

# Enhancing a Geographic Regression Discontinuity Design Through Matching to Estimate the Effect of Ballot Initiatives on Voter Turnout\*

Luke Keele<sup>†</sup>      Rocío Titiunik<sup>‡</sup>      José Zubizarreta<sup>§</sup>

November 21, 2013

## Abstract

Ballot initiatives allow the public to vote directly on public policy. A literature in political science has attempted to document whether the presence of an initiative can increase voter turnout. We study this question for an initiative that appeared on the ballot in 2008 in Milwaukee, WI, using a natural experiment based on geography. This form of natural experiment exploits variation in geography where units in one geographic area receive a treatment while units in another area do not. When assignment to treatment via geographic location creates as-if random variation in treatment assignment, adjustment for baseline covariates is unnecessary. In many applications, however, some adjustment for baseline covariates may be necessary. As such, analysts may wish to combine identification strategies—using both spatial proximity and covariates. We propose a matching framework to flexibly incorporate information about both geographic proximity and observed covariates which allows us to minimize spatial distance while preserving balance on observed covariates. This framework is also applicable to regression discontinuity designs not based on geography. We find that the initiative on the ballot in Milwaukee does not appear to have increased turnout.

**Keywords:** matching, regression discontinuity design, ballot initiative

---

\*For comments and suggestions, we thank the Joint Editor Arnaud Chevalier, the two anonymous reviewers, Matias Cattaneo, Don Green, Paul Poast, Marc Ratkovic, Paul Rosenbaum, Marshall Joffe, and Dylan Small.

<sup>†</sup>Associate Professor, Department of Political Science, 211 Pond Lab, Penn State University, University Park, PA 16802 Phone: 814-863-1592, Email: ljk20@psu.edu.

<sup>‡</sup>Assistant Professor, Department of Political Science, University of Michigan, 5700 Haven Hall, 505 South State Street, Ann Arbor, MI 48109-1045, Email: titiunik@umich.edu.

<sup>§</sup>Assistant Professor, Division of Decision, Risk and Operations, Graduate School of Business, Columbia University, 3022 Broadway, 417 Uris Hall, New York, NY 10027-6902, Email: zubizarreta@columbia.edu.

# 1 Introduction

In twenty-four of the states in the U.S., citizens can place legislative statutes directly on the ballot for passage by the electorate. In the political science literature, these ballot initiatives are believed to increase voter turnout by stimulating voters' interest in the election. Early work, however, found little evidence that initiatives increased turnout (Everson 1981; Magleby 1984). Although later research did find a positive correlation between ballot initiatives and turnout (Tolbert et al. 2001; Smith and Tolbert 2004), some stipulated that the effect was conditional on the type of election (Daniel and Yohai 2008). All these studies, however, relied on comparisons between states with and without the initiatives process, and are therefore subject to confounding from state-level factors such as election administration laws and political culture. We revisit this question by studying the turnout effects of a municipal-level ballot initiative in Milwaukee, Wisconsin, an intervention that is assigned based on geography and allows us to avoid cross-state comparisons. Treatments that vary with geography are especially common in federal systems where subnational government units such as states, counties or municipalities often have considerable latitude in the adoption of specific policies. Understanding the effects of such treatments often must rely on observational studies since experimentation may be infeasible. Any research design intended to make inferences about the effects of geographically varying treatments must compare units in the treated area to units in a control area.

One possible research design states that conditional on a set of observed pretreatment covariates, treated and control areas are comparable, and uses statistical methods such as matching or regression to adjust for these measured covariates (Keele and Titiunik 2013a). The risk with this design is that unmeasured confounders may bias the treatment effect estimate. An alternative research design exploits geographic proximity. If units sort around a boundary between treated and control areas with error or the boundary between treated and control areas is drawn arbitrarily, a local treatment effect is identifiable under a regression

discontinuity framework (Keele and Titiunik 2013b). Under this design, treated and control groups near the boundary are good counterfactuals for each other because placement in the treated or control areas can be thought to be “as if random” very near the boundary. We explore how analysts might blend these two designs and base inferences on observations that are (i) in a small neighborhood around the geographic boundary that separates treatment and control areas and (ii) still require adjustment for pretreatment covariates. In applications where the assumptions behind this combined strategy are plausible, researchers can use it to obtain estimates of the treatment effect of interest in a neighborhood around the boundary. We implement this design using matching, an intuitive and flexible form of statistical adjustment that can easily accommodate our combined design.

## 2 The 2008 Initiative in Milwaukee

We study the effect of ballot initiatives on voter turnout through an analysis of the initiative process in Milwaukee, WI, where the city has the initiative process but the state of Wisconsin does not. In 2008, a coalition of local labor, educational and community organizations led by the National Association of Working Women, helped place an initiative on the ballot that mandated all private employers in the city of Milwaukee provide one hour of sick leave for every 30 hours worked. The initiative passed receiving slightly more than 68% of the vote, but it was struck down by the courts shortly after the election. On the county wide ballot, citizens also voted on a sales tax increase which also passed.

Figure 1 contains a map of Milwaukee county. The area in yellow comprises the city of Milwaukee which is surrounded by 17 suburban areas that are considered Minor Civil Divisions—the equivalent of a municipality—by the U.S. Census Bureau. The paid sick leave initiative appeared on the ballot in the city of Milwaukee (area in yellow), but was not on the ballot in any of the surrounding suburbs (area in blue).

Basic comparisons of Milwaukee to the suburban municipalities that surround it demonstrate that the city is more ethnically diverse, has lower housing prices, and lower socio-

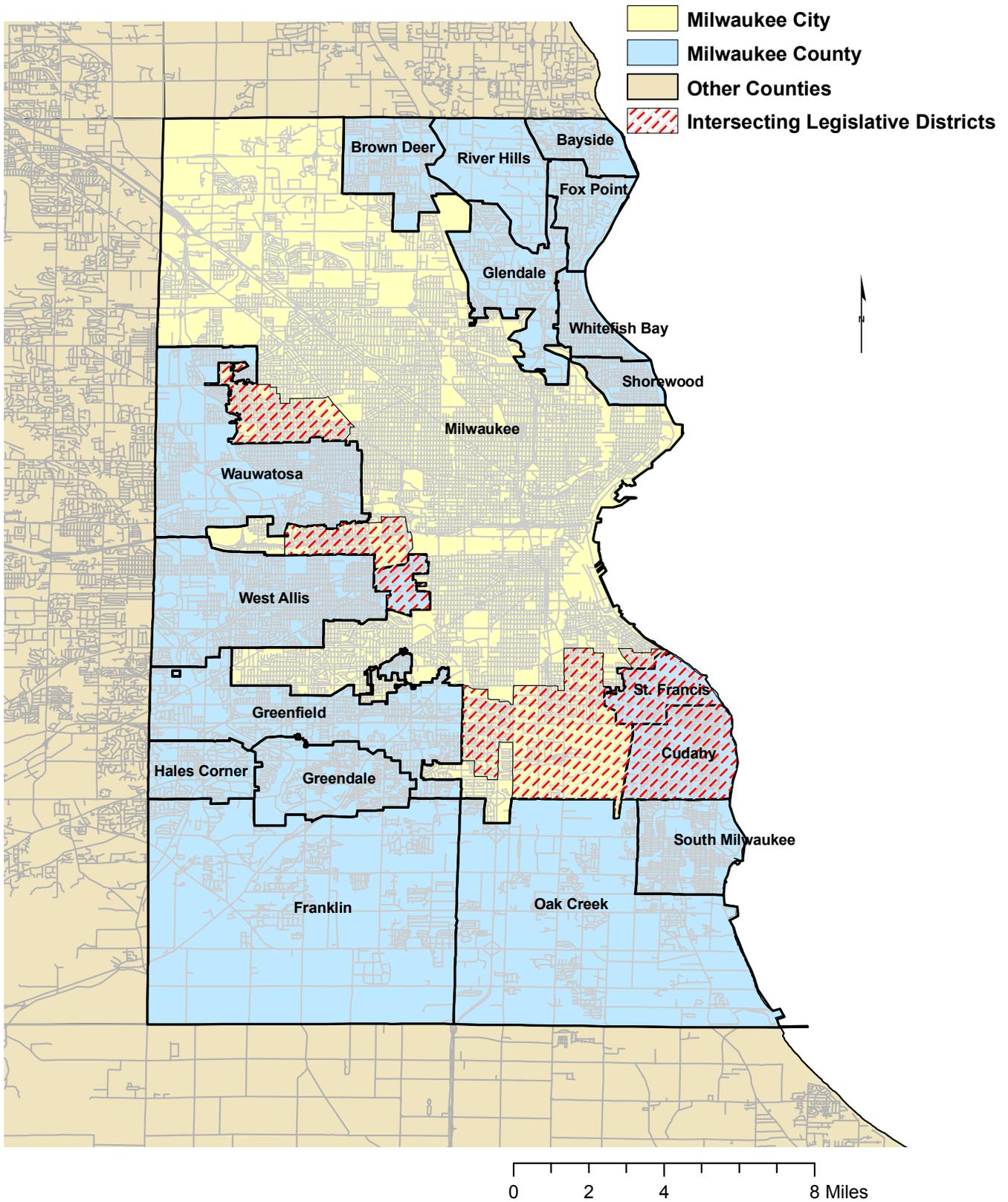


Figure 1: The Milwaukee metro area within Milwaukee county. The ballot initiative of interest was placed on the ballot within the city of Milwaukee highlighted in yellow. Suburbs to the city highlighted in green did not have the initiative on the ballot. Areas outside Milwaukee County did not have any initiatives on the ballot.

economic status. Using census data from 2000, we observe that median household income in Milwaukee is just under \$34,000, while it is nearly \$54,000 in the suburbs. The percentage of African-American residents of voting age in Milwaukee is 29%, while it is less than 1.5% in the suburbs. The difference in median housing value is nearly \$60,000. Nearly 21% of residents in the suburbs have a college degree while just over 12% in the city do. Moreover, examination of State Assembly, State Senate and U.S. House district maps in Milwaukee county reveals that most of the Milwaukee city limit is used as a boundary between legislative districts. There are only two areas where the same State Assembly, State Senate and U.S. House districts contain both treated and control voters from the city and suburbs. This suggests that state legislators use the city limit to separate voters. In Figure 1, we highlight the areas where legislative districts straddle the city limit using dashed areas on the map.

## **2.1 Data: Covariates, Housing Values, and Distances**

Our main data source is the Wisconsin Voter File, the database of registered voters maintained by the state of Wisconsin for administrative purposes. This file contains date of birth, gender, voting history, voters addresses, and the legislative districts in which each voter's address is included. We also acquired county records for the nearly 30,000 houses sold in Milwaukee county from 2006 to 2008, including house characteristics and sales price.

To determine distances between voters, we converted each voter's address into latitude and longitude coordinates, which we then used to calculate the spatial distance between voters' residences. We also calculated for each voter the median value of houses sold within a 500m radius of her residence; we use this as a voter-level measure of housing values. Finally, we recorded each voter as either residing in the treated area—Milwaukee—, or residing in the control area—one of the Milwaukee suburbs—based on their address in the voter file. We now turn to our framework for causal inference.

### 3 Statistical Framework and Designs

We adopt the potential outcomes framework, and assume that unit  $j$  has two potential outcomes,  $Y_{Tj}$  and  $Y_{Cj}$ , which correspond to levels of treatment  $Z_j = 1$  and  $Z_j = 0$ , respectively. In our geographic context, we compare units in a *treated area* to units in a *control area*, which we denote by  $A_T$  and  $A_C$ , respectively; thus,  $Z_j = 1$  when  $j$  is within  $A_T$  and  $Z_j = 0$  when  $j$  is within  $A_C$ . Units also have a vector of covariates  $\mathbf{X}_j$  determined before treatment is assigned. The triplet of observed random variables is  $(Y_j, Z_j, \mathbf{X}_j)$ . The individual-level treatment effect is  $Y_{Tj} - Y_{Cj}$ , the observed outcome is  $Y_j = Z_j Y_{Tj} + (1 - Z_j) Y_{Cj}$ , and the fundamental problem of causal inference is that we cannot observe both  $Y_{Tj}$  and  $Y_{Cj}$  simultaneously for any given unit (Neyman 1990; Rubin 1974; Holland 1986). Nonetheless, different identifiable estimands based on aggregates or averages can be defined; in Section 5, we focus on the attributable effect. We assume the Stable Unit Treatment Value Assumption (SUTVA) throughout, that is, that the potential outcomes of one unit do not depend on the treatment status of other units (Cox 1958; Rubin 1986).

In our application, the most plausible form of interference would occur if voters in the treated area encouraged their neighbors in control areas to vote due to the enthusiasm induced by the ballot initiative. However, a recent experiment on voter turnout found little evidence of treatment spillovers even within households (Sinclair et al. 2012), so we suspect that there will be little interference across voters even when the households are adjacent. Moreover, a SUTVA violation of this kind would tend to bias the effect towards zero, so any positive effects should be conservative estimates.

#### 3.1 Design 1: Conditioning on Observables

The first design we consider is based on the assumption that treatment is ignorable conditional on a set of observed covariates:

**Assumption 1** (Conditional Geographic Treatment Ignorability). *The potential outcomes are independent of treatment assignment conditional on observed covariates  $\mathbf{X}_j$ . That is,*

$$(Y_{Tj}, Y_{Cj}) \perp\!\!\!\perp Z_j \mid \mathbf{X}_j.$$

We assume that  $\mathbf{X}_j$  does not include any measures of distance to the boundary between  $A_C$  and  $A_T$ . If Assumption 1 holds, statistical inferences can be obtained in a straightforward way after adjusting for observed differences. There is, however, no *a priori* reason to suspect that the geographic variation of treatment will justify this assumption, and we suspect that unobserved confounders will contribute to why someone might live in the city of Milwaukee as opposed to one of its immediate suburbs.

### 3.2 Design 2: The Geographic Regression Discontinuity Design

An alternative design uses adjacent areas and exploits spatial proximity to the border between  $A_C$  and  $A_T$ , under the assumption that units either select locations around this boundary with error or the boundary is drawn in a haphazard fashion, unrelated to the units' characteristics. In this design, every unit  $j$  can be thought of as having a score or running variable  $\mathbf{S}_j = (S_{j1}, S_{j2})$  that uniquely defines its geographic location and allows us to compute its distance to any point  $(b_1, b_2)$  on the boundary. Assignment of treatment  $Z_j$  is then a deterministic function of this score, which has a discontinuity at the known boundary between  $A_C$  and  $A_T$ . This is a form of the regression discontinuity (RD) design, in which units are assigned to treatment or control groups based on whether their value of an observed covariate exceeds a known cutoff. The RD design was first introduced by Thistlethwaite and Campbell (1960), and the seminal paper by Hahn et al. (2001) provided identification results. See Imbens and Lemieux (2008) and Lee and Lemieux (2010) for reviews, and Calonico et al. (2013) for recent results in robust nonparametric inference of RD effects. Hereafter, we refer to this geographic RD design as the geographic regression discontinuity (GRD) design following Keele and Titiunik (2013b), where this design is discussed in detail. Identification of the average treatment effect at the boundary in the GRD design is based on the following assumption:

**Assumption 2** (Continuity in two-dimensional score). *The conditional regression functions*

are continuous in  $(S_{j1}, S_{j2})$  at all points  $(b_1, b_2)$  on the boundary:

$$\lim_{(z_1, z_2) \in A_C \rightarrow (b_1, b_2)} E \{Y_{Cj} | (S_{j1}, S_{j2}) = (z_1, z_2)\} = E \{Y_{Cj} | (S_{j1}, S_{j2}) = (b_1, b_2)\}$$

$$\lim_{(z_1, z_2) \in A_T \rightarrow (b_1, b_2)} E \{Y_{Tj} | (S_{j1}, S_{j2}) = (z_1, z_2)\} = E \{Y_{Tj} | (S_{j1}, S_{j2}) = (b_1, b_2)\},$$

for all points  $(b_1, b_2)$  on the boundary.

Identification under Assumption 2 requires that people cannot precisely self-select locations around the boundary in a way that makes potential outcomes discontinuous. The validity of this assumption may be threatened since people often select their place of residence based on administrative boundaries. For example, the quality of schools may vary sharply from one school district to the next, and many people use this information when choosing where to buy a house. In Milwaukee, there is no place along the border where the boundaries of school districts do not perfectly coincide with the Milwaukee city limit—that is, every suburb has its own school district. If residence in a given school district affected voter turnout, this would threaten identification. However, if we believed that school district is correlated with voter turnout only due to its correlation with income, and we believed that income varies smoothly at the boundary, then identification of the treatment effect of interest would still be possible.

### 3.3 Design 3: Conditioning on Observables and The Discontinuity

When there appears to be strong self-selection around the border of interest, one alternative is to combine designs and assume that, after conditioning on covariates, treatment assignment is as-if randomized for those who live near the city limit. Thus, we propose a combined design, where geographic distance between treated and control observations is minimized while balance in pretreatment covariates is also enforced. To formalize this idea, we let  $N$  collect the set of coordinates corresponding to a small geographic neighborhood around each point on the boundary between  $A_T$  and  $A_C$ . We assume that there exists a small neighborhood  $N$  where potential outcomes and treatment assignment are conditionally independent

given pre-determined covariates:

**Assumption 3** (Conditional Geographic Treatment Ignorability in Local Neighborhood).

*For each point on the boundary, there exists a neighborhood  $N$  such that  $(Y_{Tj}, Y_{Cj}) \perp\!\!\!\perp Z_j \mid \mathbf{X}_j$  for all  $j$  with  $(S_{j1}, S_{j2}) \in N$ .*

This assumption is in the spirit of an interpretation developed by Lee (2008), who argued that RD designs can be seen as local experiments near the cutoff where treatment status changes. A formalization of this local randomization idea within a randomization inference framework for the standard RD design was proposed by Cattaneo et al. (2013). The combination of designs that we propose generalizes and expands these ideas to consider a small geographic neighborhood around the boundary that separates the treatment and control areas, where an as-if randomization or independence condition holds after conditioning on pre-determined covariates—but possibly does not hold unconditionally as in the standard interpretation.

This design takes from a GRD design based on Assumption 2 the notion that the treated and control groups will be valid counterfactuals as the score approaches the cutoff and from a selection on observables assumption the notion that valid counterfactuals can only be obtained after conditioning pretreatment covariates. Note that Assumption 3 is weaker than Assumption 1 (because it requires conditional independence only for a subset of the population), but it is not necessarily weaker than Assumption 2. Note that while Assumption 2 is concerned with identification of a treatment effect only at each boundary point, Assumption 3 would allow for identification of the treatment effect not only at these boundary points, but also at all internal points included in the geographic neighborhood where it holds. Therefore, Design 3 will not be always preferable to Design 2. But when the assumptions behind Design 3 hold, making inferences based on this design will allow for estimation of the treatment effect for a (small) geographic area around the boundary that separates treated and control areas, as opposed to only at this boundary. Assumption 3 might be plausible when treatment assignment is based on geography, as subjects are typically strategic in choosing where

to reside but within small geographic areas may base their strategic decisions in observable quantities such as housing prices and neighborhood amenities.

Using our application, Figure 2 illustrates a situation in which Design 3 may be plausibly invoked by showing the treated-control mean differences in house prices as distance to the Milwaukee city limit decreases. Here, each treated voter is matched with the control voter who is geographically closest to her—but no pre-determined covariates are used to form the matches. As can be seen, even though the difference in house prices decreases approximately monotonically with geographic distance, it remains significantly different from zero even for the smallest distance of 50 meters. In this case, we believe that house prices, and more generally income, are highly correlated with turnout. Given that covariate balance improves as we near the city limit, but since those imbalances are not entirely removed, this provides a reason to invoke Design 3. In our analysis under Design 3 presented below, we invoke Assumption 3 including house prices in the conditioning set  $\mathbf{X}_j$ , which entails assuming that the unobservables (the variables not in  $\mathbf{X}_j$ ) follow a pattern similar to the one Figure 2, except that their differences do eventually vanish at the city limit.

## 4 A Matching Framework to Combine Designs

We propose a matching framework to implement Design 3 and combine the identification strategies based on observed covariates and geographic distances. While our application is characterized by a geographic discontinuity, this matching framework readily generalizes to standard (i.e., non-geographic) RD designs. To implement Design 3, one possibility would be to use standard matching methods and find close matches on some covariate distance while considering geographic proximity as an additional covariate. For this task, we could use a number of different matching algorithms, although with most types of matching it would be difficult to enforce different forms of balance on different covariates.

Matching via integer programming allows us to enforce different forms of balance for different covariates (Zubizarreta et al. 2013). In applications, there are typically key discrete

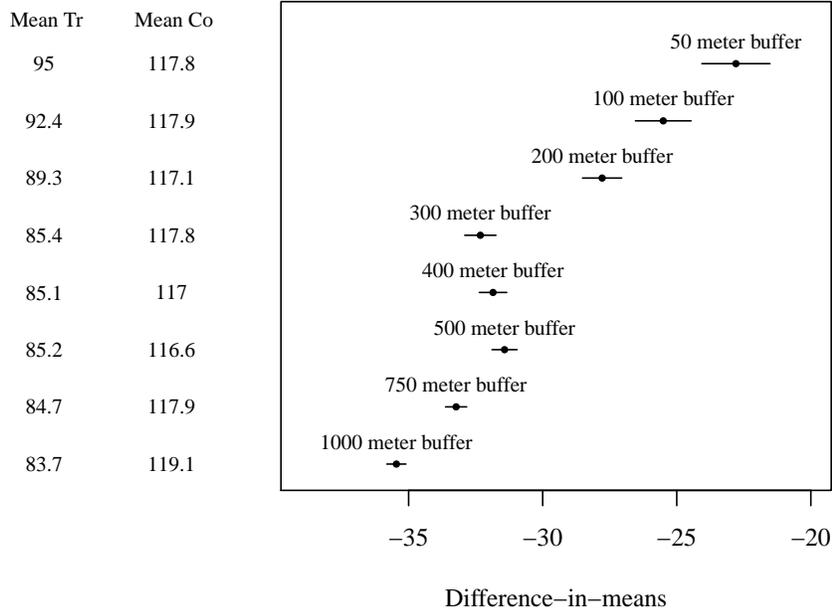


Figure 2: Difference-in-means in house prices at the individual level between treatment and control groups for different buffers around the Milwaukee city limit, matching on geographic distance within each buffer. A buffer is a narrow band around the border. A 100 meter buffer is a band that extends 100 meters from either side of the city limit. Units are in thousands of dollars. Dots are difference-in-means and bars are 95% confidence intervals based on paired t-tests.

covariates on which we may need to match exactly; other covariates of secondary importance on which we may wish to match with fine balance (i.e., balance their marginal distributions exactly in aggregate but without constraining who is matched to whom); and there are other covariates for which we may only want to minimize differences in means. See Rosenbaum et al. (2007) for a discussion of fine balance and Rosenbaum (2010, Part II) for a discussion of different forms of covariate balance. With integer programming, we can flexibly match subjects in the treated and control areas to minimize their relative geographic distances while also balancing their observed covariates with different forms of balance.

However, it may be the case that for a given treated unit there is no control unit that is both near in geographic distance and satisfies the balance constraints; this is known as a lack of common support. A caliper is one method that could be used ensure that common support holds. With a caliper, if a match cannot be made within some tolerance, the treated unit that cannot be matched is discarded. The difficulty with a caliper is that treated units

are not discarded in an optimal fashion. To deal with this problem, we apply the technique of optimal subset matching which optimally seeks to retain the largest number of treated subjects for which common support holds (Rosenbaum 2012). To that end, our matching framework implements optimal subset matching using integer programming to select the maximum number of matched pairs in relation to their total sum of distances that satisfies the balance constraints described above.

#### 4.1 Optimal Subset Matching with Integer Programming

Let  $j_t$  index the subjects in the treated area,  $A_T$ , and similarly let  $j_c$  index the subjects in  $A_C$ . Define  $d_{j_t, j_c}$  as the geographic distance between treated unit  $j_t$  and control  $j_c$ . To enforce specific forms of covariate balance, define  $e \in \mathcal{E}$  as the index of the covariates for which it is needed to match exactly, and  $b_e \in \mathcal{B}_e$  as the categories that covariate  $e$  takes, so that  $x_{j_t; e}$  is the value of nominal covariate  $e$  for treated unit  $j_t$  with  $x_{j_t; e} \in \mathcal{B}_e$ . Similarly, define  $f \in \mathcal{F}$  as the index of the nominal covariates for which it is required to match with fine balance, and  $b_f \in \mathcal{B}_f$  as the categories of covariate  $f$ , with  $x_{j_t; f}$ , the value of covariate  $f$  for treated unit  $j_t$ , and  $x_{j_c; f}$ , the value of covariate  $f$  for control unit  $j_c \in A_C$ . Finally, let  $m \in \mathcal{M}$  be the index of the covariates for which it is desired to balance their means, so that  $x_{j_t; m}$  is the value of covariate  $m$  for treated unit  $j_t$ , and  $x_{j_c; m}$  is the value of covariate  $m$  for control  $j_c$ .

To solve our problem optimally, we introduce binary decision variables

$$a_{j_t, j_c} = \begin{cases} 1 & \text{if treated unit } j_t \text{ is matched to control unit } j_c, \\ 0 & \text{otherwise,} \end{cases}$$

and, for a given scalar  $\lambda$ , we minimize

$$\sum_{j_t \in A_T} \sum_{j_c \in A_C} d_{j_t, j_c} a_{j_t, j_c} - \lambda \sum_{j_t \in A_T} \sum_{j_c \in A_C} a_{j_t, j_c} \quad (1)$$

subject to pair matching and covariate balancing constraints. Under this penalized match, if geographic distance can be minimized it will be, and if it cannot be minimized in every case, it will be minimized as often as possible. In particular, the pair matching constraints require each treated and control subject to be matched at most once,

$$\sum_{j_c \in A_C} a_{j_t, j_c} \leq 1, \quad \forall j_t \in A_T, \quad (2)$$

$$\sum_{j_t \in A_T} a_{j_t, j_c} \leq 1, \quad \forall j_c \in A_C. \quad (3)$$

This implies that we match without replacement, which we do to simplify inference. The covariate balancing constraints are defined as follows

$$\sum_{j_t \in A_T} \sum_{j_c \in A_C} \left| \mathbb{1}_{\{x_{j_t; e} = b_e\}} x_{j_t; e} - \mathbb{1}_{\{x_{j_c; e} = b_e\}} x_{j_c; e} \right| a_{j_t, j_c} = 0, \quad \forall e \in \mathcal{E}, \quad (4)$$

$$\sum_{j_t \in A_T} \sum_{j_c \in A_C} a_{j_t, j_c} \mathbb{1}_{\{x_{j_t; f} = b_f\}} - \sum_{j_t \in A_T} \sum_{j_c \in A_C} a_{j_t, j_c} \mathbb{1}_{\{x_{j_c; f} = b_f\}} = 0, \quad \forall b_f \in \mathcal{B}_f, f \in \mathcal{F}, \quad (5)$$

$$\left| \sum_{j_t \in A_T} \sum_{j_c \in A_C} a_{j_t, j_c} x_{j_t; m} - \sum_{j_t \in A_T} \sum_{j_c \in A_C} a_{j_t, j_c} x_{j_c; m} \right| \leq \varepsilon_m \sum_{j_t \in A_T} \sum_{j_c \in A_C} a_{j_t, j_c}, \quad \forall m \in \mathcal{M}, \quad (6)$$

where  $\mathbb{1}$  is the indicator function.

These constraints enforce exact matching, fine balance and mean balance, respectively. More precisely, (4) requires exact matching on the covariates  $e \in \mathcal{E}$  by matching each treated subject to a control with the same values for the covariates in  $\mathcal{E}$ ; (5) constrains the marginal distributions of the covariates in  $\mathcal{F}$  to be exactly balanced in aggregate, but without constraining who is matched to whom; and finally (6) forces the differences in means after matching to be less than or equal to the scalar  $\varepsilon_m$  for all  $m \in \mathcal{M}$ . See Zubizarreta (2012) for a discussion of these and other covariate balance constraints in the context of a more general mixed integer program. Generally, the covariates on which we wish to match exactly, with fine balance and mean balance, and the allowed discrepancy in means as represented by the scalar  $\varepsilon_m$ , should be chosen by the analyst based on substantive knowledge

of the problem at hand.

We incorporated optimal subset matching into the integer programming framework in the objective function (1) via the  $\lambda$  parameter. The first term in (1) is the total sum of geographic distances between matched pairs, and second term is the total number of matched pairs. Therefore,  $\lambda$  emphasizes the total number of matched pairs in relation to the total sum of distances and, according to (1), it is preferable to match additional pairs if on average they are at a smaller distance than  $\lambda$ . In our application, we choose  $\lambda$  to be equal to the median geographic distance between treated and control subjects so, according to (1), it is preferable to match additional pairs if on average they are at a smaller distance than the typical distance (as measured by the median). Subject to the pair matching constraints (2) and (3) and the covariate balancing constraints (4)–(6), this form of penalized optimization addresses the lack of common support problem in the distribution of observed covariates of the treated and control groups.

Including this penalty allows us to keep the largest number of matched pairs for which distance is minimized and the balance constraints are satisfied. This implies that as we alter the distances or the balance constraints, the number of treated and control subjects retained changes. In particular, for stricter constraints we tend to retain a smaller number of subjects. Although this is not ideal, discarding observations to deal with samples that have limited overlap is a common practice (Crump et al. 2009).

## 4.2 Three Matched Designs

We illustrate each of the designs discussed above with three different matching procedures. Throughout we use the R package `mipmatch` (Zubizarreta 2012). For all three matches, we first restrict our comparisons to treated and control voters that reside in the same legislative districts—the dashed areas in Figure 1. Given that state legislators often draw legislative districts based on the city limit, these districts are themselves important covariates. Therefore, we match exactly on the legislative districts: *Legislative District Exact Match I* includes

only treated and control voters in the 4th U.S. Congressional district, the 7th State Senate district, and the 20th State Assembly district, and *Legislative District Exact Match II* includes only treated and control voters in the 4th U.S. Congressional district, the 5th State Senate district, and the 15th State Assembly district. As we show in the Supplemental Appendix, exact matching on legislative districts significantly decreases covariate imbalance. Within each triplet of State Assembly, State Senate and U.S. House districts, we restricted all our analyses to observations within 750 meters of the Milwaukee city limit. As a practical matter, we also exactly matched on gender to ease the computation. Once we exactly match on legislative districts, gender is balanced, so differences in this covariate should have no effect on the matched estimates.

#### **4.2.1 Design 1: A Conventional Matching on Covariates**

The first match, based on Design 1, balances observable covariates, ignoring geographic distances. The covariate distances between units were obtained from a rank-based Mahalanobis distance matrix (see Rosenbaum 2010, §8.3). We also imposed a number of balance constraints. For each voter, we have binary indicators for whether they voted in 2004 and 2006, which we used to create a five-level categorical measure of voting history. See the appendix for details on how we constructed this measure. We match exactly on each of these categories since we suspect voting history is of critical importance. We constrained the means of age and housing value to differ by less than 1 year and \$1,000 dollars, respectively, between the treated and control areas. While the constraint on housing prices forces the mean differences to be similar, we also want the distribution of housing values across the treated and control groups to be similar. We therefore enforced a fine balance constraint on housing values so that house prices have the same distribution in treated and control groups without constraining how units are matched. We matched with fine balance for seven categories of housing prices to capture the somewhat long tail in the upper end of the distribution.

### 4.2.2 Design 2: Geographic Distance Matching

This second match hews as closely as possible to the GRD design where only geographic distance is needed for identification. Therefore, for the voters in the two overlapping legislative district triplets, we minimized the total sum of geographic distances between matched pairs. We also modified the distances  $d_{i,j_c}$  to penalize matching subjects residing at more than two kilometers of distance. The question is whether imbalances remain in observed covariates once geographic distance has been minimized in the matches.

### 4.2.3 Design 3: Combining Geographic Distance and Covariates

This last matching implements Design 3 using the integer programming optimal matching framework described above. The algorithm minimized geographic distances between matched pairs as in Design 2 while matching for sex, age, voting history, and housing values in the same manner as in Design 1.

## 4.3 Three Matched Comparisons

Table 1 shows housing prices and geographic distances for a design where we only exactly match on legislative districts and sex, and for the three matching designs, for both legislative district triplets. The table shows means and absolute standardized differences in means (difference in means divided by the pooled standard deviation between groups before matching) for housing prices, and average and median geographic distances between treated and matched control voters. When evaluating covariate balance, we focus on housing prices because they reflect important neighborhood characteristics such as quality of schools, safety, and household income. The importance of housing prices is discussed in detail in the hedonic pricing literature, where house values are used to infer the implicit prices of housing attributes and environmental characteristics (see Malpezzi (2002) for a review). The appendix contains additional balance test results.

In *Legislative District Exact Match I*, the median distance between treated and control observations in the unmatched data is a little more than one and a half kilometers. House

Table 1: Design Comparison for Covariate Balance in Three Matched Comparisons

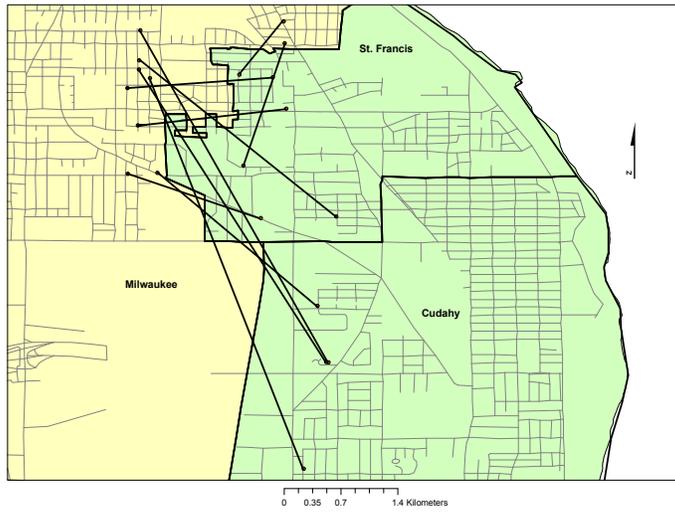
	House Value (\$)			Distance (km)		
	Mean Treated	Mean Control	Abs. Std. Diff.	Median	Mean	Pairs
	Legislative District Exact Match I					
Unmatched	167458	157663	0.44	3.72	3.54	–
Design 1 Covariates Only Match	156070	157051	0.04	2.87	3.28	2704
Design 2 Distance Only Match	164070	151135	0.56	0.88	1.04	2524
Design 3 Covariates and Distance Match	154259	153261	0.04	0.88	1.02	1939
	Legislative District Exact Match II					
Unmatched	158567	144692	0.69	6.58	5.78	–
Design 1 Covariates Only Match	144926	144692	0.01	7.72	5.80	1667
Design 2 Distance Only Match	136049	144802	0.43	1.87	1.68	1663
Design 3 Covariates and Distance Match	140725	141720	0.05	1.96	1.80	536

Note: For all designs, exact matching was done on sex, Congressional district, State Senate district, and State Assembly district, and only for observations within 750 meters from the border of each legislative district triplet. Design 1 additionally matches exactly on voting history; it also constrains the means of age and housing price to be less or equal than 1 year and \$1,000, respectively, and matches with fine balance for seven categories of housing price all the while minimizing the total sum of covariate distances based on a rank-based Mahalanobis distance within pairs. Design 2 minimizes the total sum of geographic distances between matched pairs. Design 3 minimizes the total sum of geographic distances between matched pairs while also matching on the same covariates as in Design 1. In *Legislative District Exact Match I*, all voters are in the 4th Congressional district, the 7th State Senate district, and the 20th State Assembly district. In *Legislative District Exact Match II*, all voters are in the 4th Congressional district, the 5th State Senate district, and the 13th State Assembly district. Abs. Std. Diff.= absolute standardized difference. Distance is from control voter’s resident to treated voter’s residence measured in kilometers. In the unmatched designs, *Pairs* shows the available number of pairs based on the total number of treated units; the original number of controls is 7396 in Legislative District Exact Match I and 9089 in Legislative District Exact Match II.

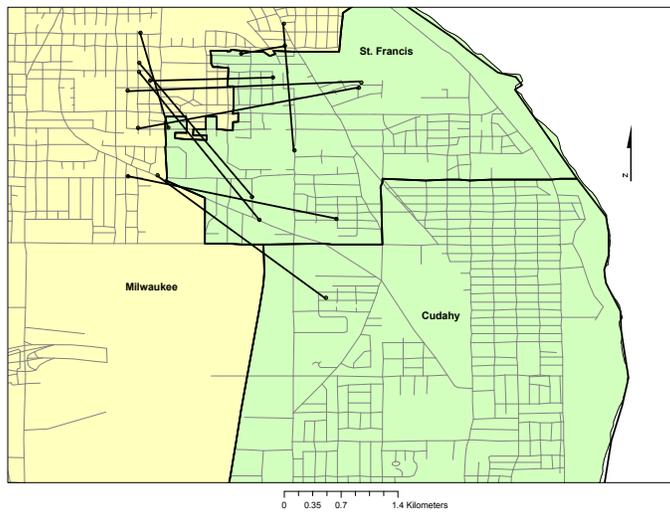
prices in Design 1 are very well balanced, with average house prices differing by just 500 dollars. This improved covariate balance, however, comes at the expense of distance. In Design 1, where covariate imbalance is minimized without regard to geographic distance, the median distance between matched pairs is nearly 3 kilometers, a full kilometer larger than in the unmatched data. In Design 2, which only minimizes geographic distance, the median geographic distance is reduced to 0.88 kilometers. But this improvement in geographic distance comes at the expense of covariate balance: balance on housing prices is now worse than in the unmatched data. For example, the mean difference in housing values is slightly less than \$10,000 in the unmatched data, but in Design 2 this difference increases to nearly \$13,000. As shown in Table 2 in the appendix, a similar pattern holds for age, where mean differences also increase in Design 2 relative to the unmatched data. Design 3, however, enforces both restrictions simultaneously. The standardized difference for housing value in Design 3 equals that in Design 2, while the median distance within matched pairs is identical to Design 2.

We see a very similar pattern in *Legislative District Exact Match II*, shown in the bottom panel of Table 1. Once again, Design 1 reduces imbalance in housing prices relative to the unmatched dataset (the standardized difference is reduced by 97%), but at the price of increasing the median distance between treated and controls from about 2 to 3.5 kilometers. While Design 2 decreases this median distance, it also increases the difference in house prices, with a standardized mean difference that is 430% larger than in Design 1. Again, Design 3 minimizes these differences while also restricting the comparison to observations that are geographically very close to each other; the difference in housing prices drops to less than \$1,000 while the median distance within matched pairs exceeds that in Design 2 by less than a tenth of a kilometer.

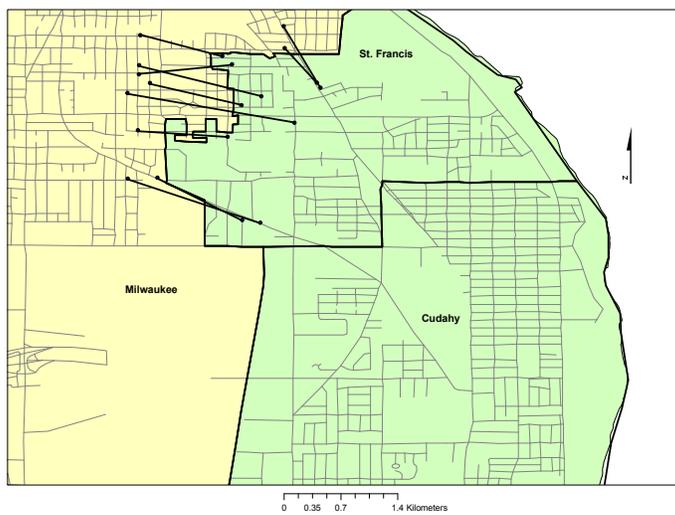
The matching results from the three different designs are illustrated in Figure 3, which shows ten matched pairs randomly chosen from each of the three designs in *Legislative District Exact Match I*. In the figures, the treated units are held fixed, and we show how the



(a) Design 1 - Covariates Only Match



(b) Design 2 - Distance Only Match



(c) Design 3 - Covariates and Distance Match

Figure 3: Ten Pairs of Matches Randomly Sampled from *Legislative District Exact Match I*

distance to the matched controls varies across the three designs. As seen in Figure 3(a), when geographic distance is not incorporated in the matching procedure, matched pairs in Design 1 are far away from each other. Incorporating geographic distance leads to matched pairs that are much closer together, as shown in Figure 3(b). When we match on both covariates and distance, the smaller distances remain as seen in Figure 3(c), but we also gain better balance in observables. Figure 3(c) embodies the strategy behind Assumption 3, which makes comparisons conditional on important covariates between units that are in a small geographic neighborhood of the boundary. The matching procedure implemented in Design 3 is able to reduce both distance and covariate imbalance by discarding observations from the analysis via the optimal subsetting. While Design 3 uses about 24% less observations than Design 1 in *Legislative District Exact Match I*, it uses about 62% less observations in *Legislative District Exact Match II*. As mentioned above, this loss of observations is expected given the stricter constraints imposed by Design 3 and is necessary to ensure common support holds.

## 5 Estimating the Effect of Ballot Initiatives on Turnout and Sensitivity to Unmeasured Confounders

### 5.1 Effects, Inference and Sensitivity Analysis with Binary Responses under Randomization Inference

We estimate the effect of the Wisconsin ballot initiative on turnout. We use a randomization inference framework, where potential outcomes are seen as fixed quantities and the only source of randomness is the assignment of treatment. The randomization-based framework requires interpreting Assumption 3 in terms of fixed quantities, as in Cattaneo et al. (2013). We outline this framework following Rosenbaum (2002b), and explain in detail how we conduct estimation and inference. In our analysis, there are  $I$  matched pairs,  $i = 1, \dots, I$ , with two subjects,  $j = 1, 2$ , one treated and one control for  $2I$  total subjects. Treatment assignment, potential outcomes, and observed outcomes are respectively  $Z_{ij}$ ,  $y_{Tij}$ ,  $y_{Cij}$ , and

$Y_{ij}$ —where we use lowercase to denote fixed variables. We write  $\boldsymbol{\delta}$  for the  $2I$ -dimensional vector of treatment effects,  $\boldsymbol{\delta} = (\delta_{11}, \dots, \delta_{I2})^T$ . In this case,  $\delta_{ij} \in \{-1, 0, 1\}$  for each  $i, j$  pair, and below we test Fisher’s sharp null hypothesis of no treatment effect on  $(y_{Tij}, y_{Cij})$  which stipulates that  $H_0 : y_{Tij} = y_{Cij}$  for all  $i, j$  and may be expressed as  $H_0 : \boldsymbol{\delta} = \mathbf{0}$ .

We test Fisher’s sharp null hypothesis using McNemar’s test, which is based on the number of discordant pairs in matched outcomes. In the case of matched pairs with binary responses, pair  $i$  is discordant if it contains exactly one person who voted,  $Y_{i1} + Y_{i2} = 1$ . McNemar’s statistic is the number of votes,  $T$ , among treated subjects in discordant pairs,  $T = \sum_{i \in D} \sum_{j=1}^2 Z_{ij} Y_{ij}$ , where  $D$  is a set of indices for the  $I^* \leq I$  discordant pairs. Some of the votes recorded in  $T$  may have been caused by the presence of the ballot initiative and others might have occurred whether there was an initiative on the ballot or not. The unobservable quantity  $T_c = \sum_{i,j} Z_{ij} y_{Cij}$  is the number of votes that would have occurred without an initiative on the ballot. Fisher’s sharp null hypothesis,  $H_0 : \boldsymbol{\delta} = \mathbf{0}$ , says that no votes were caused or prevented by the ballot initiative, implying that  $T = T_c$ . Therefore, this hypothesis may be tested by comparing  $T$  with the randomization distribution of  $T_c$ , which follows a binomial distribution with sample size  $I^*$  and probability of success  $1/2$ .

In an observational study, we can base a test of the sharp null hypothesis on the randomization distribution of  $T_c$  (Rosenbaum 2002a). In particular, if every unit  $j$  in pair  $i$  has the same probability of receiving treatment,  $\Pr(Z_{ij} = 1) = 1/2$ . This mode of treatment assignment would be true by construction in a pair randomized experiment since we would choose one unit at random from each pair to receive treatment. In our analysis, we assume this model of treatment assignment holds after conditioning on  $\mathbf{x}_j$ . One model for a sensitivity analysis of this assumption stipulates that  $1/(1 + \Gamma) \leq \Pr(Z_{ij} = 1 | \mathbf{x}_j) \leq \Gamma/(1 + \Gamma)$  for a specified value of  $\Gamma$  greater than one, such that randomization inference corresponds to  $\Gamma = 1$ ; see Rosenbaum (2002b, §4) for a discussion. We use values of  $\Gamma > 1$  to compute a range of possible inferences, which indicates the magnitude of bias due to an unobserved covariate that would need to be present to alter the conclusions reached when we assume

that random assignment of the treatment holds given the observed covariates.

To estimate an effect parameter and a one-sided confidence region, we use  $\boldsymbol{\delta}_0$ , which is a  $2I$ -dimensional vector with elements  $\delta_{0ij} \in \{-1, 0, 1\}$ . We consider hypotheses of the form  $H_0 : \boldsymbol{\delta} = \boldsymbol{\delta}_0$ . There are many hypotheses of the form  $H_0 : \boldsymbol{\delta} = \boldsymbol{\delta}_0$  that we could test, and it is generally not practical to test them all. We can summarize the testing of multiple hypotheses using the attributable effect, which is a scalar and unobserved quantity. The attributable effect  $\Delta = \sum_{i,j} Z_{ij}\delta_{ij}$  is the number of votes due to the ballot initiative, so that  $T - \Delta$  is the number of treated subjects who would have voted even in the absence of treatment (Rosenbaum 2002a).

If  $H_0 : \boldsymbol{\delta} = \boldsymbol{\delta}_0$  is true, we can define  $\Delta_0 = \sum_{i,j} Z_{ij}\delta_{0ij}$  to estimate an effect parameter for the ballot initiative. We use the method of Hodges and Lehmann (1963) to obtain a point estimate for  $\Delta$  by equating McNemar’s statistic to its null expectation. In the case of matched binary outcomes, we use a table of matched outcomes and adjust it until it is exactly without treatment effect. Specifically, the effect parameter is the value of  $\Delta_0$  such that the two off-diagonal cells of discordant pairs in the table of matched outcomes are equal to each other. The effect parameter represents the number of votes attributable to the treatment, which we express as the percentage of treated votes attributable to treatment.

We also calculate a one-sided confidence set for  $\boldsymbol{\delta}$  by testing every  $H_0 : \boldsymbol{\delta} = \boldsymbol{\delta}_0$  and retaining compatible hypotheses not rejected by the test (Rosenbaum 2002a). For example, if we reject the null hypothesis at the 5% level if  $\Delta_0 < a$  and accept if  $\Delta_0 \geq a$ , then a one-sided 95% confidence set for  $\boldsymbol{\delta}$  is the set of all  $\boldsymbol{\delta}_0$  compatible with  $\Delta_0 = \sum Z_{ij}\delta_{0ij} \geq a$ . Therefore, the 95% one-sided confidence set for  $\boldsymbol{\delta}$  is the set of all treatment effects with at least  $a$  responses among the subjects actually caused by the treatment.

## 5.2 Does Turnout Increase Due to a Ballot Initiative?

For the outcome analysis, we combined the data from the two different exact legislative matches into a single data set of matched pairs for each design. Table 2 contains cross-

Table 2: Voting Patterns in Designs 1, 2 and 3.

		Lived in Milwaukee Suburb		
		$Z_{ij} = 0$		
		Didn't Vote, $Y_{ij} = 0$	Voted, $Y_{ij} = 1$	
Design 1 <sup>a</sup>	Lived in Milwaukee	Didn't Vote, $Y_{ij} = 0$	212	690
	$Z_{ij} = 1$	Voted, $Y_{ij} = 1$	814	2655
Design 2 <sup>b</sup>	Lived in Milwaukee	Didn't Vote, $Y_{ij} = 0$	199	683
	$Z_{ij} = 1$	Voted, $Y_{ij} = 1$	782	2523
Design 3 <sup>c</sup>	Lived in Milwaukee	Didn't Vote, $Y_{ij} = 0$	118	401
	$Z_{ij} = 1$	Voted, $Y_{ij} = 1$	421	1535

<sup>a</sup> The one-sided  $p$ -value from McNemar's test is 0.008. The upper bound on the  $p$ -value is 0.046 for  $\Gamma = 1.08$ . An estimated 124 votes are attributable to treatment with an upper-bound of 193 votes.

<sup>b</sup> The one-sided  $p$ -value from McNemar's test is 0.005. The upper bound on the  $p$ -value is 0.035 for  $\Gamma = 1.04$ . An estimated 99 votes are attributable to treatment with an upper-bound of 166 votes.

<sup>c</sup> The one-sided  $p$ -value from McNemar's test is 0.254. An estimated 30 votes are attributable to treatment with an upper-bound of 70 votes.

tabulations of the matched pairs from each design. The table for Design 1, where the matches ignore distance, shows the counts of matched pairs—the number of discordant pairs are in the off-diagonal cells. For Design 1, there are  $I^* = 814 + 690 = 1504$  discordant pairs, and the one-sided  $p$ -value is calculated by comparing 814 votes among treated voters to a binomial distribution with 1504 trials and probability 1/2. If there is no hidden bias,  $\Gamma = 1$ , then the sharp null hypothesis of no treatment effect is implausible as the  $p$ -value from the test is 0.0075. However, the upper bound on the  $p$ -value is 0.046 for  $\Gamma = 1.08$  and 0.067 for  $\Gamma = 1.09$ , which indicates that even a weak confounder might alter our conclusions. The estimated effect is the value of  $\Delta_0$  which makes the McNemar's test statistic equal to its null distribution; this occurs when the numbers of discordant pairs in each of the two off-diagonal cells are equal to each other. Thus, the effect estimate is  $814 - 690 = 124$  or  $124/4371 \approx 2.8\%$  of the votes among the treated are attributable to the ballot initiative. To form a confidence interval, we test all hypotheses  $H_0: \delta = \delta_0$  and retain the set of values of  $\Delta_0$  not rejected at the 5% level. We find that  $\Delta_0 = 192$  attributable votes are accepted with one-sided

significance level 0.0519, while 193 votes are rejected with one-sided significant level 0.0494. Therefore with 95% confidence, we can say that no more than  $193/4371 \approx 4.4\%$  of the votes among the treated were due to the presence of the ballot initiative.

In Design 2, using McNemar’s test, the p-value of the one-sided sharp null test is 0.0052, so if  $\Gamma = 1$  the sharp null hypothesis of no treatment effect is implausible. The upper bound on the  $p$ -value is 0.035 for  $\Gamma = 1.04$  and 0.052 for  $\Gamma = 1.05$ , so the effect in Design 2 is even more sensitive to bias from a hidden confounder than in Design 1. In Design 2, the effect indicates that 99 or  $99/4187 \approx 3.0\%$  of the votes among the treated are attributable to the ballot initiative. In the absence of hidden bias, all hypotheses with  $H_0: \boldsymbol{\delta} = \boldsymbol{\delta}_0$  with  $\Delta_0 = 167$  attributable votes are accepted with significance level 0.0511. Thus with 95% confidence, no more than  $166/4187 \approx 3.0\%$  of the votes among the treated were due to the treatment. In sum, the conclusions we draw from Designs 1 and 2 are quite similar. Under both designs, we would conclude that approximately 3% of the among the treated were caused by the presence of an initiative on the ballot. For both of these designs, however, the sensitivity analysis indicates that these inferences could easily be reversed by a weak confounder.

In Design 3, the test of the sharp null hypothesis yields a p-value of 0.254, so if there is no hidden bias, it is plausible that the treatment did not cause any votes. In terms of the point estimate, 30 or  $30/2475 \approx 1.2\%$  of the votes among the treated are attributable to the ballot initiative, and with 95% confidence, no more than  $71/2475 \approx 2.9\%$  of the votes among the treated were due to treatment. Thus, based on Design 3, there is little evidence that ballot initiatives caused an increase in turnout.

## 6 Summary: Enhancing Regression Discontinuity Designs Through Matching

We use a penalized integer program to combine two identification strategies in a principled manner. Our approach allows us to find voters that are close geographically but that are also similar in terms of observable characteristics. Matching on just distance or observables alone

produced inferior matches in terms of observed balance. Thus we are able to produce more comparable matches while retaining as many matched pairs as possible in relation to their distances as regulated by  $\lambda$  subject to balance constraints. While our application focused on a geographic discontinuity, our method of matching could be applied to standard regression discontinuity designs. To our knowledge, this is the first application of matching methods to a discontinuity design.

We found that a ballot initiative did not increase turnout in the Milwaukee election, a result also found by Keele and Titiunik (2013a). Using randomization inference, we estimated that among the balanced subset of the data only 1.2% of the treated votes could be attributed to the presence of an initiative on the ballot, but the one-sided p-value of 0.254 indicated that this point estimate is consistent with a null effect. Our results are consistent with a previous finding in the literature that shows that ballot initiatives do not appear to have turnout effects in presidential elections. Thus, our analysis is consistent with the thesis that initiatives only increase turnout in midterm election years (Daniel and Yohai 2008).

## References

- Calonico, S., M. Cattaneo, and R. Titiunik (2013). Robust nonparametric confidence intervals for regression-discontinuity designs. Unpublished Manuscript.
- Cattaneo, M., B. Frandsen, and R. Titiunik (2013). Randomization inference in the regression-discontinuity design: An application to party advantages in the u.s. senate. Unpublished Manuscript.
- Cox, D. R. (1958). *Planning of Experiments*. New York: Wiley.
- Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96(1), 187–199.
- Daniel, S. and I. Yohai (2008). How initiatives don't always make citizens: Ballot initiatives in the american states, 1978-2004. *Political Behavior* 30(4), 469–489.
- Everson, D. (1981). The effects of initiatives on voter turnout: A comparative state analysis. *Western Political Quarterly* 34(3), 415–425.
- Hahn, J., P. Todd, and W. van der Klaauw (2001). Identification and estimation of treatments effects with a regression-discontinuity design. *Econometrica* 69(1), 201–209.
- Hodges, J. L. and E. Lehmann (1963, June). Estimates of location based on ranks. *The Annals of Mathematical Statistics* 34(2), 598–611.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* 81(396), 945–960.
- Imbens, G. W. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142(2), 615–635.
- Keele, L. J. and R. Titiunik (2013a). Natural experiments based on geography. Unpublished Manuscript.
- Keele, L. J. and R. Titiunik (2013b). Natural experiments based on geography. Unpublished Manuscript.
- Lee, D. S. (2008). Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics* 142(2), 675–697.
- Lee, D. S. and T. Lemieux (2010, June). Regression discontinuity designs in economics. *Journal of Economic Literature* 48(2), 281–355.
- Magleby, D. B. (1984). *Direct Legislation: Voting on Ballot Propositions in the United States*. Baltimore: Johns Hopkins University Press.
- Malpezzi, S. (2002). Hedonic pricing models and house price indexes: A select review. In K. Gibb and A. O'Sullivan (Eds.), *Housing Economics and Public Policy: Essays in Honour of Duncan MacLennan*, pp. 67–89. Oxford: Blackwell Publishing.

- Neyman, J. (1923 [1990]). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science* 5(4), 465–472. Trans. Dorota M. Dabrowska and Terence P. Speed.
- Rosenbaum, P. R. (2002a, March). Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association* 97(457), 1–10.
- Rosenbaum, P. R. (2002b). *Observational Studies* (2nd ed.). New York, NY: Springer.
- Rosenbaum, P. R. (2010). *Design of Observational Studies*. New York: Springer-Verlag.
- Rosenbaum, P. R. (2012, January). Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics* 21(1), 57–71.
- Rosenbaum, P. R., R. N. Ross, and J. H. Silber (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association* 102(477), 75–83.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 688–701.
- Rubin, D. B. (1986, December). Which ifs have causal answers. *Journal of the American Statistical Association* 81(396), 961–962.
- Sinclair, B., M. McConnell, and D. P. Green (2012, October). Detecting spillover in social networks: Design and analysis of multilevel experiments. *American Journal of Political Science* 56(4), 1055–1069.
- Smith, D. A. and C. 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.
- Thistlethwaite, D. L. and D. T. Campbell (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology* 51(6), 309–317.
- Tolbert, C. J., J. A. Grummel, and D. A. Smith (2001, November). The effects of ballot initiatives on voter turnout in the american states. *American Politics Research* 29(6), 625–648.
- Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association* 107(500), 1360–1371.
- Zubizarreta, J. R., D. S. Small, N. K. Goyal, S. A. Lorch, and P. R. Rosenbaum (2013). Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *Annals of Applied Statistics* 7(1), 25–50.