Geographic Boundaries as Regression Discontinuities*

Luke Keele† Rocío Titiunik‡

March 27, 2011

Abstract

We explore the use of geographic boundaries as regression discontinuities (RD), studying designs where the assignment variable is distance to a political boundary and subjects on either side of this boundary are compared. We develop the identification assumptions behind RD designs of this type and suggest that the key assumption is more likely to be violated, since agents are better able to sort around the discontinuity. Moreover, we show that geographic RD designs that employ a naive notion of distance as the assignment variable fail to recover the treatment effects of interest, and develop an new estimator that is faithful to the inherently spatial qualities of the design. We illustrate our argument and method with an application to voter turnout that investigates whether ballot initiatives increase turnout by exploiting a political boundary as a regression discontinuity. We focus on a 2008 initiative that was on the ballot in the city of Milwaukee but not in Milwaukee county.

1 Introduction

Selection and endogeneity are often key threats to inference with observational data. Recently, analysts have turned to natural experiments and quasi-experimental methods as one way overcome these obstacles in observational studies. Among these quasi-experimental techniques, the regression discontinuity (RD) design has been revived with great fanfare, particularly in economics, but also in political science. Lee and Lemieux (2010) summarize the promise that surrounds RD designs: "Another reason for the recent wave of research

*Authors are in alphabetical order. We thank Matias Cattaneo and Marc Meredith for comments and discussion. We thank Mark Grebner for assistance with acquiring the Wisconsin Voter File.

†Associate Professor, Department of Political Science, 2137 Derby Hall, Ohio State University, Columbus, OH 43210 Phone: 614-247-4256, Email: keele.4@polisci.osu.edu

‡Assistant Professor, Department of Political Science, P.O. Box 1248, University of Michigan, Ann Arbor, MI 48106 Phone: 734-936-2939, Email: titiunik@umich.edu

is the belief that the RD design is not 'just another' evaluation strategy and that causal inferences from RD designs are potentially more credible than those from typical 'natural experiment' strategies (e.g., differences-in-differences or instrumental variables), which have been heavily employed in applied research in recent decades." Recently, RD designs have gained further credibility by recovering experimental benchmarks (Green et al. 2009; Cook et al. 2008)

In the simplest version of the RD design, we observe a dichotomous treatment assignment that is a deterministic function of a single, observed, continuous covariate or "score." Treatment is assigned to those individuals whose score crosses a known threshold. The average outcomes for individuals just below the threshold are assumed to represent a valid counterfactual for treated individuals just above the threshold.

The use of RD designs has exploded in economics recently. Lee and Lemieux (2010) count 78 applications of RD designs in economics, and the design is spreading quickly in the political science literature (Butler 2009; Butler and Butler 2006; Broockman 2009; Eggers and Hainmueller 2009; Gerber et al. 2011; Hopkins and Gerber 2009). One particular type of RD design exploits discontinuities in geography. In this form of the RD design, which we interchangeably call the *geographic RD* or the *geographic discontinuity (GD)* design, the discontinuity threshold is a geographic boundary such as a school district or national border. In economics, such designs are often used to estimate the effect of school quality on house prices(Black 1999; Bayer et al. 2007; Lavy 2006). In political science, political boundaries are often associated with variation in key treatments such as national or state institutions. Important variation in political boundaries has led analysts to often adopt this geographic discontinuity design (Posner 2004; Miguel 2004; Krasno and Green 2008; Berger 2009).

We argue that the geographic RD design is not just another RD design. In their seminal paper, Hahn et al. (2001) briefly suggest that geographic discontinuities are identified under the same assumptions as classic RD designs based on scholarship cutoffs on standardized tests. Contrary to this view, we prove that the GD design requires different identification

assumptions than standard RD designs. When the sharp RD design is applied to geographic boundaries, the assignment variable must be defined as a two-dimensional distance between points on a map. Once a second dimension is added to the assignment variable, the standard continuity assumptions required for identification must be generalized, and a consequence of this generalization is that the RD design now identifies an infinite number of treatment effects, one at every point on the discontinuity boundary. We develop a nonparametric estimation method for treatment effects that is specially suited to deal with these issues.

We also argue that the key assumption of continuity that belies the RD design is often unlikely to hold when applied to geographic boundaries. That is, with geographic discontinuities we expect that agents will be able to sort very precisely around geographic boundaries, which may undermine identification. More generally, we argue that great care and considerable substantive knowledge is needed to successfully exploit political boundaries as RD designs. We explore the complications of geographic discontinuity designs within the context of a substantive example on turnout. Specifically, we explore whether ballot initiatives can increase turnout. We study the effect of an initiative on the ballot in the city of Milwaukee that did not appear on the ballot in the rest of the county.

2 Identification with a Geographic Discontinuity

In a regression discontinuity design, assignment of a binary treatment, T, is a function of a known covariate, S, usually referred to as the *forcing variable* or *score*. In the sharp RD design, treatment assignment is a deterministic function of the score, where all units with score less than the known cutoff $S = \bar{s}$ are assigned to the control condition ($T = 0$) and all units above the cutoff are assigned to the treatment condition ($T = 1$). In contrast, in a fuzzy design the assignment to treatment is a random variable given the score, but the probability of receiving treatment conditional on the score, $P(T = 1|S)$, still jumps discontinuously at $\bar{s}$. In both cases, the crucial aspect of the design is that the probability of receiving treatment jumps discontinuously at the known cutoff $\bar{s}$. In what follows, we focus on the sharp RD

design, since all examples of geographic discontinuities have a deterministic assignment by definition.

We follow the literature and adopt the potential outcomes framework, and assume that individual i has two potential outcomes, Y_{i1} and Y_{i0}, which correspond to both levels of treatment, $T_i = 1$ and $T_i = 0$, respectively. The observed outcome is $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$, and the fundamental problem of causal inference is that we cannot observe both Y_{i1} and Y_{i0} simultaneously for any given individual. A regression discontinuity design provides a possible way to identify the parameter of interest, at least locally. In the sharp RD design, $T_i = \mathbf{1}\{S_i > \bar{s}\}$, where $\mathbf{1}\{\cdot\}$ is the indicator function.

Hahn et al. (2001) demonstrate that the key condition needed for identification is that the potential outcomes are a *continuous* function of the score. Under this continuity assumption, the potential outcomes can be arbitrarily correlated with the score, so that, for example, people with higher scores might have higher potential gains from treatment. Focusing on the regression function, this assumption can be formally stated as follows:

A1: Continuity in one-dimensional score. The conditional regression functions are continuous in s at $\bar{s}$:

$$\lim_{s \to \bar{s}} E(Y_{i0}|S_i = s) = E(Y_{i0}|S_i = \bar{s})$$

$$\lim_{s \to \bar{s}} E(Y_{i1}|S_i = s) = E(Y_{i1}|S_i = \bar{s}).$$

Since $Y_i = Y_{i1}$ when $T_i = 1$, $Y_i = Y_{i0}$ when $T_i = 0$, and $T_i = \mathbf{1}\{S_i \geq \bar{s}\}$, assumption **A1** implies

$$\lim_{s \to \bar{s}^+} E(Y_i|S_i = s) = E(Y_{i1}|S_i = \bar{s})$$

and

$$\lim_{s \to \bar{s}^-} E(Y_i|S_i = s) = E(Y_{i0}|S_i = \bar{s}),$$

which is a formal statement of the intuition that individuals very close to the cutoff and on

opposites sides of it are comparable or good counterfactuals for each other. From this, it follows that

$$\lim_{s \to \bar{s}^+} E(Y_i|S_i = s) - \lim_{s \to \bar{s}^-} E(Y_i|S_i = s) = E(Y_{i1} - Y_{i0}|S_i = \bar{s}).$$

Thus, continuity of the conditional regression function is enough to identify the average treatment effect *at the cutoff*. That is, the RD design identifies a *local* average treatment effect for the subpopulation of individuals whose value of the score is (near) $\bar{s}$. Without further assumptions, such as constant treatment effects, the effect at $\bar{s}$ might or might not be similar to the effect at different values of S.

Lee (2008) provides a behavioral interpretation for the continuity assumption in the RD design. He demonstrates that when agents are able to precisely manipulate their value of S continuity of the conditional regression function is unlikely to hold. Formally, $S = Z + e$, where Z comprises efforts by agents to sort above and below $\bar{s}$ and e is a stochastic component. When e is small and agents are able to precisely sort around the threshold, the RD design may not identify the parameter of interest. This behavioral interpretation of the continuity will prove useful for the evaluation of GD designs.

Analysts who use geographic boundaries as regression discontinuities have relied on the identification conditions for the classic sharp RD design. Typically, GD designs compare two adjacent areas; in one of the areas, all individuals residing in that area are assigned to the control condition, and in the other, all individuals are assigned to the treatment condition. Henceforth, we call these areas the *control area* and the *treatment area*, and we denoted them by A_c and A_t, respectively. In the GD design, treatment assignment jumps discontinuously along the boundary that separates A_c and A_t. In most applications of the GD design, the score S is defined as the shortest distance to the boundary, and units that are close to the boundary in terms of this distance but on opposite sides of the boundary are taken as valid counterfactuals for each other. In this setup, individual i has distance $S_i = d$ if the

distance from i's location to the point on the boundary that is closest to i is equal to d. A serious limitation of this strategy, however, is that it ignores the spatial nature of geographic locations. As illustrated in Figure 2, the shortest distance from individual i's location to the boundary does not determine the exact location of i in the map, since two individuals i and j in different locations can both have $S_i = S_j = d$. That is, this naive distance does not account for distance *along* the border. As one can see in Figure 2, a naive implementation of the RD design along a geographic boundary that does not take into account both dimensions would treat individuals i and j in the control area as equally distant from individual k in the treatment area, when in fact j is much closer to k than i. This problem will be exacerbated when the boundary is longer; in Figure 2, as the boundary becomes longer, the distance between control unit i and treated unit k can be made arbitrarily large even as $S_i = d$ remains constant, by moving i down along the dotted line.

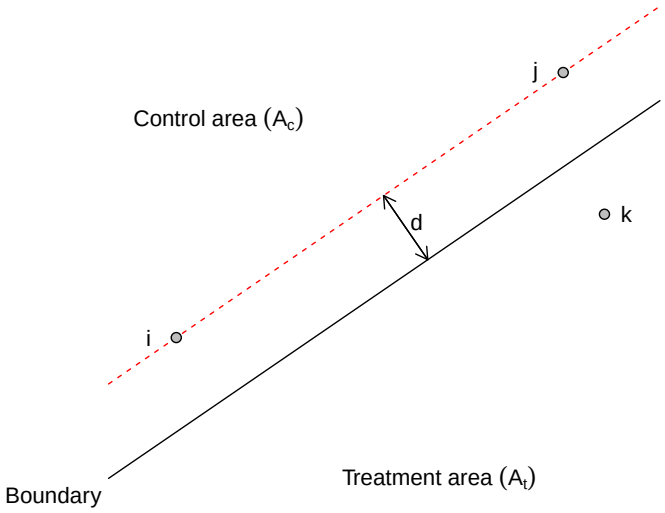


Figure 1: Failure of one-dimensional distance to identify boundary points

This suggests that applying the sharp RD design to geographic boundaries requires gen-

eralizing the score S_i to be a two-dimensional distance between points on a map. We reformulate the score so that $\mathbf{S}_i = (S_{i1}, S_{i2})$, where the score is now a function of two points that uniquely define i's location, such as the latitude and longitude.[1] Once we add a second dimension to the score, the continuity assumption required for identification must be generalized to the following assumption:

A2: Continuity in two-dimensional score. The conditional regression functions are continuous in (s_1, s_2) at all points $(\bar{s}_1, \bar{s}_2)$ on the boundary:

$$\lim_{(s_1,s_2)\to(\bar{s}_1,\bar{s}_2)} E\left\{Y_{i0}|(S_{i1}, S_{i2}) = (s_1, s_2)\right\} = E\left\{Y_{i0}|(S_{i1}, S_{i2}) = (\bar{s}_1, \bar{s}_2)\right\}$$

$$\lim_{(s_1,s_2)\to(\bar{s}_1,\bar{s}_2)} E\left\{Y_{i1}|(S_{i1}, S_{i2}) = (s_1, s_2)\right\} = E\left\{Y_{i1}|(S_{i1}, S_{i2}) = (\bar{s}_1, \bar{s}_2)\right\},$$

for all points $(\bar{s}_1, \bar{s}_2)$ on the boundary.

Note that in the case of a two-dimensional score, the left and right limits are no longer defined, since now any set of points $(\bar{s}_1, \bar{s}_2)$ on the boundary can be approached from an infinite number of directions. Moreover, there is no longer a single point at which treatment jumps discontinuously, but rather an infinite collection of points – the collection of all points on the boundary or line that separates the treatment and control areas. This has the important implication that the parameter identified by a geographic RD is not unidimensional but rather infinite-dimensional as it is a curve on a plane. In other words, since the cutoff is not a point but a *boundary*, under the appropriate two-dimensional continuity assumptions the GD design will identify the treatment effect at *each* of the boundary points.

Assuming **A2**, for any point $(\bar{s}_1, \bar{s}_2)$ on the boundary, we have

$$\lim_{(s_1,s_2)\to(\bar{s}_1,\bar{s}_2)} E\left\{Y_i|(S_{i1}, S_{i2}) = (s_1, s_2)\right\} = E\left\{Y_{i0}|(S_{i1}, S_{i2}) = (\bar{s}_1, \bar{s}_2)\right\}, \quad \text{for all} (s_1, s_2) \in A_c$$

[1] An alternative set of points could be easting and northing which are geographic Cartesian coordinates in the Universal Transverse Mercator coordinate system.

and

$$\lim_{(s_1,s_2)\to(\bar{s}_1,\bar{s}_2)} E\left\{Y_i|(S_{i1},S_{i2})=(s_1,s_2)\right\} = E\left\{Y_{i1}|(S_{i1},S_{i2})=(\bar{s}_1,\bar{s}_2)\right\}, \quad \text{for all}(s_1,s_2)\in A_t$$

Thus, under **A2**, the GD identifies a (possibly different) treatment effect *for every point on the boundary*:

$$\begin{aligned}\tau(\bar{s}_1,\bar{s}_2) &= \lim_{(s_1,s_2)\in A_t\to(\bar{s}_1,\bar{s}_2)} E\left\{Y_i|(S_{i1},S_{i2})\right\} - \lim_{(s_1,s_2)\in A_c\to(\bar{s}_1,\bar{s}_2)} E\left\{Y_i|(S_{i1},S_{i2})\right\} \\ &= E\left\{Y_{i1}-Y_{i0}|(S_{i1},S_{i2})=(\bar{s}_1,\bar{s}_2)\right\}\end{aligned}$$

The effect $\tau(\bar{s}_1,\bar{s}_2)$ evaluated at all all points $(\bar{s}_1,\bar{s}_2)$ on the boundary defines the (infinite dimensional) treatment effect curve. Not surprisingly, the treatment effect in a GD design is actually a spatial construct as it can change in space. Once the two-dimensional structure of the problem is recognized, the problem illustrated in Figure 2 can be easily avoided. Since there is an infinite number of distinct locations from where an individual can be equally close to the boundary, the one-dimensional distance to the closest point on the boundary does not identify a point on this boundary. Thus, implementing a GD design with a naive one-dimensional distance will not identify a treatment effect at a discontinuity point, since any such point can only be identified with two coordinates on the plane. Thus, applying a sharp RD design to a geographic discontinuity requires a modification of the identifying assumption, and the average treatment effect of interest must be defined as a curve rather than as a single parameter.

It is useful at this point to consider the GD design in the context of the behavioral interpretation of the continuity assumption. First, the move from **A1** to **A2** does not change the fact that continuity of the conditional regression function is less likely to hold if agents can precisely sort around the threshold. Second, when the discontinuity is a geographic boundary between cities, counties, school districts, etc., and the units of analysis are individuals who

reside in these areas, assuming that **A2** holds amounts to assuming that people cannot precisely sort around the boundary in a way that makes potential outcomes discontinuous. The assumption about the ability of agents to sort around the threshold may be strong in any RD design, but in a GD design we might expect that people will often be able to carefully select their place of residence based on the boundary of interest. That is, features such as the quality of schools, crime rates, distance to public transportation, the price of housing may all vary discontinuously at the boundary. In short, depending on the application, we may have reason to suspect that the stochastic component of the score, e, is quite small. If true, assumption **A2** will be violated. This contrasts with many conventional applications of RD, where often it can be assumed that sorting occurs relatively imprecisely. Thus, in any application of the GD design, analysts must carefully consider the ability of people to sort around the boundary. Analysts will need to carefully check whether pre-determined characteristics have the same distribution along the boundary. If the means of such characteristics jump discontinuously, there is little reason to think parameters of interest are identified. Substantive knowledge of the geographic boundary under consideration will also prove useful in understanding the ability of agents to sort around the discontinuity of interest. We now outline our application of interest, where the issues mentioned in this section will become apparent.

3 Application: Ballot Initiatives and Voter Turnout

One feature of the political arena in some states is the initiative process. While the method by which direct legislation is implemented varies, in 24 states citizens can place legislative statutes directly on the ballot for passage by the electorate. While the initiative process is often decried as populism run amok in the popular press, the consequences of initiatives are thought to be benign to favorable in much of the academic literature (Matsusaka 2004; Lupia and Matsusaka 2004; Smith and Tolbert 2004). For good or ill, few doubt that direct legislation changes outcomes across states, particularly directly on the issue area in question.

It is also thought, however, that initiatives have spill over effects on outcomes unrelated to the policy issue on the ballot. In particular, it is thought that ballot initiatives increase voter turnout (Tolbert et al. 2001; Smith and Tolbert 2004; Tolbert and Smith 2005). In fact both political parties often strategically sponsor ballot initiatives in hopes of boosting turnout among key constituencies (Gertner 2006).

Assessing the effect of initiatives on the behavior of citizens is a difficult task. Given that the initiative process was not randomly assigned across states and that states are very heterogeneous, we must exercise great caution before assuming initiatives *cause* a particular outcome. Both of these obstacles are common when attempting to make causal inferences with observational data. The problem is that it is quite likely a confounder exists which might be correlated with the presence of direct legislation and the outcome in question. That is states with a particularly progressive civic culture are probably more likely to adopt reforms like initiatives as well being more likely to vote. We must account for baseline differences across states before any valid comparisons can be made across states with and without direct legislation. We argue that past studies have drawn erroneous conclusions about the effects of the initiative process because they have not accounted for such baseline differences across states or taken into account causal heterogeneity.

First, we discuss why we might expect states with initiatives to have higher levels of voter turnout. Explanations for why citizens vote (or fail to vote) tend to be based on one of three general models of political participation: the socioeconomic status model, the rational choice model, or the mobilization model. The most controversial of these three models is the rational choice model, which tends to focus on the instrumental calculations behind the decision to participate in politics. Under this perspective, one votes if the following is true:

$$PB - C > 0$$

where P is the probability that one's vote is decisive, B is the net benefit from having one's preferred candidate win, and C is the net cost of voting. This model describes the "calculus

of voting" as one where voting occurs if the benefits of voting and the probability of being decisive outweigh the costs of voting (Downs 1957; Riker and Ordeshook 1968). While it is difficult to precisely define the size of B and C, P is equal to $1/n$, where n is the size of the electorate. As such, the probability of being decisive in most elections is very small; making it unlikely that the benefits ever outweigh the costs no matter their size. This model of participation has been widely criticized, and even Riker and Ordeshook acknowledged that purely instrumental calculations are insufficient to cause people to vote. They introduced a term for the experiential benefits of voting to account for the fact that the decision to vote is more than a cost benefit analysis on the part of citizens. While the calculus of voting model has been criticized on many fronts and revised in a number of ways, it remains a useful model for understanding voter turnout as it helps focus attention on the incentives for participation.[2]

The calculus of voting model provides an explanation for why the presence of ballot initiatives might increase participation in elections. The presence of a ballot initiative might change the calculus of voting in two ways. First, an initiative may reduce the size of P. Ballot initiatives reduces the size of the electorate participating and therefore increases P. The reduction in P, however, is likely to be trivial in most instances. Even in Wyoming, nearly 200,000 citizens voted in 2004. While P might be quite small in local elections, it is unlikely that initiatives reduce the probability of being decisive enough to matter in state elections. Ballot initiatives can, however, increase the benefits of voting and more importantly make those benefits more salient.

In a presidential or Congressional election, voter estimates of B must be imprecise. Electoral promises from candidates are often necessarily general and possibly ambiguous. Even if a candidate were to promise a large tax cut or a large increase in targeted public goods, once elected the politician may renege on the promise and even presidents can do little immediately without Congressional approval. Thus electoral victory does not ensure the payoff

[2]Whiteley (1995) provides a useful overview of the debate over rational choice models of participation.

of B. In contrast, initiatives often have precise payoffs (a reduction in taxes or a ban on smoking) and become law in a relatively short period of time if not immediately after the election. Therefore, initiatives are more likely to provide immediate and precise payoffs to voters which makes the benefits of voting more salient.

While it would appear that initiatives can increase the benefits of voting, the nature of initiatives makes the incentive to vote based on initiatives conditional. This conditionality is due to the differing content of initiatives: not every initiative promises clearly defined benefits. While Proposition 13 in California offered an obvious payoff to a well-defined constituency through lower property taxes, the benefits of many initiatives are diffuse and not well defined. For example, Proposition 60 in California required that all parties participating in a primary election would advance their candidate with the most votes to the general election. Passage of this initiative would provide a diffuse to negligible benefit for most voters. The promised benefit of an initiative such as Proposition 60 may not be enough to outweigh the costs of voting in that election for many voters. The conditional nature of initiative content also implies that the number of ballot initiatives is not necessarily indicative of increased turnout. Five initiatives without defined benefits may not increase turnout as much as a single initiative promising an obvious benefit. Recent work has demonstrated the voters are sensitive to the costs of voting in the form of the weather, so voters may be sensitive to the benefits through initiatives (Gomez et al. 2007).

Initiatives, however, certainly do not guarantee increased levels of voter turnout. Of course, given the size of P and C, the promised benefit of any initiative may not be enough to spur one to vote beyond initial inclinations. Or perhaps only those with sufficient individual resources will understand the benefits. Moreover, passage of an initiative does not guarantee it will be enacted. Many initiatives depend on cooperation from the state legislature, and there is evidence that state politicians do not always cooperate (Gerber et al. 2001). From a theoretical standpoint, then, it is unclear whether we should expect differences in voter turnout across states with and without direct legislation. The empirical literature, however,

offers an unequivocal answer. Several studies in the extant literature have found that states with initiatives have higher levels of voter turnout than states that do not (Tolbert et al. 2001; Smith and Tolbert 2004; Tolbert and Smith 2005).

3.1 Milwaukee Ballot Initiative

One of the difficulties with trying to make causal inferences about ballot initiatives is that this political institution is confounded with a variety of other state level institutions and the political culture which lead to the adoption of the initiative process itself. Because of this, it is nearly impossible to disentangle the effect of ballot initiatives from a host of other state level institutions such as absentee voting requirements and voter registration laws like election day registration. To avoid the difficulties of trying to make counterfactuals comparisons across states, we adopt a within-state design (Keele and Minozzi N.d.). That is, we rely on a comparison within the same state. The city of Milwaukee is one of many cities with an initiative process. What distinguishes it from many other cities with the initiative process is that the state of Wisconsin does not have initiatives. Most municipalities with initiatives are also in states with initiatives. Thus municipal initiatives typically appear along state-wide initiatives. Here, we are able to focus on an initiative that appeared on the ballot within the city limits of Milwaukee but did not appear on the ballot in the municipalities that surround Milwaukee and are also within Milwaukee county.

For the 2008 election, a coalition of local labor, educational and community organizations led by 9to5, the National Association of Working Women, helped place an initiative on the ballot that mandated all private employers in the city of Milwaukee to provide one hour of sick leave for every 30 hours worked. The initiative passed receiving slightly more than 68% of the vote. It was struck down by courts shortly after the election. On the county wide ballot, citizens also voted on a sales tax increase which passed as well. Thus we are able to isolate the effect of an additional high profile initiative on the ballot. We think this particular initiative provides a useful example for understanding the effects of initiatives on turnout

more generally. The initiative was easy to understand for voters. The exact language on the ballot was as follows:

> Shall the City of Milwaukee adopt Common Council File 080420, being a substitute ordinance requiring employers within the city to provide paid sick leave to employees?

This is also an issue that affects most voters and should be highly salient and received considerable attention. We found 64 different mentions of this initiative in the local news papers from July up until election day. Many initiatives ask voters to decide on more complex or less salient policy matters. But here we focus on a ballot initiative with easy to understand consequences that would be widely felt by citizens. Another key advantage to our design is that county is held constant. Election administration is conducted at the county level, and we might worry that a wealthy suburban county may spend more on polling places or voting technology. In our design county level confounders are held constant. Figure 2 contains a map of Milwaukee county. The areas in yellow comprise the city of Milwaukee which is surrounded by 17 suburban areas that are considered Minor Civil Divisions by the Census. Six of these municipalities do not share a border with Milwaukee while the rest have contiguous borders with the city to varying degrees. While these are suburban areas, these do not represent recent movements to the suburbs. While there has been substantial growth in the suburbs this growth has occurred much farther west along the I94 corridor in Waukesha county. Unfortunately school districts do not overlap city limits. We use school performance data to evaluate whether a clear divide exists near the city limit. Basic comparisons of Milwaukee to these municipalities clearly demonstrate that the city is more ethnically diverse has lower housing prices and incomes. Census data from 2000 reveals clear divisions. Median household income in the city is just under 34,000 dollars while it is nearly 54,000 dollars in these suburbs. The percentage of African-American residents that are of voting age in Milwaukee is 29% while it is less than 1.5% in these suburbs. The difference in median housing value is nearly $60,000. While the percentage of high school graduates

is nearly identical, nearly 21% in the suburbs have a college degree while just over 12% in the city have a college degree. We could try to adjust for such differences via regression of matching. The goal here is to understand whether exploiting the city limit as a discontinuity lends credibility to our inference. We now turn to details of the analysis.

4 Data

For our analysis, we merged data from four sources. In some instances, we used geographic software to perform the data merges as we describe below. Our main data sources is the Wisconsin Voter File, the database of registered voters maintained by the State for administrative purposes. The voter file for Wisconsin contains a limited number of covariates: date of birth, gender, and voting status. Wisconsin does not record either race or party registration in the voter file. While we do not have individual level measures of education and income, voting status is recorded for past elections this forms a key covariate in our analysis.

The rest of the data we use is aggregated to differing geographies. Most of the aggregate data is Census data in one of two forms. First, we collected block level data. Blocks are the lowest unit of census geography. The number of covariates available at the block level, however, is limited. At the block level, we have measures for the percentage of African-Americans, Hispanics, and minorities, as well as median-age. For the 2000 census, block group data provide a richer set of covariates. At this level, we used the percentage with a high school degree, the percentage with a college degree, median income, the percentage of unemployed, the percentage below the poverty level, the percentage married and the percentage of foreign born. We also collected measures on housing characteristic at the block group level. These measures included median housing values, the median rent asked and the percentage of owner occupied housing units. A discontinuity in housing values along the Milwaukee city limit is one obvious indication that our design is not identified. Finally, we also collected data from the Wisconsin Board of Elections. Data from the Board

of Elections is aggregated to the ward level, the precinct equivalent in Wisconsin. Here, we collected partisan vote shares for Federal offices and the governor as well as ward level turnout measures for the 2004 and 2006 elections.

5 Analysis

We next outline the two forms of analysis that we use to evaluate the GD design. We use Geographic Information Systems (GIS) techniques for both data management as well as to locate voters and calculate the coordinates needed for measuring spatial distance to the discontinuity. Later once we completed the GIS stage of the analysis, we developed statistical techniques to assess the identification assumption, estimate treatment effects, and account for spatial patterns in the data.

5.1 Geographic

It may be possible to avoid using GIS techniques when using a GD design. We would argue, however, that without GIS the GD design is significantly weakened. GIS software allows analysts to more fully exploit geography and spatial proximity. Here, we outline the geographic analysis we performed to implement the GD design in Milwaukee County.[3] First, we geocoded the voter file. Geocoding is the process of converting addresses into a coordinate system typically that of latitude and longitude. Geocoding allows us to know the distance between voters and the city limit which forms the discontinuity of interest.[4] Geocoding allows us to develop a score that reflects the two dimensional geographic space. Once we completed the geocoding, GIS software allows us to merge the individual level data from the voter file with covariates collected from larger geographies. For example, we would like to use both census data and election data which is gathered at the block, block group, or precinct level. Once geocoding is complete, we were able to locate each voter within the appropriate block, block group, and ward. This effectively merges each voter with the data

[3] We performed all the geographic analysis in ArcgGIS 9.4.

[4] Geocoding requires taking formatted addresses for each voter. These addresses are then compared to a known database of addresses and street locations.

we collected at the three different geographies.

One might assume we also used GIS software for calculating the score, the distance to the geographic discontinuity. Others analysts have used GIS software to calculate the distance between voters and politically relevant geographic points. Brady and McNulty (2011); Haspel and Knotts (2005) use GIS software to calculate the distance between voters' addresses and their polling location. The method used by these analysts calculates the shortest distance from each voter's address to the point of interest. In our case that would be the Milwaukee city limit. Such a distance can be calculated as either the driving distance along streets or as a direct distance as the crow flies so to speak. Other work on voter turnout has found little difference between these two distances (Brady and McNulty 2011; Haspel and Knotts 2005). As we demonstrated in Section 2 this distance does not identify the GD design. This distance is not spatial in the sense that while it calculates the distance to the boundary it does not measure distance *along* the boundary. We did calculate such a distance and use it for illustrative purposes later.

We can use the latitude and longitude obtained from the geocoding to calculate the spatial distance between voter residences and the city limit. One might imagine that a simple application of the Euclidean distance with the points defined by latitude and longitude would be sufficient for calculation of the score in the GD design. This would be appropriate if voters resided in a plane, but the Earth is a sphere. Naive Euclidean distances calculated between geographic locations can severely overestimate the distance (Banerjee 2005). There are two standard alternatives to the naive Euclidean distance: the geodetic and chordal distance. We use both the geodetic and chordal distance which is a rescaling of the Euclidean distance. The chordal distance is very close to the geodetic distance for locations that are less than 2000 km apart. The additional advantage of the chordal distance is that it allows for valid calculations of spatial correlations which the geodetic does not allow (Banerjee 2005).

Finally, we used GIS software for a number of smaller tasks. First, we created what is called a buffer around the city limit. The buffer is a spatial object that records which voters

fall within a specified distance of a geographic boundary here the city limit. We used a buffer to record which voters are within 50, 100, 200, 300, 400, 500, 750 and 1000 meters from the city limit. We use the buffer first as a modified naive score. Most uses of the GD design simply compare bands a fixed distance along the border. Using a buffer as the score will not identify the design since again this is not a spatial distance. We later combine the buffer with the chordal distance as a method of pruning the data for a more local estimate. Third, we use GIS software to divide the city limit into equal parts. In a GD design, we are able to make the estimates more local into two ways. One is by using the buffer to restrict the analysis to within some distance of the city border. We can also make the estimate more local by using one part of the city limit instead of the entire boundary. For example, we might only wish to compare the part of the city limit where Milwaukee borders the inner suburb of West Allis. We can accomplish this by using GIS software to divide the city limit into equal lengths and restrict the analysis to one of these locations. Finally, even though the treatment effect estimate identified in the GD design is a plane, for practical purposes we need actual points on the city limit to use for the calculation of treatment effects. We do this by dividing the city limit into points define by latitude and longitude spaced at equal intervals of 100 meters. Armed with these geographic tools, we now turn to more standard statistical tools.

5.2 Statistical

The first goal in our statistical analysis is quite basic. While we have proven that identification of the GD design requires a two dimensional score, the last section demonstrates that development of such a score is fairly trivial. Geographic measures of chordal or geodetic distance serve this purpose quite well. We argued in Section **??** that the key assumption in an RD design, continuity of the conditional regression function, is vulnerable to violation in the GD design since we might expect that voters will precisely sort around the discontinuity. That is, voters might choose to live either in or outside of Milwaukee due to differences in

property taxes, schools, housing, or a variety of other amenities. Thus in any GD design, analysts must carefully assess the plausibility of the continuity assumption. Indeed, most of our analysis seeks to understand the plausibility of this assumption. Once we decide the assumption is plausible, estimation of treatment effects may proceed. We start by outlining one method best suited to assumption assessment but poorly suited to estimation treatment effects. We then outline a method for estimating treatment effects, with limited ability to assess the continuity assumption.

5.2.1 Balance Analysis

We are concerned that the continuity assumption does not hold due to the ability of voters to sort around the boundary of interest in this case the Milwaukee city limit. Our method for understanding the quality of this assumption is based on standard practice in RD designs where the analyst looks for changes in pre-determined characteristics at the discontinuity. In other words, one looks for jumps in the distributions of pre-determined characteristics around the discontinuity. In our context, this means looking for sharp differences in covariate values round the Milwaukee city limit. For example, the most important covariate we have is voter turnout in previous elections where there were no initiatives on the ballot. While there might be clear differences in past turnout between the city and its immediate suburbs, those differences should decrease the closer we get to points along the city limit.

Here, we use the concept of balance-the degree of discrepancy between treated and control units-on observable characteristics from the matching literature to assess the continuity assumption. With matching estimators, once matching is complete the goal is for observed pretreatment characteristics to be nearly identical across treatment and control groups. The same is true in an RD design, but balance should be a function of the score and should improve as one moves closer to the discontinuity. We apply this logic by testing whether balance improves as function of distance to the discontinuity, here, the city limit. We assess balance as a function of distance using 27 covariates. The list of covariates that we use are in Table 1. The covariates are measured at four different levels of aggregation from the

individual level up to the block group and precinct level which are fairly similar in size. In Milwaukee county there are 4,388 block groups with the smallest having a population of 389 and the largest being 6,889 with a mean population of 1,222. In Milwaukee county there are 6278 precincts with an average of nearly 1700 voters. Of course, balance in larger units like block groups should improve with diminishing returns as the distance to the city limit gets smaller. For example, moving from 300 to 200 meters will do little to improve balance at the block group level since blocks are about 200 meters and block groups be definition are composed of several blocks.

We do balance assessment six different ways. First, we simply compute balance across the entire treated and control populations. This provides a basic baseline balance assessment for comparisons. We, next, assess balance with two methods that are what we call spatially naive. That is they are based on concepts of the score that do not measure spatial distance to the discontinuity. One spatially naive method would be to simply use the buffers or distance intervals from the city limit. That is we have defined intervals of 50, 100, 200, 300, 400, 500, 750, and 100 meters on either side of the Milwaukee city limit. We calculate balance on the covariates within each of these buffers. We also use one other spatially naive method. As we mentioned earlier, one can calculate a simple nearest distance to the city limit. We use this naive distance with a simple matching algorithm. Matching on the naive distance is a slightly more refined version of using the buffers which are a form of simple exact matching within buffers.

We contrast these naive methods with methods that use a two dimensional score that accurately reflects spatial distance. First, we repeat the matching exercise on distance but replace the naive distance metric with the chordal distance. The chordal distance measures not only distance to the boundary but also along the city limit. Next, we repeat the matching on chordal distance with the buffers. That is we restrict the sample to those within the buffer intervals and then match on the chordal distance. Note we use the buffers not as strict intervals but to limit the sample to those at least that far from the city limit. For

Table 1: Balance Covariates

Covariate	Measurement Level
% Hispanic 18 yrs or older	Block
% Black 18 yrs or older	Block
Median Age	Block
Median Household Income	Block group
% College Graduates	Block group
% Foreign Born	Block group
% Owns Home	Block group
% Below Poverty Level	Block group
% Unemployed	Block group
% Urban	Block group
Median Home Price	Block group
% High School Graduates	Block group
Median Rent Asked	Block group
Median Rooms	Block group
Median Rent Asked	Block group
% Owner Occupied	Block group
President Democratic Vote Share 2004	Precinct
U.S. House Democratic Vote Share 2004	Precinct
U.S. House Democratic Vote Share 2006	Precinct
U.S. Senate Democratic Voter Share 2006	Precinct
Aggregate Turnout 2004	Precinct
Aggregate Turnout 2006	Precinct
Male	Individual
Age	Individual
Turnout 2000	Individual
Turnout 2002	Individual
Turnout 2004	Individual
Turnout 2006	Individual

example, the 300 meter band includes that are no more than 300 meters from the city limit not people that are between 200 and 300 meters from the city limit. Finally, we examine the quality of balance by increasing locality in a second dimension. Clearly by using smaller buffers, the area of estimation becomes more local. But even with the smallest buffer of 50 meters, we are using the entire city limit. We might suspect that some parts of the city limit are more comparable to some suburbs than others. For example, cursory examination of Census data indicates that the two western suburbs of West Allis and Wauwatosa are more comparable to Milwaukee than Shorewood and Glendale in the Northeast. Here, we use the city limit intervals to compare various parts of the city limit to the suburbs. We divided the city limit into ten different equal intervals. We then repeated the chordal matching with buffers within these smaller intervals around the city limit.

Importantly, we apply matching not because we are invoking the selection on observables assumption but simply as a method for assessing the quality of the continuity assumption in this GD design. The key difference between what we do and a more standard matching analysis is that we do not match on *any* of the 27 observed covariates that we think might be unbalanced due to sorting around the city limit. That is we want to see if balance improves as a function of distance to the discontinuity. Therefore, we only match on geographic distance measures in various forms. We use a simple matching method that is held constant. That is anytime we match we rely on nearest neighbor matching with ties broken randomly. We also match with replacement. While a balance analysis is useful for understanding whether the continuity assumption holds, it doesn't provide us with an obvious estimation method short of relying on matched differences. We next propose an estimation method that is faithful to the spatial nature of the GD design.

5.2.2 Local Polynomial Estimation

While standard methods of balance are perfectly suitable for trying to assess how well the continuity assumption holds, they do not provide us with a coherent strategy for estimating treatment effects. Moreover, the estimation of such effects is complicated by the fact

that in the GD design continuity doesn't identify a point estimate but instead an infinite-dimensional curve of treatment effects. We develop a nonparametric kernel density estimator that we can use to both assess balance and estimate treatment effects. Our estimator being nonparametric is relatively flexible but also allows us to remain faithful to the spatial aspects of the GD design. We use this kernel density estimator to both estimate treatment effects and assess the continuity assumption.

Kernel density smoothing is a standard form of nonparametric regression estimation that uses weighted moving average smoothing to estimate a nonparametric conditional expectation. In a standard kernel density estimator, a weighted mean is calculated within bins of the data. Data close to the middle of the bin, the focal point, are weighted more heavily than observations farther from the focal point. This requires a measure of distance from the focal point. The standard measure of distance is

$$z_i = \frac{(x_i - x_0)}{h}. \tag{1}$$

The term z_i measures the scaled and signed distance between the x-value for the ith observation and the focal point: x_0. The scale factor, h, is called the *bandwidth* of the kernel estimator, and it controls the bin width. Of course, h controls how smooth or rough the nonparametric estimate will be. A weighting or kernel function $K(\cdot)$ is applied to the signed and scaled distances, which attaches the greatest weight to the observations that are close to x_0, with weights decreasing symmetrically and smoothly as the value of $|z|$ grows. This produces the set of weights $w_i = K[z_i]$. In our application, we use the Gaussian or normal kernel, which is simply the normal density function applied to the values of z_i. The weights, w_i, are then used to calculate to the local weighted average:

$$\hat{f}(y|x) = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{2}$$

A standard generalization of kernel density estimation that reduces its poor boundary

behavior is *local polynomial regression*, which estimates the regression function by fitting a polynomial at the focal value x_0 weighting observations with the kernel weights w_i. When the polynomial is of degree one, the procedure is called *local linear regression*. We adopt a multivariate local linear regression estimation method to provide adequate estimates of the treatment effect curve identified by the GD design. First, the two-dimensional focal point becomes the latitude and longitude of a point on the discontinuity boundary and thus has two elements: $\boldsymbol{x}_0 = (x_{00}, x_{01})$. We define x_1 as a vector of latitudes and x_2 as a vector of longitudes. We calculate spatial-based kernel weights as:

$$w_i = K[(x_1 - x_{00})/h] \times K[(x_2 - x_{01})/h] \tag{3}$$

where is K remains the Normal density. We then regress the outcome Y on X, a matrix with a constant and focal-point-deviated latitudes and longitudes using weighted least squares (WLS) with w_i as the weights. The predicted value of Y from this WLS regression using only observations in the treatment area serves as a point estimate for the treated regression function at $\boldsymbol{x}_0$. Similarly, the predicted value of Y using only observations in the control area serves as a point estimate for the control regression function at $\boldsymbol{x}_0$.

We repeat this for each point of latitude and longitude on the city limit. From the two separate applications of the estimator to the treated and control populations, we define two vectors: $\hat{Y}_T$ and $\hat{Y}_C$. These vectors are indexed by the number of points of latitude and longitude along the city limit. For example, if one spaces the points at 100-meter intervals there are 1752 points, and if one uses 250-meter intervals there are 557 points . We can use the difference of $\hat{Y}_T$ and $\hat{Y}_C$ in two ways. If we use a predetermined characteristic such as pre-treatment turnout as the outcome in the WLS regression, we can assess the continuity assumption in the GD design. More usefully, we can identify areas along the discontinuity where the assumption is most likely to hold. That is, we can identify areas of the city border where pre-treatment characteristics do not differ. Using this method to assess a

large number of pre-treatment characteristics is probably infeasible. The advantage of the balance approach is a large number of covariates can be analyzed simultaneously. The kernel density method would require 27 iterations for each covariate. It is best applied to a few key covariates, like past turnout in our study. However, if we use the outcome of interest, turnout in 2008, our estimator provides a spatial estimate of the treatment effect of interest. Since our estimator iterates over a set of point along the discontinuity, it provides treatment effect estimates that vary spatially and approximates the treatment effect curve identified in the GD design.

The nonparametric nature of our estimator requires dealing with a few additional complications. One is choice of the bandwidth parameter. Here, we rely on cross-validation for bandwidth selection. Using the diagonal elements of $H = X(X'X)^{-1}X'$, h_i, for each ith observation from the WLS regression, we can calculate Δ_{cv}, the cross-validation prediction error as

$$\frac{1}{n}\sum \frac{(y_i - \hat{y}_i)^2}{(1-h_i)^2}. \tag{4}$$

Finally there is the matter of inference. The analytical standard errors from the WLS regression are invalid since the weights are non-parametrically estimated. A natural alternative is the bootstrap. Since the bootstrapped statistic is a weighted regression model, this statistic is smooth in the mean and thus bootstrap theory should hold. We are currently working on appropriate bootstrap algorithms that respect the spatial quality of the data, and a theoretical derivation of a spatially weighted asymptotic variance-covariance matrix. Since these calculations are currently in progress, the results section focuses on point estimation and not yet on statistical inference.

6 Results

Here we focus on how balance changes as a function of the distance to our discontinuity boundary, the Milwaukee city limit. In every analysis, we checked balance on the 27 different covariates in Table 1. Of these measures, we focus particularly on two: individual level

turnout in 2004 and 2006. We do this for two reasons. First, these measures represent placebo outcomes, since turnout rates in 2004 and 2006 occurred *before* the 2008 initiative was on the ballot in Milwaukee. Covariates like education are simply proxies for these measures of turnout. Second, these covariates are among the few individual level covariates in our study. As such we should be able more precisely understand how balance changes as a function of distance as opposed to measures at the block group level which comprise relatively larger geographic areas. There are a wide variety of ways in which we could report the level of balance. We focus on two basic measures. The first is the median difference in a quantile-quantile (QQ plot). We report this in two ways. First, we report the average median QQ difference across all 27 variables that we check for balance. Second, we report the average median QQ difference for just the two individual level turnout measures. While this median difference doesn't have an interpretable scale, we always compare it to the baseline from the unadjusted data. Recall that for every analysis, we hold the matching method constant by using nearest neighbor matching. The question of interest is whether balance improves as the distance between the two-dimensional score and the geographic discontinuity decreases.

We start with the unmatched data and match on a naive distance. That is, this measure only considers how far voters are to the city limit, but not their location along that border. The results are in Table 2. The numbers from the unmatched data provide our baseline. What is immediately clear is that turnout is much lower within the city of Milwaukee as compared to the suburban areas within the county. In 2006, the difference in turnout was nearly 15 percentage points. Given the disparities in education and income between the city and suburbs, this difference is not surprising. Matching on the naive distance measure does improve balance, but the differences are still substantial, as turnout in 2006 is still lower by almost eleven percentage points. Of course, as we demonstrated, this uni-dimensional measure does not identify the estimates from the GD design.

Next, we examine balance among groups of voters in bands or buffers of increasing

Table 2: Balance Results Based on Unmatched Data and Naive Geographical Distance

	Unmatched	Matched Naive Distance
Average QQ Median Difference	0.168	0.124
Turnout 2004	-7.9	-5.9
Turnout 2006	-14.5	-10.9
Average QQ Median Difference Placebo Outcomes	0.112	0.042

Note: [a]Cell entry is treated minus control difference in turnout.

distance from the city limit. That is, we start with all voters that live within 1000 meters of the city limit. We then narrow this buffer to widths of 500, 400, 300, 300, 100 and 50 meters. Using all voters within 1000 meters of the city limits produces results that are better than the unmatched data and comparable to the use of the naive distance measure. For example, the ratio for the median QQ difference on the two turnout measures is nearly one-to-one when comparing the largest buffer to the naive distance. Smaller buffers, however, improve balance considerably. We see the balance on the turnout variables is best for the 200 meter buffer. The difference in turnout percentages is now just under two and five points for 2004 and 2006 respectively, though overall balance is little better than that based on the naive distance. Most existing designs based on a geographic discontinuity use bands of this type. For example, Black (1999) uses a buffer of two-tenths of a mile which is slightly larger than 300 meters. Of course, as we have shown, such design need to recover the treatment effect even if continuity holds in two dimensions. That said, even this naive method substantially improves balance on the turnout measures, which suggests that even a naive consideration of distance may be useful in some applications.

In Table 4 we account for spatial distance to the discontinuity. First, we do this by simply matching on the spatial distance instead of the naive distance. This allows us to account for distance both from and along the Milwaukee city limit. Next, we combine the spatial distance with the buffers zones along the city limit. For buffers of 1000 and 500

Table 3: Balance Results Based on Decreasing Zones Around the Milwaukee City Limit

	1000m Buffer	500m Buffer	400m Buffer	300m Buffer	200m Buffer	100m Buffer	50m Buffer
Average QQ Median Difference	0.157	0.156	0.153	0.144	0.122	0.109	0.09
Turnout 2004[a]	-5.1	-4.6	-4.4	-4.8	-1.8	-3.5	-5.4
Turnout 2006[a]	-8.1	-7.5	-7	-7.1	-4.5	-5.2	-6.1
Average QQ Median Difference Placebo Outcomes	0.034	0.031	0.029	0.030	0.016	0.022	0.029

Note: [a]Cell entry is treated minus control difference in turnout.

meters overall balance is much improved compared to buffers without the spatial distance. For 1000 meters, the average median QQ distance is 0.058 compared to 0.156. Thus, balance is better by nearly a factor of three. Once we move to a buffer of 400 meters, balance on the turnout variables improves dramatically. For a buffer of 300 meters, the difference in turnout is one and less than three percentage points. Note that overall balance does not improve much as the buffers shrink. This is not surprising given that it is impossible to make balance better in the block group level measures for smaller distances. The balance in this analysis is impressive. The reader should keep in mind that we have not matched on *any* of the 27 measured covariates, and yet we see impressive improvements in balance. Thus we might have some confidence that the geographic RD design is identified. Or at the very least we might conclude that considering spatial distance might add something beyond only using observable covariates.

Often, using a more local estimator increase the interval validity of an design. In fact, the RD design is built on this very concept. In the geographic RD, we can make the estimates more local in two ways. One method is by decreasing the width of the buffer around the border. Clearly a 100 meter buffer will produce a more local estimate than a 500 meter buffer. As we observed in Table 4, the more local design produced a better counterfactual. Within the geographic RD design, we can increase locality in another way. Thus far, we have used the entire length of the Milwaukee city limit. But we might suspect that by only

Table 4: Matched Balance Results

	Chordal Distance	1000m Buffer	500m Buffer	400m Buffer
Average QQ Median Difference	0.050	0.058	0.086	0.073
Turnout 2004[a]	8.5	0.9	2.4	2.5
Turnout 2006[a]	26.6	13.5	9.9	1.9
Average QQ Median Difference Placebo Outcomes	0.088	0.036	0.032	0.011
	300m Buffer	200m Buffer	100m Buffer	50m Buffer
Average QQ Median Difference	0.065	0.067	0.071	0.067
Turnout 2004[a]	2.7	3.4	1.3	-2.1
Turnout 2006[a]	1	4	4.5	1.7
Average QQ Median Difference Placebo Outcomes	0.009	0.019	0.014	0.01

Note: [a]Cell entry is treated minus control difference in turnout.

using parts of the boundary balance might improve. For example, along parts of the western part of the Milwaukee city limit are older suburbs that appear more comparable than the more affluent areas along the northeastern part of the city limit. We use a basic method of subclassification to make the estimates more local in the second spatial dimension. We divided the city limit into ten subclasses of equal distance. Dividing the city limit up into ten subclasses makes for sub-boundaries of length just over 17.5 kilometers. In Table 5, we repeat the balance analysis used on the entire metropolitan boundary for one of these zones. The results are comparable to those to based on the entire boundary. We might expect a smaller distance along the boundary to bring about further improvement. For example, we might instead use twenty subclasses.

Next, we look at results based on the local linear regression estimator. As mentioned above, this approach involves estimating the regression function at different points on the discontinuity boundary to produce a treatment effect curve. We used this estimation procedure to estimate the treatment effect curve for 2004 and 2006 turnout shares, two crucial covariates since they are pre-treatment realizations of our outcome of interest (2008 turnout

Table 5: Zone 6 Balance Results

	Chordal Distance	1000m Buffer	500m Buffer	400m Buffer
Average QQ Median Difference	0.061	0.059	0.076	0.079
Turnout 2004[a]	8.6	8.3	3.8	2.4
Turnout 2006[a]	4.6	12.4	7.8	3.8
Average QQ Median Difference Placebo Outcomes	0.038	0.052	0.029	0.016
	300m Buffer	200m Buffer	100m Buffer	50m Buffer
Average QQ Median Difference	0.080	0.089	0.081	0.099
Turnout 2004[a]	1.5	1.9	6.4	14.1
Turnout 2006[a]	1.7	-1.6	3.2	5.5
Average QQ Median Difference Placebo Outcomes	0.009	0.009	0.024	0.049

Note: [a]Cell entry is treated minus control difference in turnout percentage.

shares). We report the results for 2006 turnout; similar results for 2004 turnout rates are available upon request. We estimated the treatment-control difference in 2006 turnout, $\hat{Y}_T - \hat{Y}_C$, at 557 different points on the Milwaukee city boundary. These points were obtained by applying 250-meter intervals along the city boundary, excluding the fractions of the boundary that overlap with the Milwaukee County boundary and consequently do not have appropriate close control observations. We estimated all differences within the 300-meter buffer.

Figure 3 reports a histogram of the absolute values of the estimated 557 treatment-control differences. Since turnout is expressed in shares, the absolute value of these differences ranges from 0 to 1. As can be seen, there is wide variation across city boundary points. While some points have treatment-control differences of less than 1 percentage point, others have differences as large as 80 or 90 percentage points. Since this is a pre-treatment covariate, *if the geographic RD assumptions held at all points on the boundary*, one would expect all these differences to vanish and become statistically indistinguishable from zero. Since our standard errors are still not available, we use the unadjusted treatment-control difference as a

baseline. This is the simple (absolute value of the) difference in means in 2006 turnout rates between treatment and control areas within the 300-meter buffer, which is equal to 0.083. As shown in the histogram, 46% of the points have treatment-control differences that are smaller than the unadjusted difference. This suggests two important conclusions. First, explicitly incorporating the spatial local of individuals can increase balance across treatment and control areas even within a small buffer around the boundary, and thus provide observable evidence about the plausibility of the continuity assumptions necessary for the identification of the treatment effect curve of interest. Second, it suggests that this approach will detect any heterogeneity in the plausibility of these assumptions across different points in the boundary, since points where balance in crucial pre-treatment covariates is never achieved by means of a spatially weighted local linear regression are evidence against these assumptions. In other words, this method, by estimating treatment effects for a large number of points along the discontinuity boundary, provides a way to detect areas where the geographic RD may be more or less likely to hold. Both suggestions stem from the same principle: if indeed identification of the effects of interest is coming from the geographic discontinuity, balance measures should improve when counterfactual groups are obtained by explicitly incorporating a two-dimensional measure of distance.

Since each treatment-control difference is attached to a point on the city boundary, we can effectively use this geographic information to place the estimated treatment effects on the map. This is done in Figure 4, where a map of Milwaukee County displays the estimated treatment-control differences for the 557 points on the city boundary where we estimated these effects. Black points are those where the estimated absolute value of the treatment-control difference is less than the unadjusted treatment-control difference in the 300-meter buffer; hollow points are those where this estimated difference is less than the unadjusted treatment-control difference. As can be seen, by plotting the treatment effect curve on the map, we are able to locate in space the points along the boundary where balance improves when a two-dimensional notion of space is incorporated in the estimation of treatment-control

differences.

A current limitation of this map is that it now reports point estimates, which are interesting on their own right but do not answer the question of whether the null hypothesis of no difference between treatment and control areas can be rejected at every point. When standard errors are available, we will replicate the map in Figure 4, but distinguishing points where the null hypothesis cannot be rejected from points where the treatment-control differences in this important covariate are statistically significant. This will give us a better idea of the portions along the boundary where the geographic RD assumptions are more likely to hold. Nonetheless, the current map illustrates the general idea behind our approach.

7 Conclusion

The design-based approach holds at its core that unless analysts have an experiment, natural or randomized, an observable selection process, a discontinuity or some other strong research design it is difficult to make a compelling case that an estimated correlation is causal. Among these designs, use of the regression discontinuity design has grown rapidly and is often viewed as more compelling than other quasi-experimental designs. Lee and Lemieux (2010) argue that the reason RD design is compelling is that, like an experiment, it is a design and not an estimation method. For almost any research study, a regression model of some type can be devised where the outcome is a function of treatment status and other covariates. In an RD design, either a discontinuity exists or it does not. Moreover, the design predicts that pre-treatment covariates should not change at the discontinuity, which provides a clear testable implication of the key assumption. Of course, true discontinuities are somewhat rare. Geographic boundaries would seem to be one promising avenue given that treatments of interest often change sharply at relevant geographic boundaries.

The difficulty is that while geographic discontinuities are relatively common, this design is particularly vulnerable to a violation of the continuity assumption. Lee's (2008) behavioral interpretation of the continuity assumption brings to sharp focus the key weakness of the GD

design. Quite often agents are able to sort very precisely around the boundary that forms the discontinuity in the design. New York City provides an insightful example. No one will mistake an apartment in Manhattan for one in the Bronx. Even the difference between Manhattan south of Houston street is a salient division and few are likely to mistake being north of Houston for being south of Houston. Thus, understanding whether a specific GD design is identified requires substantive knowledge and careful evaluation of how observables behave as distance to the boundary decreases.

This is not to say that the GD design does not hold considerable promise. In our application, we find that balance improves considerably near the Milwaukee city limit. We demonstrate that balance on key covariates improves as a function of the spatial distance to the city limit. We found that even naive approaches to distance improved balance. This suggests that in this application the Milwaukee city limit is worth exploiting as a discontinuity instead of simply relying on the ubiquitous selection on observables assumption. Again New York city provides a relevant example of the promise of the GD design. Fernandez (2011) details that the line between Queens and Brooklyn is one where few residents really know its exact location which makes sorting between the two boroughs difficult, but this boundary does create important administrative differences. In short, while GD designs may be vulnerable to violation of the continuity assumption, geographic discontinuities may also prove to be strong designs. Again this decision can only be made on a case-by-case basis. There is little hope that the GD design can be mass produced as it requires careful attention to not only the statistical analysis needed to justify the continuity assumption but the geographic analysis needed to exploit the discontinuity.

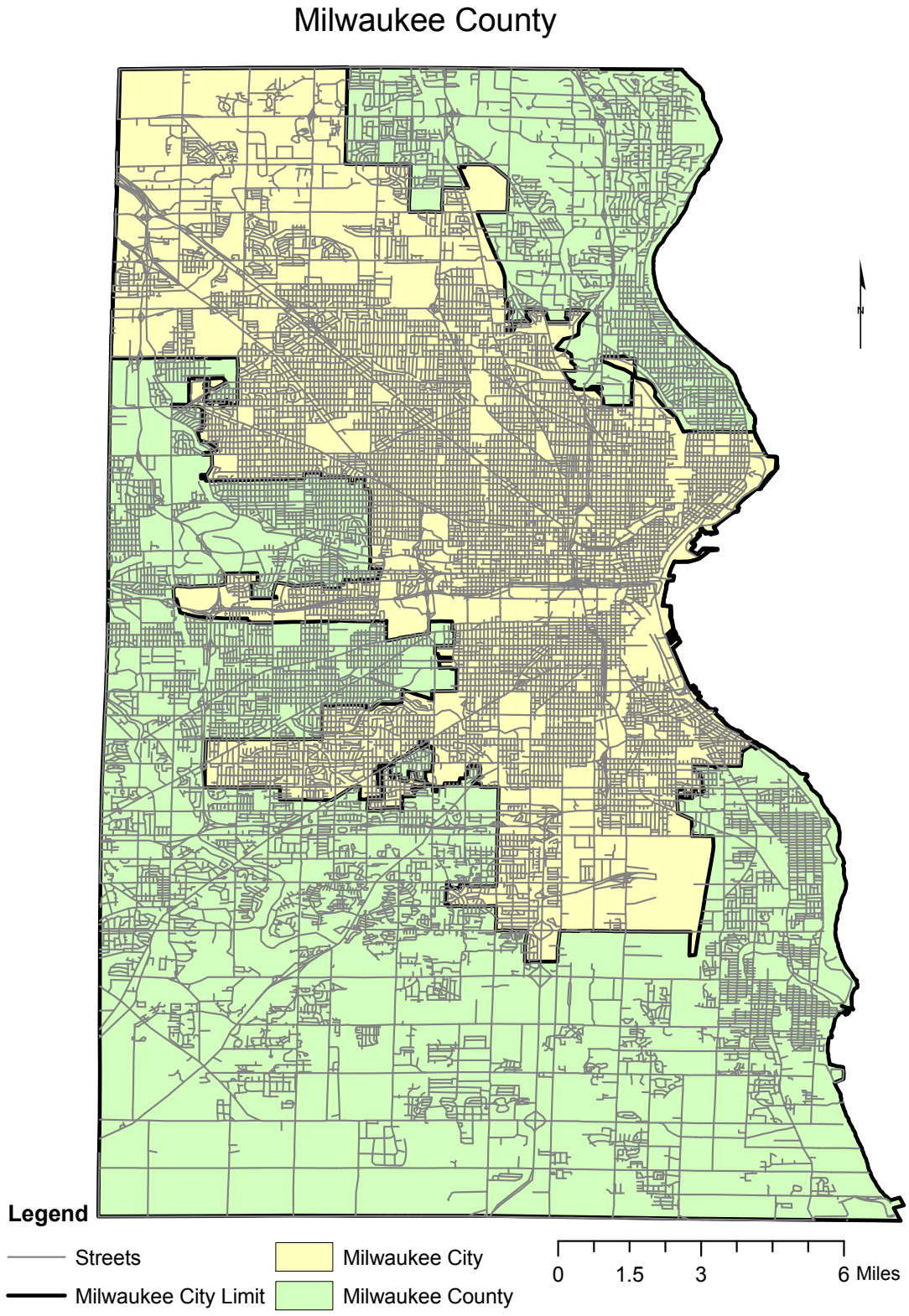


Figure 2: Milwaukee County with Geographic Discontinuity Based on Milwaukee City Limit Highlighted

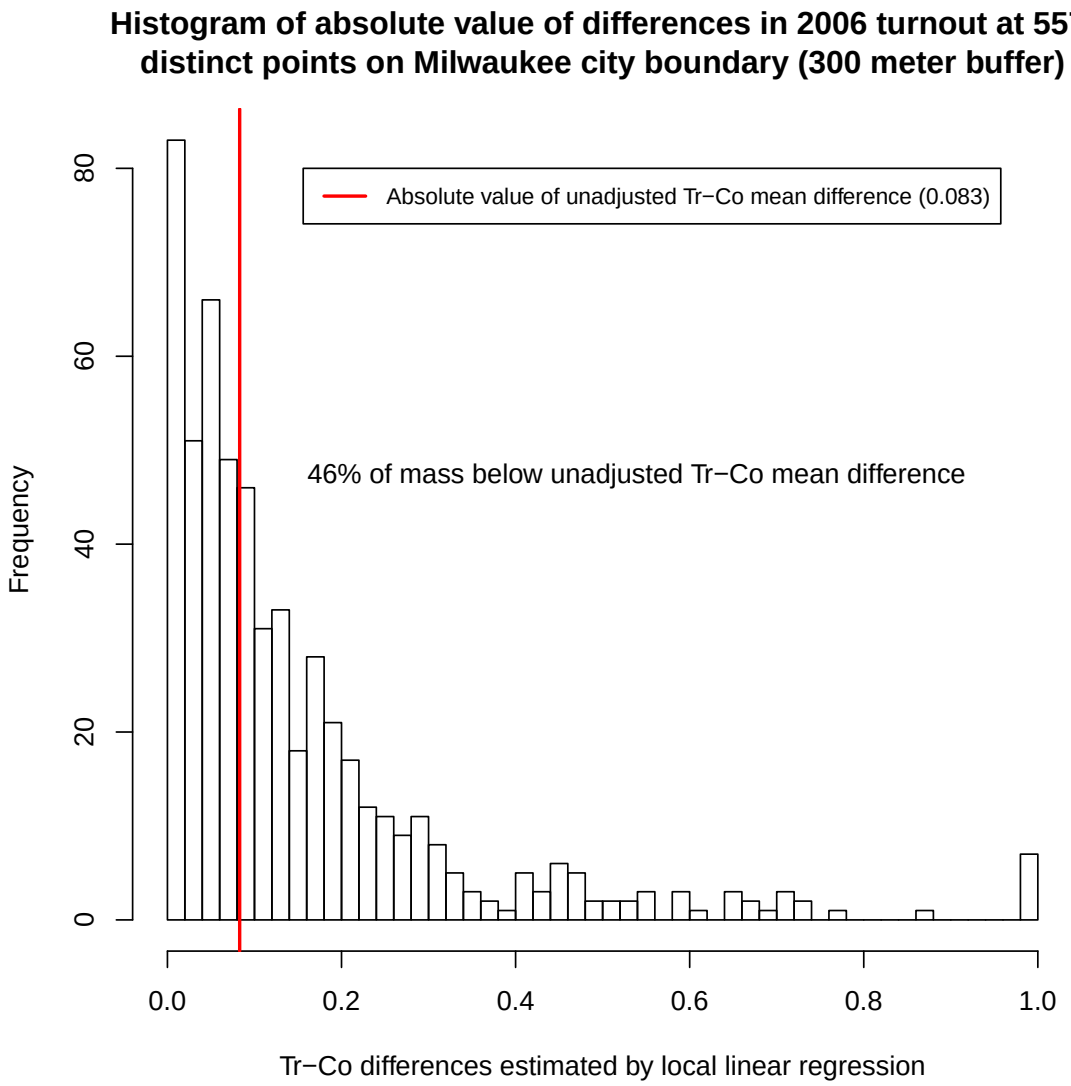
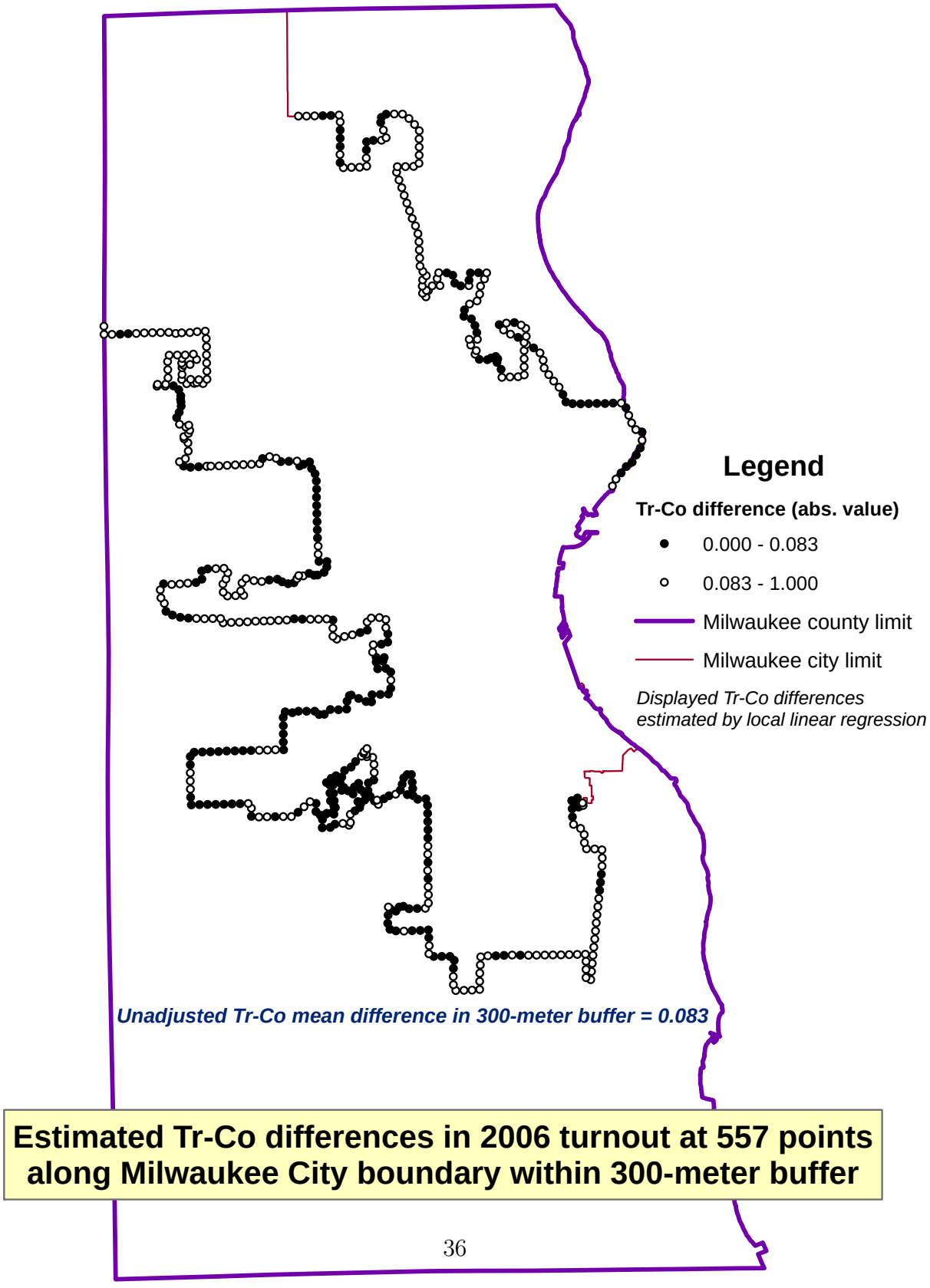


Figure 3: Estimated treatment-control differences in 2006 turnout

Figure 4: Estimated Tr-Co differences in 2006 turnout along Milwaukee City boundary



References

Banerjee, Sudipto. 2005. "On Geodetic Distance Computations in Spatial Modeling." *Biometrics* 61 (2): 617–625.

Bayer, Patrick, Fernando Ferreira, and Robert McMillan. 2007. "A Unified Framework for Measuring Preferences for Schools and Neighborhoods." *Journal of Political Economy* 115 (4): 588–638.

Berger, Daniel. 2009. "Taxes, Institutions and Local Governance: Evidence from a Natural Experiment in Colonial Nigeria.". Unpublished Manuscript.

Black, Sandra E. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education." *The Quarterly Journal of Economics* 114 (2): 577–599.

Brady, Henry E. and John E. McNulty. 2011. "Turning Out To Vote: The Costs of Finding and Getting to the Polling Place." *American Political Science Reviewm* Forthcoming.

Broockman, David E. 2009. "Do Congressional Candidates Have Reverse Coattails? Evidence from a Regression Discontinuity Design." *Political Analysis* 17 (4): 418–434.

Butler, Daniel M. 2009. "A Regression Discontinuity Design Analysis of the Incumbency Advantage and Tenure in the U.S. House." *Electoral Studies* 28 (1): 123–128.

Butler, Matthew J. and Daniel M. Butler. 2006. "Splitting The Difference? Causal Inference and Theories of Split-Party Delegations." *Political Analysis* 14 (4): 439–455.

Cook, Thomas D., William R. Shadish, and Vivian C. Wong. 2008. "Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons." *Journal of Policy Analysis and Management* 27 (4): 724–750.

Downs, Anthony. 1957. *An Economic Theory of Democracy.*. New York: Harper and Row.

Eggers, Andrew C. and Jens Hainmueller. 2009. "MPs for Sale? Returns to Office in Postwar British Politics." *American Political Science Review* 103 (4): 513–533.

Fernandez, Manny. 2011. "Of Queens and Kings.". New York Times.
URL `http://www.nytimes.com/interactive/2010/12/12/nyregion/20101212-border-info-gallery.html`

Gerber, Alan S., Daniel P. Kessler, and Marc Meredith. 2011. "The Persuasive Effects of Direct Mail: A Regression Discontinuity Based Approach." *Journal of Politics* Forthcoming.

Gerber, Elisabeth R., Arthur Lupia, Mathew D. McCubbins, and D. Roderick Kiewiet. 2001. *Stealing the Initiative: How State Government Responds to Direct Democracy*. Upper Saddle River, NJ: Prentice-Hall.

Gertner, Jon. 2006. "What Is a Living Wage?". New York Times.

Gomez, Brad T., Thomas G. Hansford, and George A. Krause. 2007. "The Republicans Should Pray for Rain: Weather Turnout, and Voting in U.S. Presidential Elections." *Journal of Politics* 69 (3): 649–663.

Green, Donald P., Terence Y. Leong, Holger Kern, Alan S. Gerber, and Christopher W. Larimer. 2009. "Testing the Accuracy of Regression Discontinuity Analysis Using Experimental Benchmarks." *Political Analysis* 17 (4): 400–417.

Hahn, Jinyong, Petra Todd, and Wilbert van der Klaauw. 2001. "Identification and Estimation of Treatments Effects with a Regression-Discontinuity Design." *Econometrica* 69 (1): 201–209.

Haspel, Moshe and H. Gibbs Knotts. 2005. "Location, Location, Location: Precinct Placement and the Costs of Voting." *Journal of Politics* 67 (2): 560–573.

Hopkins, Daniel J. and Elisabeth R. Gerber. 2009. "When Mayors Matter: Estimating the Impact of Mayoral Partisanship on City Policy.". Unpublished Manuscript.

Keele, Luke J. and William Minozzi. N.d. "How Much is Minnesota Like Wisconsin? States as Counterfactuals.". Unpublished Manuscript.

Krasno, Jonathan S. and Donald P. Green. 2008. "Do Televised Presidential Ads Increase Voter Turnout? Evidence from a Natural Experiment." *Journal of Politics* 70 (1): 245–261.

Lavy, Victor. 2006. "From Forced Busing to Free Choice in Public Schools: Quasi-Experimental Evidence of Individual and General Effects.". National Bureau of Economic Research Working Paper 11969.

Lee, David S. 2008. "Randomized Experiments From Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142 (2): 675–697.

Lee, David S. and Thomas Lemieux. 2010. "Regression Discontiuity Designs in Economics." *Journal of Economic Literature* 48 (2): 281–355.

Lupia, Arthur and John G. Matsusaka. 2004. "Direct Democracy: New Approaches to Old Questions." *Annual Review of Political Science* 7: 463–82.

Matsusaka, John G. 2004. *For The Many Or The Few: The Initiative, Public Policy, and American Democracy*. Chicago, IL: Chicago University Press.

Miguel, Edward. 2004. "Tribe or Nation? Nation Building and Public Goods in Kenya Versus Tanzania." *World Politics* 56 (3): 327–362.

Posner, Daniel N. 2004. "The Political Salience of Cultural Difference: Why Chewas and Tumbukas Are Allies in Zambia and Adversaries in Malawi." *The American Political Science Review* 98 (4): 529–545.

Riker, William H. and Peter C. Ordeshook. 1968. "A Theory of the Calculus of Voting." *American Political Science Review* 62 (1): 25–42.

Smith, Daniel A. and Caroline J. Tolbert. 2004. *Educated By Initiative: The Effects of Direct Democracy On Citizens And Political Organizations In The American States*. Ann Arbor, MI: University of Michigan Press.

Tolbert, Caroline J., John A. Grummel, and Daniel A. Smith. 2001. "The Effects of Ballot Initiatives on Voter Turnout In The American States." *American Politics Research* 29 (6): 625–648.

Tolbert, Caroline J. and Daniel A. Smith. 2005. "The Educative Effects of Ballot Initiatives on Voter Turnout." *American Politics Research* 33 (2): 283–309.

Whiteley, Paul F. 1995. "Rational Choice and Political Participation-Evaluating the Debate." *Political Research Quarterly* 48 (1): 211–233.