See McKenzie (2016) for an extended discussion of statistical power in regression discontinuity designs.

Sampling for spatial discontinuity designs Similarly, in spatial discontinuity designs, differences between units are most credibly attributable to the boundary for units very close to the boundary. Common specifications discussed in Section 1.2 involved restricting to observations within small distances of the boundary, and observations further from the boundary do not contribute to estimates of treatment effects in such specifications. Under a fixed sample size, sampling observations within a small distance of the boundary therefore maximizes power.

In many cases, no sampling is needed; data may be available for all potential units of observation, as in Dube et al. (2016) for US counties, or when remotely sensed outcomes are of interest.

In other contexts, sampling may be possible from administrative data or a census. For instance, Card and Krueger (1994) sample *all* fast food restaurants from four chains within their discontinuity sample (eastern Pennsylvania and New Jersey), using white-pages telephone listings. However, in many contexts, such a "census" may not be available.

Constraints on sampling for spatial discontinuity designs However, in many important contexts, fully censusing units of observation and identifying their locations may not be practical. This was our case in JKLM and we discuss our context in Section 2.3. In a related example, Lowes et al. (2017) oversample communities in which individuals are disproportionately likely to have ancestry in their discontinuity sample, in a context (Kananga) in which no census of ancestral origins is available and conducting such a census would have been prohibitively costly. In general, the appropriate approach will vary across contexts.

2.2 How to sample?

Below, we summarize the approach to sampling plots, and surveying their associated plot managers, near the boundary of the command areas of the hillside irrigation schemes in JKLM. In addition, we discuss the generalization of the approach to sampling for spatial discontinuity designs more broadly, including when particular steps are and are not necessary.

Step-by-step sampling for spatial discontinuity designs

Context Section 2.3 describes the setting of JKLM, and how it motivated spatial sampling.

Step 1: The boundary Section 2.4.1 describes administrative data on the spatial discontinuity.

> Identify a shape file for the administrative boundary and sampleable region (e.g., non-water). Restrict to segment of boundary that is within b meters of sampleable region on both sides of boundary.

Step 2: Sampling space Section 2.4.2 describes the process of sampling GPS coordinates in the neighborhood of the spatial discontinuity.

> Draw a b-meter buffer around the discontinuity. Restrict buffer to sampleable region. Draw a m-meter grid of points within the restricted buffer. Sample n_1 points from the grid.

Step 3: From space to a listing Section 2.4.3 describes the process of interpretably describing each sampled GPS coordinate and identifying its appropriate respondent.

> Identify respondent associated with each point in the grid.

Step 4: From a listing to households Section 2.4.4 describes the process of sampling from this list to construct a survey sample of households and descriptions of their associated GPS coordinates.

> Identify unique respondents on grid. Sample n_2 unique respondents, and sample 1 point per respondent.

Step 5: Surveying households Section 2.4.5 describes the process of surveying this sample.

> Survey the respondent associated with each sampled point on the sampled point.

Testing exogeneity Section 2.5 describes the process of testing exogeneity.

Interpreting estimates Section 2.6 describes how the employed sampling strategy shapes the interpretation of estimates under the approaches to estimation described in Section 1.2.

2.3 Context, and why to (spatially) sample?

In JKLM, the ideal discontinuity sample around the boundary of hillside irrigation schemes was exceptionally local. Each command area, the area of the scheme in which farmers are able to access irrigation from the canal, was between 1 and 3 square kilometers. The hillside irrigation schemes lay in areas with substantial geographic heterogeneity, reflected in the photograph of one scheme in Figure 2: the schemes were built into rolling hills, with canals that were far from neither the tops of the hills nor the marshlands in the valley.

Figure 2: A hillside irrigation scheme

Notes: Appendix Figure A1 of JKLM, a photograph of Karongi 12 hillside irrigation scheme.

While censuses of households were available at the village-level, rural Rwanda is a context where many households have plots in villages that are not where they live (Kondylis, 2008). As a result, sampling individuals in villages located at the boundary would result in a low probability of sampling individuals with plots within a small distance from the boundary.

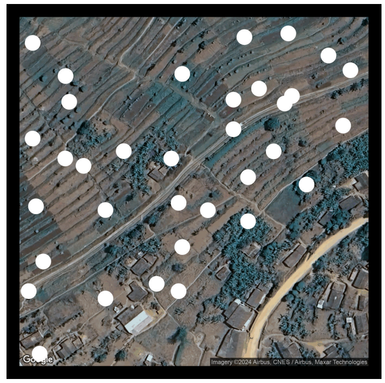
Alternatively, a census of households with land in command area was available. Unfortunately, this data did not identify the proximity of households' plots to the boundary. And perhaps more crucially, this data did not identify households with plots outside the command area.

This context motivated our "spatial sampling" approach described in Section 2.4: broadly, we directly sampled plots located near the boundary, and followed up with their managers.

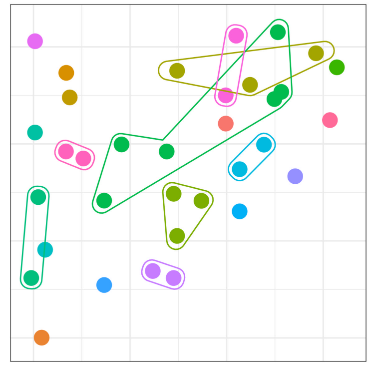
Figure 3: Implementing spatial sampling

(a) Step 1: Boundary

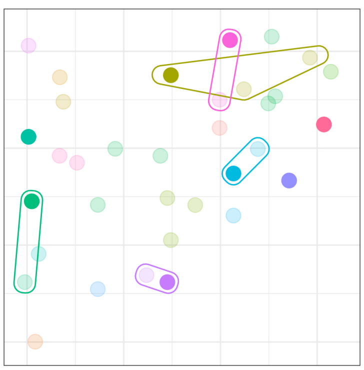
(b) Step 2: Sample space



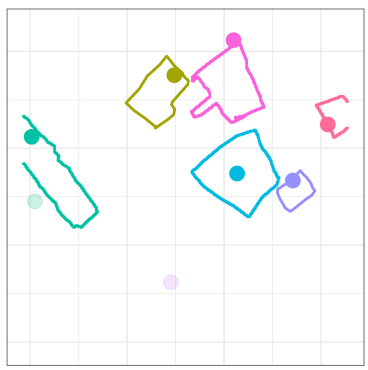
(c) Step 3: List respondents



(d) Step 4: Sample respondents



(e) Step 5: Survey respondents



(f) Plot maps from surveys

Notes: The steps of our spatial sampling procedure in JKLM are visualized in this figure along a segment of the command area boundary, from the boundary shapefile in 3a to the sample plot maps along that segment in 3f. Additional details are in Section 2.4.

2.4 Implementing spatial sampling

2.4.1 Administrative data on boundary

To sample plots within a small distance of the boundary, we began with a shapefile of the command area boundary, an example of which is presented in Figure 3a.

2.4.2 Sampling GPS coordinates

In a second step, we used the boundary shapefile to construct three sampling areas in each of the hillside irrigation sites with a discontinuity: the area within 50 meters of the boundary both inside and outside the command area, and the remainder of the command area. We then placed a grid of GPS coordinates at 2 meter resolution over the site, and sampled coordinates from each of the three areas (with an equal number of points sampled within 50 meters of the boundary inside and outside the command area).

An example of our sample of coordinates is presented in Figure 3b. We observe coordinates both outside the command area (on the bottom right) and coordinates inside the command area (on the top left). All coordinates outside the command area lie within 50 meters of the boundary, while some additional coordinates were sampled more than 50 meters from the boundary inside the command area.

2.4.3 From GPS coordinates to listing

In a third step, we provided enumerators with GPS devices with the coordinates of each sampled point (uploaded to the GPS devices as `.gpx` files), and asked enumerators to visit each sampled point; while visiting points, enumerators used the "record" feature of their GPS device, which allowed verification that enumerators visited the intended points. At each point that fell on cultivable land (80.1%), enumerators worked with a local informant (often farmers working nearby) to identify the cultivator responsible for the plot (name, village of residence, and phone number), and describe the plot with sufficient detail that the cultivator would be able to identify the exact plot described. Names of identified cultivators were then fuzzily matched, with the fuzzy matching confirmed through phone calls with village leaders.

An example of the output of this listing is visualized in Figure 3c; each of the points sampled in Section 2.4.2 (which, in this case, all happened to fall on cultivable land) is now colored based on the identity of the cultivator, with points belonging to a common cultivator enclosed in that cultivator's color. Many cultivators had multiple points sampled, which in some cases lay on multiple distinct plots.

2.4.4 From listing to survey sample

In a fourth step, we randomly sampled one point among those each cultivator was responsible for. An example output of this sampling is visualized in Figure 3d, in which only the sampled points are left darkened. Some cultivators have no sampled points in Figure 3d, as their sampled point lay in another area of the hillside irrigation scheme.

The description of the plot at the sampled point, which we refer to as the "sample plot", along with the identity of the cultivator, comprise our survey sample. Note that points, rather than plots (of which we did not have a comprehensive census), were sampled – as a consequence, very large plots were oversampled, which we return to in Section 2.6.

2.4.5 Surveying

In a final step, we surveyed cultivators on their sample plot, and also on their household and their other plots. Fuzzy matching the initial names and confirming through phone calls with village leaders was particularly important, and we were able to reach 90.2% (1,695 households) of our initial sample. Of these 1,695, we successfully mapped the sample plot for 91.6% (1,552).[4] An example of the mapped sample plots, and their associated sampled points, is presented in Figure 3e, with the plot maps overlaid with satellite imagery in Figure 3f.

2.5 Testing exogeneity

Our goal of sampling plots close to the boundary in Section 2.4 was to estimate the impacts of access to irrigation, by comparing plots just inside and just outside the command area, using the methods described in Section 1.2. These estimates recover the causal effect of access to irrigation under exogeneity conditional on proximity to the discontinuity: absent access to irrigation, outcomes would have been the same on plots just inside and just outside the command area.

In JKLM, we implement a number of tests of exogeneity: balance of household characteristics, balance of plot characteristics, balance of remotely sensed outcomes on sampled plots, and balance of attrition in follow up surveys. These are tests of exogeneity conditional on our sampling strategy. Such tests may therefore reject balance if either outcomes would have discontinuously varied at the boundary absent access to irrigation, or alternatively if our sampling strategy differentially sampled households or plots discontinuously at the boundary.

[4] Attrition was often due to data corruption during transfer from the GPS devices. Modern CAPI software enables plot mapping with the tablet used for surveying, which can be used with a bluetooth GPS receiver. This simplifies plot mappint protocols, and has eliminated data corruption as an issue.

2.6 Interpreting estimates

Under exogeneity conditional on proximity to the discontinuity, what weighted average of treatment effects do comparisons recover under this sampling strategy? Our sampling strategy, as mentioned in Section 2.4.4, oversamples large plots. To interpret this oversampling, it is helpful to consider two limiting cases. First, suppose that each household holds a vanishingly small share of total area: in this case, our procedure would identify exactly one point per cultivator, cultivators would be sampled with probability proportional to their total area managed, and plots within cultivator would be sampled with probability proportional to their area. Second, suppose that we sampled sufficiently densely that at least one point was sampled from each household: in this case, all cultivators would have been sampled, but plots within cultivator would be sampled with probability proportional to their area. In JKLM, we averaged 4.2 points per cultivator and identified multiple points for 63% of cultivators, suggesting that we sampled sufficiently densely (or that there were sufficiently few total cultivators) that in practice we were closer to sampling all cultivators. However, in general, this procedure will sample cultivators with probability increasing in their total area managed, and will sample plots within cultivator proportional to their area.

As a consequence, estimated impacts on sample plots using this sampling strategy reflect average effects on land within household,[5] with households weighted by their sampling probability. In JKLM, for instance, our discussion above would suggest that estimates reflect average impacts across households, with plots weighted proportionally to their area within household.

3 Putting it all together

We present example reproducible code for sampling GPS coordinates along an administrative boundary in Listing 1. In the example, 300 points are sampled within 10km of the boundary between Durham and Orange counties in North Carolina from a 100 meter resolution grid. We simulate listing and surveying households at each point, by treating each Census Tract as if it is a single household that is the respondent for all

[5]The interpretation of these estimated impacts, distinct from the weighted average of comparisons they reflect, is shaped by across plot spillovers, which we discuss at length in JKLM.

points within the Census Tract.[6] We implement the steps in Section 2.4 as follows:

- **Step 1: The boundary** We use the administrative boundary between Durham and Orange counties.
- **Step 2: Sampling space** We draw a 10km buffer around the boundary. We then restrict the buffer to Durham and Orange counties. We then draw a 100 meter resolution grid within the buffer, and sample 300 points from the grid.
- **Step 3: From space to a listing** We intersect all points with the Census Tract boundaries for Durham and Orange counties.
- **Step 4: From a listing to households** We identify the unique Census Tracts (62) intersected by our 300 points, and sample 1 point per Census Tract.
- **Step 5: Surveying households** We merge data on Census Tract median household income.

In addition, we present examples of estimating spatial discontinuity designs following the approaches laid out in Section 1.2, and visualize identifying variation under each design.

[6] Since, of course, data on all Census Tracts is observed, one could have considered simply using the centroid of each Census Tract. The code below should therefore be interpreted as simulating the process of sampling when data collection is needed.

Listing 1: Example R code for sampling and estimating a spatial discontinuity design

```r
library(sf)
library(tidyverse)
library(tidycensus)
library(fixest)

#### SETTING CONSTANTS ####
w <- 10000       # BUFFER WIDTH
gw <- 100        # GRID RESOLUTION
n <- 300         # SAMPLED POINTS
#### COORDINATE REFERENCE SYSTEMS ####
crs <- 32119     # PROJECTION FOR NC
gpscrs <- 4267   # GPS COORDINATES PROJECTION
#### REGRESSION PARAMETERS ####
sfeknearest <- 5 # NO. OF NEAREST NEIGHBORS FOR SFE

#### READ DATASETS ####
# NC SHAPEFILE
nc <- st_read(system.file("shape/nc.shp", package="sf"))
# NC CENSUS TRACT DATA
ncdata <- get_acs(
state = "NC", county = c("Durham", "Orange"), geography = "tract",
variables = "B19013_001", geometry = T, year = 2020
)
ncdata <- ncdata %>% rename(medhhinc = estimate)
# APPLY PROJECTIONS
nc <- st_transform(nc, crs)
ncdata <- st_transform(ncdata, crs)

#### SAMPLING SPACE ####
# SUBSET TO DURHAM AND ORANGE COUNTIES
nc <- nc %>% filter(NAME %in% c("Durham", "Orange"))
# IDENTIFY BOUNDARY BETWEEN DURHAM AND ORANGE
ncbdry <- st_intersection(st_boundary(nc[1,]), st_boundary(nc[2,]))
# CONSTRUCT w BUFFER AROUND BOUNDARY
ncbuffer <- st_buffer(ncbdry, w)
```

```
# INTERSECT BUFFER WITH SAMPLE REGION
ncsample <- st_intersection(nc, ncbuffer)
# CONSTRUCT GRID WITH RESOLUTION gw
g <- st_bbox(ncsample)
g[1:2] <- floor(g[1:2] / gw) * gw
g[3:4] <- ceiling(g[3:4] / gw) * gw
g <- expand.grid(x = seq(g[1], g[3], gw), y = seq(g[2], g[4], gw))
# INTERSECT GRID WITH RESTRICTED BUFFER
g <- st_as_sf(g, coords = c("x", "y"), crs = crs) %>%
st_intersection(st_union(ncsample))
# SAMPLE n FROM GRID
set.seed(1234)
gsample <- g[sample(1:nrow(g), n, replace = F),]

#### "LISTING" RESPONDENTS AND "SURVEYING" SPACE ####
# "LISTING" = MATCH CENSUS TRACT DATA TO SAMPLE
gsample <- st_join(gsample, ncdata)
# SAMPLE ONE OBSERVATION PER TRACT
set.seed(5678)
gsample <- gsample %>%
group_by(NAME) %>%
mutate(sampled = row_number() == sample(n(), 1)) %>%
ungroup
# "SURVEYING" = SET NON-SAMPLED OBSERVATIONS TO MISSING
gsample <- gsample
mutate(medhhinc = ifelse(!sampled, NA, medhhinc))

#### IMPLEMENTING SPATIAL DISCONTINUITY DESIGN ####
# CONSTRUCT DURHAM INDICATOR
gsample$d <- st_intersects(gsample, nc %>%
filter(NAME == "Durham"), sparse = F)
# CONSTRUCT DISTANCE INDICATOR
gsample$m <- st_distance(gsample, ncbdry) %>% as.numeric
# MAKE DISTANCE NEGATIVE IN DURHAM
gsample$m <- gsample$m * ifelse(gsample$d, 1, -1)
```

```
# CONSTRUCT SPATIAL FIXED EFFECT DIFFERENCED VARIABLES
gsample_sfe <- gsample %>% filter(!is.na(medhhinc)) %>% mutate(intercept = 1)
# WEIGHT MATRIX IS 1 FOR sfeknearest OBSERVATIONS AND ZERO OTHERWISE
gsample_weights <- st_distance(gsample_sfe, gsample_sfe) %>%
apply(MARGIN = 2, function(x) x <= sort(x, decreasing = F)[sfeknearest]) %>% t
# CONSTRUCTING SFE DEMEANED Y AND D
gsample_sfe$logmedhhincsfe <- log(gsample_sfe$medhhinc) -
(gsample_weights %*% log(gsample_sfe$medhhinc) /
gsample_weights %*% gsample_sfe$intercept)
gsample_sfe$dsfe <- gsample_sfe$d -
(gsample_weights %*% gsample_sfe$d /
gsample_weights %*% gsample_sfe$intercept)
# CONSTRUCTING DISTANCE-DEMEANED D
gsample_sfe$dresid <- feols(d ~ m + m:d, data = gsample_sfe)$resid
# MERGING BACK TO SAMPLE
gsample_sfe <- gsample_sfe %>%
select(GEOID, NAME, sampled, logmedhhincsfe, dsfe, dresid) %>% data.frame
gsample <- gsample %>% left_join(gsample_sfe,
by = c("GEOID", "NAME", "sampled"))

#### VISUALIZING SPATIAL DISCONTINUITY ####
# VISUALIZING SAMPLE, DISTANCE, AND BOUNDARY
ggplot() +
geom_sf(data = ncbdry) +
geom_sf(data = gsample, aes(col = m, shape = d)) +
geom_sf(data = gsample %>% subset(!sampled), aes(shape = d), col = "white") +
scale_color_gradient2(low = "#6BAED6", mid = "#4A4A4A", high = "#FD8D3C",
midpoint = 0, limits = c(-w, w))
# VISUALIZING CONDITIONAL ON DISTANCE TO DISCONTINUITY
ggplot() +
geom_sf(data = ncbdry) +
geom_sf(data = gsample %>% subset(sampled & !is.na(dresid)), aes(col = dresid))
scale_color_distiller(palette = "RdBu", limits = c(-0.5, 0.5))
# VISUALIZING CONDITIONAL ON SPATIAL FIXED EFFECTS
ggplot() +
geom_sf(data = ncbdry) +
```

```
geom_sf(data = gsample %>% subset(sampled & !is.na(dsfe)), aes(col = dsfe)) +
scale_color_distiller(palette = "RdBu", limits = c(-0.6, 0.6))

#### ESTIMATING A SPATIAL DISCONTINUITY DESIGN ####
# EXOGENEITY CONDITIONAL ON PROXIMITY TO DISCONTINUITY
feols(log(medhhinc) ~ d, cluster = "NAME", data = gsample)
# EXOGENEITY CONDITIONAL ON DISTANCE TO DISCONTINUITY
feols(log(medhhinc) ~ d + m + d:m, cluster = "NAME", data = gsample)
# EXOGENEITY CONDITIONAL ON SPATIAL FIXED EFFECTS
feols(logmedhhincsfe ~ dsfe, cluster = "NAME", data = gsample)
```

3.1 Exercises

1. Reproduce the code in Listing 1.

2. Pick two counties in a different state that share a boundary.

   – Use `tidycensus::get_acs` to sample all Census Tracts in those counties and their boundaries.

   – Use `sf::st_union` to construct polygons for each county.

   – Replicate the code above for the two counties you selected.

References

Card, David and Alan B Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *The American Economic Review*, 1994, *84* (4), 772–793.

Cattaneo, Matias D and Rocio Titiunik, "Regression discontinuity designs," *Annual Review of Economics*, 2022, *14* (1), 821–851.

_, _, and Gonzalo Vazquez-Bare, "Power calculations for regression-discontinuity designs," *The Stata Journal*, 2019, *19* (1), 210–245.

Dell, Melissa, "The persistent effects of Peru's mining mita," *Econometrica*, 2010, *78* (6), 1863–1903.

Dube, Arindrajit, T William Lester, and Michael Reich, "Minimum wage shocks, employment flows, and labor market frictions," *Journal of Labor Economics*, 2016, *34* (3), 663–704.

Gelman, Andrew and Guido Imbens, "Why high-order polynomials should not be used in regression discontinuity designs," *Journal of Business & Economic Statistics*, 2019, *37* (3), 447–456.

Jones, Maria, Florence Kondylis, John Loeser, and Jeremy Magruder, "Factor market failures and the adoption of irrigation in rwanda," *American Economic Review*, 2022, *112* (7), 2316–2352.

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

Keele, Luke, Scott Lorch, Molly Passarella, Dylan Small, and Rocio Titiunik, "An Overview of Geographically Discontinuous Treatment Assignments with an Application to Children's Health Insurance ," *Regression discontinuity designs: Theory and applications*, 2017, pp. 147–194.

Kondylis, Florence, "Agricultural outputs and conflict displacement: Evidence from a policy intervention in Rwanda," *Economic Development and Cultural Change*, 2008, *57* (1), 31–66.

_ **and John Loeser**, "Spatial jumps," 2019.

Lowes, Sara, Nathan Nunn, James A Robinson, and Jonathan L Weigel, "The evolution of culture and institutions: Evidence from the Kuba Kingdom," *Econometrica*, 2017, *85* (4), 1065–1091.

Magruder, Jeremy R, "High unemployment yet few small firms: The role of centralized bargaining in South Africa," *American Economic Journal: Applied Economics*, 2012, *4* (3), 138–166.

_ , "Can minimum wages cause a big push? Evidence from Indonesia," *Journal of Development Economics*, 2013, *100* (1), 48–62.

McKenzie, David, "Power calculations for regression discontinuity evaluations: Part 3," 2016.

Pei, Zhuan, David S Lee, David Card, and Andrea Weber, "Local polynomial order in regression discontinuity designs," *Journal of Business & Economic Statistics*, 2022, *40* (3), 1259–1267.