ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2015)

Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout

Luke Keele,

Penn State University, University Park, USA

Rocío Titiunik

University of Michigan, Ann Arbor, USA

and José R. Zubizarreta

Columbia University, New York, USA

[Received October 2012. Final revision November 2013]

Summary. Ballot initiatives allow the public to vote directly on public policy. The 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, Wisconsin, 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 whereas 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 incorporate information about both geographic proximity and observed covariates flexibly which allows us to minimize spatial distance while preserving balance on observed covariates. This framework is also applicable to regression discontinuity designs that are not based on geography. We find that the initiative on the ballot in Milwaukee does not appear to have increased turnout.

Keywords: Ballot initiative; Matching; Regression discontinuity design

1. Introduction

In 24 of the states in the USA, 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

Address for correspondence: Luke Keele, Department of Political Science, 211 Pond Lab, Penn State University, University Park, PA 16802, USA.
E-mail: ljk20@psu.edu

 0964–1998/15/178000

question by studying the turnout effects of a municipal level ballot initiative in Milwaukee, Wisconsin, an intervention that is assigned on the basis of 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 that is intended to make inferences about the effects of geographically varying treatments must compare units in the treated area with units in a control area.

One possible research design states that, conditionally on a set of observed pretreatment covariates, the entire 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 (RD) 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

(a) in a small neighbourhood around the geographic boundary that separates treatment and control areas and
(b) 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 neighbourhood around the boundary. We implement this design by using matching, an intuitive and flexible form of statistical adjustment that can easily accommodate our combined design.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

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, Wisconsin, where the city has the initiative process but the state of Wisconsin does not. In 2008, a coalition of local labour, educational and community organizations led by the National Association of Working Women helped to place an initiative on the ballot that mandated all private employers in the city of Milwaukee to provide 1 h of sick leave for every 30 h 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 countywide ballot, citizens also voted on a sales tax increase which also passed.

Fig. 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 US Census Bureau. The paid sick leave initiative appeared on the ballot in the city of Milwaukee (the area in yellow) but was not on the ballot in any of the surrounding suburbs (the area in blue).

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

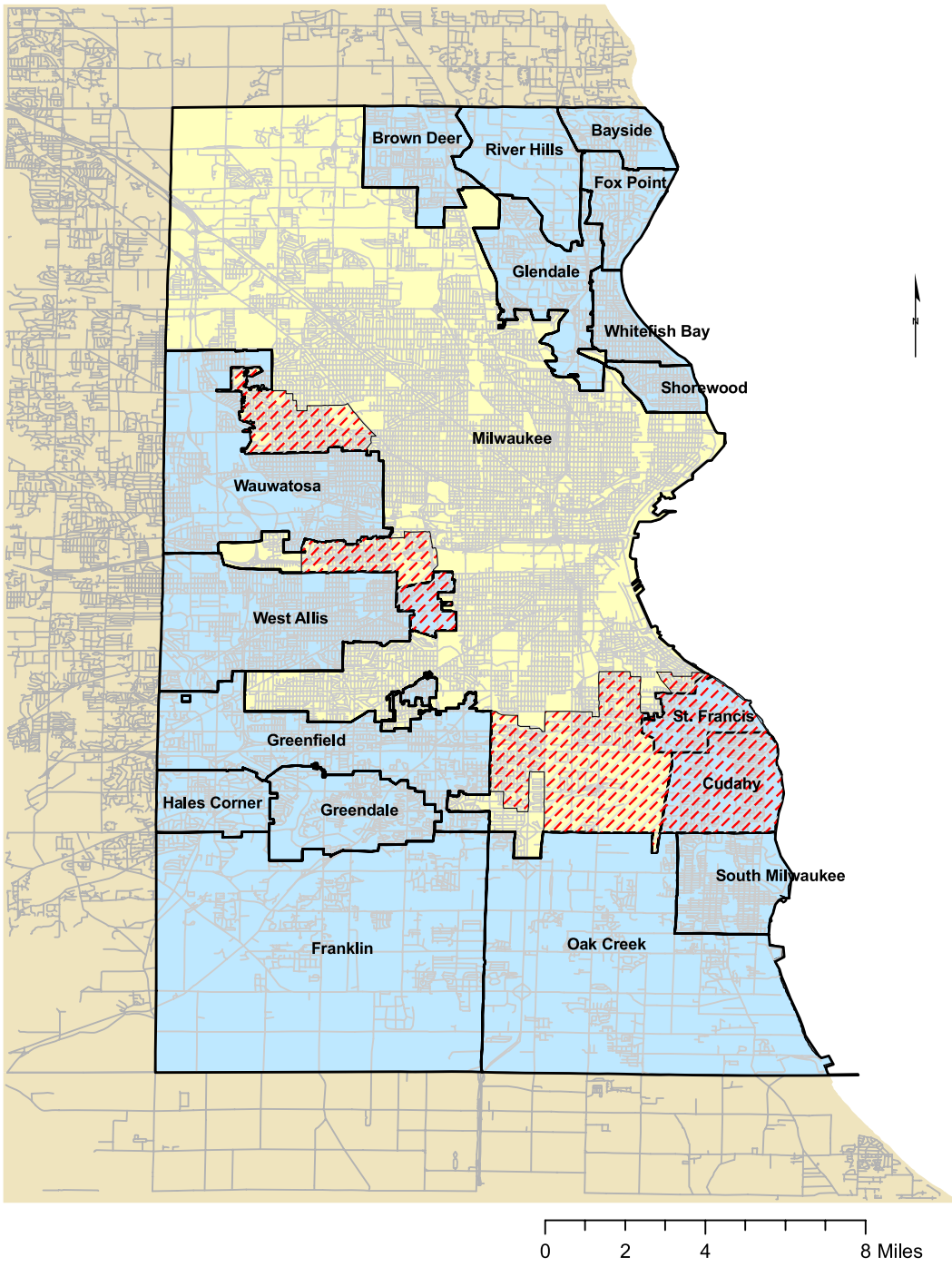


Fig. 1. Milwaukee metropolitan area within Milwaukee county: the ballot initiative of interest was placed on the ballot within the city of Milwaukee (□); suburbs to the city (□) did not have the initiative on the ballot; areas outside Milwaukee (□) did not have any initiatives on the ballot (the figure indicates areas where legislative districts straddle the city limit (▨))

status. Using census data from 2000, we observe that the median household income in Milwaukee is just under $34000, whereas it is nearly $54000 in the suburbs. The percentage of African-American residents of voting age in Milwaukee is 29%, whereas it is less than 1.5% in the suburbs. The difference in median housing value is nearly $60000. Nearly 21% of residents in the suburbs have a college degree whereas just over 12% in the city do. Moreover, examination of State Assembly, State Senate and US 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 US 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 Fig. 1, we highlight the areas where legislative districts straddle the city limit by using diagonally shaded areas on the map.

2.1. Data: covariates, housing values and distances

Our main source of data is the Wisconsin voter file, which is 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 30000 houses that were 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 co-ordinates, 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 500-m 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—on the basis of their address in the voter file.

3. Statistical framework and designs

We adopt the potential outcomes framework and assume that unit j has two potential outcomes, $Y_{\mathrm{T}j}$ and $Y_{\mathrm{C}j}$, 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* with 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 $Y_{\mathrm{T}j} - Y_{\mathrm{C}j}$, the observed outcome is $Y_j = Z_j Y_{\mathrm{T}j} + (1 - Z_j) Y_{\mathrm{C}j}$ and the fundamental problem of causal inference is that we cannot observe both $Y_{\mathrm{T}j}$ and $Y_{\mathrm{C}j}$ simultaneously for any given unit (Rubin, 1974; Holland, 1986; Neyman, 1990). 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 throughout, i.e. 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 neighbours in control areas to vote because of the enthusiasm that is 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 stable unit treatment value assumption violation of this kind would tend to bias the effect towards 0, so any positive effects should be conservative estimates.

3.1. Design 1: conditioning on observables

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

Assumption 1 (conditional geographic treatment ignorability). The potential outcomes are independent of treatment assignment conditional on observed covariates $\mathbf{X}_j$, i.e. $(Y_{\mathrm{T}j}, Y_{\mathrm{C}j}) \perp\!\!\!\perp Z_j|\mathbf{X}_j$.

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 RD design, in which units are assigned to treatment or control groups on the basis of whether their value of an observed covariate exceeds a known cut-off. 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 non-parametric inference of RD effects. Hereafter, we refer to this design as the geographic RD design following Keele and Titiunik (2013b), where this design is discussed in detail. Identification of the average treatment effect at the boundary in the geographic RD design is based primarily on the following assumption.

Assumption 2 (continuity in two-dimensional score). The conditional regression functions $E\{Y_{\mathrm{C}j}|(S_1, S_2)\}$ and $E\{Y_{\mathrm{T}j}|(S_1, S_2)\}$ are continuous in (S_1, S_2) at 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 on the basis of 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—i.e. 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 because of 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 to 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 $N(b_1, b_2)$ collect the set of co-ordinates corresponding to a small geographic neighbourhood around each point (b_1, b_2) on the boundary between A_T and A_C. We assume that there is a small neighbourhood where potential outcomes and treatment assignment are conditionally independent given predetermined covariates.

Assumption 3 (conditional geographic treatment ignorability in local neighbourhood). For each point (b_1, b_2) on the boundary, there is a neighbourhood $N(b_1, b_2)$ such that $(Y_{Tj}, Y_{Cj}) \perp\!\!\!\perp Z_j | \mathbf{X}_j$ for all j with (S_{j1}, S_{j2}) in $N(b_1, b_2)$.

This assumption is in the spirit of an interpretation that was developed by Lee (2008), who argued that RD designs can be seen as local experiments near the cut-off 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 neighbourhood around the boundary that separates the treatment and control areas, where an as-if randomization or independence condition holds after conditioning on predetermined covariates—but possibly does not hold unconditionally as in the standard interpretation.

This design takes from a geographic RD design based on assumption 2 the notion that the treated and control groups will be valid counterfactuals as the score approaches the cut-off 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 for only a subset of the population), but it is not necessarily weaker than assumption 2. Note that, whereas 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 neighbourhood 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 neighbourhood amenities.

Using our application, Fig. 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 predetermined 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 0 even for the smallest distance of 50 m. 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 that is presented below, we invoke assumption 3 including house prices in the

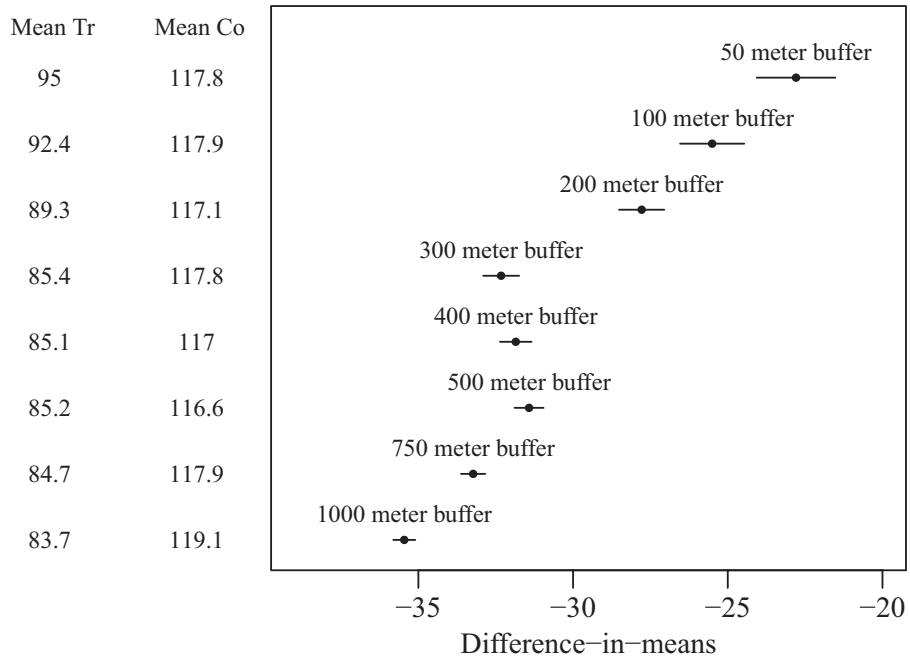


Fig. 2. Difference in means (•) in house prices at the individual level between treatment and control groups for various buffers around the Milwaukee city limit, matching on geographic distance within each buffer: a buffer is a narrow band around the border; a 100-m buffer is a band that extends 100 m from either side of the city limits; units are in thousands of dollars (—, 95% confidence intervals based on paired t-tests)

conditioning set $\mathbf{X}_j$, which entails assuming that the unobservables (the variables that are not in $\mathbf{X}_j$) follow a pattern similar to that in Fig. 2, except that their differences do eventually vanish near 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. Although 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 to find close matches on some covariate distance while considering geographic proximity as an additional covariate. For this task, we could use various matching algorithms, although with most types 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 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. to balance their marginal distributions exactly in aggregate but without constraining who is matched to whom); and there are other covariates for which we may want to minimize only 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 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 to 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). For that, 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 that were 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 $d_{j_{\mathrm{t}},j_{\mathrm{c}}}$ 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 $x_{j_{\mathrm{t}};e}$ is the value of nominal covariate e for treated unit j_{t} with $x_{j_{\mathrm{t}};e} \in \mathcal{B}_e$. 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 $x_{j_{\mathrm{t}};f}$, the value of covariate f for treated unit j_{t}, and $x_{j_{\mathrm{c}};f}$, the value of covariate f for treated unit $j_{\mathrm{c}} \in A_{\mathrm{C}}$. Finally, let $m \in \mathcal{M}$ be the index of the covariates for which it is desired to balance their means, so that $x_{j_{\mathrm{t}};m}$ is the value of covariate m for treated unit j_{t}, and $x_{j_{\mathrm{c}};m}$ is the value of covariate m for control j_{c}.

To solve our problem optimally, we introduce binary decision variables

$$a_{j_{\mathrm{t}},j_{\mathrm{c}}} = \begin{cases} 1 & \text{if treated unit } j_{\mathrm{t}} \text{ is matched to control unit } j_{\mathrm{c}}, \\ 0 & \text{otherwise,} \end{cases}$$

and, for a given scalar λ, we minimize

$$\sum_{j_{\mathrm{t}} \in A_{\mathrm{T}}} \sum_{j_{\mathrm{c}} \in A_{\mathrm{C}}} d_{j_{\mathrm{t}},j_{\mathrm{c}}} a_{j_{\mathrm{t}},j_{\mathrm{c}}} - \lambda \sum_{j_{\mathrm{t}} \in A_{\mathrm{T}}} \sum_{j_{\mathrm{c}} \in A_{\mathrm{C}}} a_{j_{\mathrm{t}},j_{\mathrm{c}}} \tag{1}$$

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,

$$\sum_{j_{\mathrm{c}} \in A_{\mathrm{C}}} a_{j_{\mathrm{t}},j_{\mathrm{c}}} \leqslant 1, \qquad \forall j_{\mathrm{t}} \in A_{\mathrm{T}}, \tag{2}$$

$$\sum_{j_{\mathrm{t}} \in A_{\mathrm{T}}} a_{j_{\mathrm{t}},j_{\mathrm{c}}} \leqslant 1, \qquad \forall j_{\mathrm{c}} \in A_{\mathrm{C}}. \tag{3}$$

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

$$\sum_{j_{\mathrm{t}} \in A_{\mathrm{T}}} \sum_{j_{\mathrm{c}} \in A_{\mathrm{C}}} |\mathbb{1}_{\{x_{j_{\mathrm{t}};e} = b_e\}} x_{j_{\mathrm{t}};e} - \mathbb{1}_{\{x_{j_{\mathrm{c}};e} = b_e\}} x_{j_{\mathrm{c}};e}| a_{j_{\mathrm{t}},j_{\mathrm{c}}} = 0, \qquad \forall e \in \mathcal{E}, \tag{4}$$

$$\sum_{j_{\mathrm{t}} \in A_{\mathrm{T}}} \sum_{j_{\mathrm{c}} \in A_{\mathrm{C}}} a_{j_{\mathrm{t}},j_{\mathrm{c}}} \mathbb{1}_{\{x_{j_{\mathrm{t}};f} = b_f\}} - \sum_{j_{\mathrm{t}} \in A_{\mathrm{T}}} \sum_{j_{\mathrm{c}} \in A_{\mathrm{C}}} a_{j_{\mathrm{t}},j_{\mathrm{c}}} \mathbb{1}_{\{x_{j_{\mathrm{c}};f} = b_f\}} = 0, \qquad \forall b_f \in \mathcal{B}_f, \qquad f \in \mathcal{F}, \tag{5}$$

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

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

Constraints (4), (5) and (6) enforce exact matching, fine balance and mean balance respectively. More precisely, constraint (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 $\mathcal{E}$; constraint (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 constraint (6) forces the differences in means after matching to be less than or equal to the scalar ε_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 programme. 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 ε_m, should be chosen by the analyst on the basis of substantive knowledge of the problem at hand.

We incorporated optimal subset matching into the integer programming framework in the objective function (1) via the λ-parameter. The first term in equation (1) is the total sum of geographic distances between matched pairs, and the second term is the total number of matched pairs. Therefore, λ emphasizes the total number of matched pairs in relation to the total sum of distances and, according to equation (1), it is preferable to match additional pairs if on average they are at a smaller distance than λ. In our application, we choose λ to be equal to the median geographic distance between treated and control subjects so, according to equation (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 that were 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 who reside in the same legislative districts—the red diagonally shaded areas in Fig. 1. Given that state legislators often draw legislative districts on the basis of 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 fourth US Congressional district, the seventh State Senate district and the 20th State Assembly district, and *legislative district exact match II* includes only treated and control voters in the fourth US Congressional district, the fifth State Senate district and the 15th State Assembly district. As we show in the on-line supplemental appendix, exact matching on legislative districts significantly decreases covariate imbalance. Within each triplet of State Assembly, State Senate and US House districts, we restricted all our analyses to observations within 750 m 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), section 8.3). We also imposed some 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 on-line appendix for details on how we constructed this measure. We match exactly on each of these categories since we suspect that voting history is of critical importance. We constrained the means of age and housing value to differ by less than 1 year and $1000 respectively between the treated and control areas. Although 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

The second match hews as closely as possible to the geographic RD 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_{j_t, j_c} to penalize matching subjects residing at more than 2 km 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

The last matching implements design 3 using the integer programming optimal matching framework that was 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 exactly match only on legislative districts and sex, and for the three matching designs, for both legislative district triplets. Table 1 shows means and absolute standardized differences in means (differences 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 neighbourhood 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 on-line appendix contains additional balance test results.

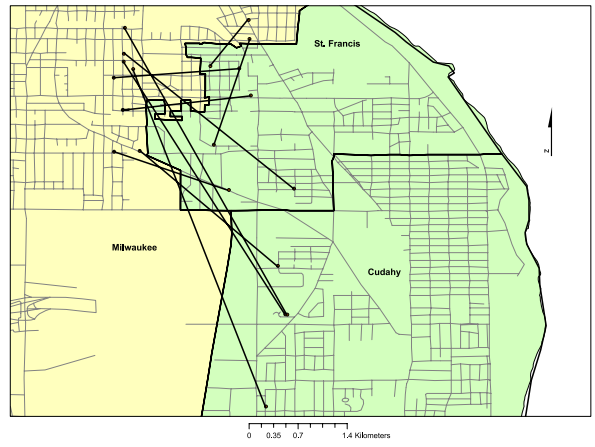
Table 1. Design comparison for covariate balance in three matched comparisons†

Design	*House value ($)*			*Distance (km)*		
	Mean treated	*Mean control*	*Absolute standardized difference*	*Median*	*Mean*	*Pairs*
Legislative district exact match I						
Unmatched	167458	157663	0.44	3.72	3.54	—
1, covariates-only match	156070	157051	0.04	2.87	3.28	2704
2, distance-only match	164070	151135	0.56	0.88	1.04	2524
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	—
1, covariates-only match	144926	144692	0.01	7.72	5.80	1667
2, distance-only match	136049	144802	0.43	1.87	1.68	1663
3, covariates and distance match	140725	141720	0.05	1.96	1.80	536

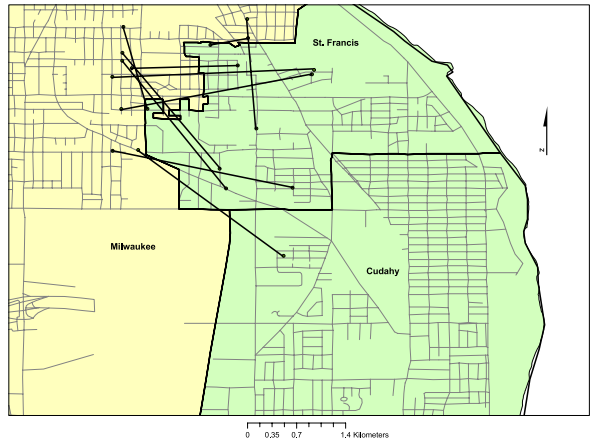
†For all designs, exact matching was done on sex, Congressional district, State Senate district and State Assembly district, and only for observations within 750 m of the border of each legislative district triplet. Design 1 additionally matches exactly on voting history; it also constrains the means of age and housing price to be less than or equal to 1 year and $1000 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 fourth Congressional district, the seventh State Senate district and the 20th State Assembly district. In *legislative district exact match II*, all voters are in the fourth Congressional district, the fifth State Senate district and the 13th State Assembly district. Distance is from the control voter's residence to the treated voter's residence measured in kilometres. 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.

In legislative district exact match I, the median distance between treated and control observations in the unmatched data is a little more than $1\frac{1}{2}$ km. House prices in design 1 are very well balanced, with average house prices differing by just $500. 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 km, a full kilometre larger than in the unmatched data. In design 2, which minimizes only geographic distance, the median geographic distance is reduced to 0.88 km. 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 $10000 in the unmatched data, but in design 2 this difference increases to nearly $13000. As shown in Table 2 in the on-line 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, whereas the median distance within matched pairs is identical to that in design 2.

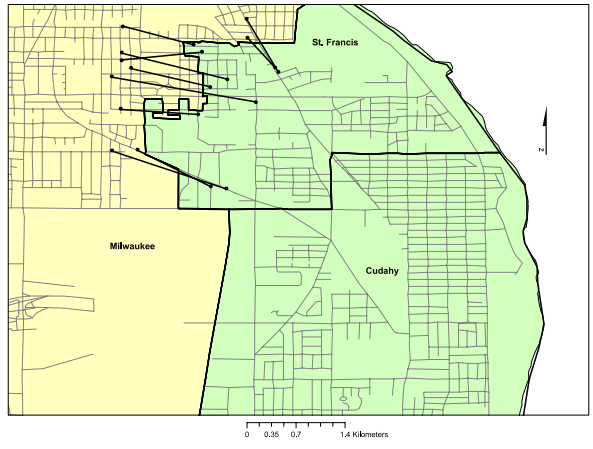
We see a very similar pattern in legislative district exact match II, which is shown in the bottom panel of Table 1. Once again, design 1 reduces imbalance in housing prices relative to the unmatched data set (the standardized difference is reduced by 97%), but at the price of increasing the median distance between treated and controls from about 2 to 3.5 km. Although design 2 decreases this median distance, it also increases the difference in house prices, with a



(a)



(b)



(c)

Fig. 3. 10 pairs of matches randomly sampled from legislative district exact match I: (a) design 1, covariates-only match; (b) design 2, distance-only match; (c) design 3, covariates and distance match

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 \$1000 whereas the median distance within matched pairs exceeds that in design 2 by less than $\frac{1}{10}$th of a kilometre.

The matching results from the three different designs are illustrated in Fig. 3, which shows 10 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 distance to the matched controls varies across the three designs. As seen in Fig. 3(a), when geographic distance is not incorporated in the matching procedure, matched pairs in design 1 are far from each other. Incorporating geographic distance leads to matched pairs that are much closer, as shown in Fig. 3(b). When we match on both covariates and distance, the smaller distances remain as seen in Fig. 3(c), but we also gain better balance in observables. Fig. 3(c) embodies the strategy behind assumption 3, which makes comparisons conditional on important covariates between units that are in a small geographic neighbourhood of the boundary. The matching procedure that is implemented in design 3 can reduce both distance and covariate imbalance by discarding observations from the analysis via the optimal subsetting. Whereas design 3 uses about 24% fewer observations than design 1 in legislative district exact match I, it uses about 62% fewer observations in legislative district exact match II. As mentioned above, this loss of observations is expected given the stricter constraints that are imposed by design 3 and is necessary to ensure that common support holds in our sample.

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 (2002a) and explain in detail how we conduct estimation and inference. In our analysis, there are I matched pairs, $i = 1, \ldots, I$, 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}, $y_{\mathrm{T}ij}$, $y_{\mathrm{C}ij}$ and Y_{ij}—where we use lower-case letters to denote fixed variables. We write $\boldsymbol{\delta}$ for the $2I$-dimensional vector of treatment effects: $\boldsymbol{\delta} = (\delta_{11}, \ldots, \delta_{I2})^{\mathrm{T}}$. 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 $(y_{\mathrm{T}ij}, y_{\mathrm{C}ij})$ which stipulates that $H_0 : y_{\mathrm{T}ij} = y_{\mathrm{C}ij}$ for all i and j and may be expressed as $H_0 : \boldsymbol{\delta} = 0$.

We test Fisher's sharp null hypothesis by 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, $T = \Sigma_{i \in D} \Sigma_{j=1}^{2} Z_{ij} Y_{ij}$, where D is a set of indices for the $I^* \leqslant I$ discordant pairs. Some of the votes that are 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 $T_{\mathrm{c}} = \Sigma_{i,j} Z_{ij} y_{\mathrm{C}ij}$ is the number of votes that would have occurred without an initiative on the ballot. Fisher's sharp null hypothesis, $H_0 : \boldsymbol{\delta} = 0$, says that no votes were caused or prevented by the ballot initiative, implying that $T = T_{\mathrm{c}}$. 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 $\frac{1}{2}$.

In an observational study, we can base a test of the sharp null hypothesis on the randomization distribution of T_c (Rosenbaum, 2002b). The randomization distribution for T_c is valid if every unit j in pair i has the same probability of receiving treatment, $\Pr(Z_{ij}=1)=\frac{1}{2}$. 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 that this model of treatment assignment holds after conditioning on $\mathbf{X}_j$. One model for a sensitivity analysis of this assumption stipulates that $1/(1+\Gamma) \leqslant \Pr(Z_{ij}=1|\mathbf{X}_j) \leqslant \Gamma/(1+\Gamma)$ for a specified value of Γ greater than 1, such that randomization with no hidden bias corresponds to $\Gamma=1$; see Rosenbaum (2002a), chapter 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 that are 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 by using the attributable effect, which is a scalar and unobserved quantity. The attributable effect $\Delta = \Sigma_{i,j} Z_{ij}\delta_{ij}$ is the number of votes due to the ballot initiative, so $T-\Delta$ is the number of treated subjects who would have voted even in the absence of treatment (Rosenbaum, 2002b).

If $H_0: \boldsymbol{\delta}=\boldsymbol{\delta}_0$ is true, we can define $\Delta_0 = \Sigma_{i,j} Z_{ij}\delta_{0ij}$ to estimate an effect parameter for the ballot initiative. We use the method of Hodges and Lehmann (1963) to obtain a point estimate for Δ 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 Δ_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 that are 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 that are not rejected by the test (Rosenbaum, 2002b). For example, if we reject the null hypothesis at the 5% level if $\Delta_0 < a$ and accept if $\Delta_0 \geqslant a$, then a one-sided 95% confidence set for $\boldsymbol{\delta}$ is the set of all $\boldsymbol{\delta}_0$ that are compatible with $\Delta_0 = \Sigma Z_{ij}\delta_{0ij} \geqslant a$. 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 because of 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-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 with a binomial distribution with 1504 trials and probability $\frac{1}{2}$. 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 Δ_0 which makes the

Table 2. Voting patterns in designs 1–3

Design			*Lived in Milwaukee suburb, $Z_{ij}=0$*	
			Did not vote, $Y_{ij}=0$	*Voted, $Y_{ij}=1$*
1†	Lived in Milwaukee, $Z_{ij}=1$	Did not vote, $Y_{ij}=0$	212	690
		Voted, $Y_{ij}=1$	814	2655
2‡	Lived in Milwaukee, $Z_{ij}=1$	Did not vote, $Y_{ij}=0$	199	683
		Voted, $Y_{ij}=1$	782	2523
3§	Lived in Milwaukee, $Z_{ij}=1$	Did not vote, $Y_{ij}=0$	118	401
		Voted, $Y_{ij}=1$	421	1535

†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.
‡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.
§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.

McNemar 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 $H_0:\boldsymbol{\delta}=\boldsymbol{\delta}_0$ and retain the set of values of Δ_0 that are not rejected at the 5% level. We find that $\Delta_0=192$ attributable votes are accepted with one-sided significance level 0.0519, whereas 193 votes are rejected with one-sided significance 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 level of significance 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 that we draw from designs 1 and 2 are quite similar. Under both designs, we would conclude that approximately 3% of those 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, on the basis of 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 programme to combine two identification strategies in a principled manner. Our approach allows us to find voters who are close geographically but who 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 can produce more comparable matches while retaining as many matched pairs as possible in relation to their distances as regulated by λ subject to balance constraints. Although our application focused on a geographic discontinuity, our method of matching could be applied to standard RD 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, which is a result that was 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).

Acknowledgements

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

References

Calonico, S., Cattaneo, M. and Titiunik, R. (2013) Robust nonparametric confidence intervals for regression-discontinuity designs. *Manuscript*. University of Michigan, Ann Arbor.

Cattaneo, M., Frandsen, B. and Titiunik, R. (2013) Randomization inference in the regression-discontinuity design: an application to party advantages in the U.S. senate. *Manuscript*. University of Michigan, Ann Arbor.

Cox, D. R. (1958) *Planning of Experiments*. New York: Wiley.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2009) Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, **96**, 187–199.

Daniel, S. and Yohai, I. (2008) How initiatives don't always make citizens: ballot initiatives in the american states, 1978-2004. *Polit. Behav.*, **30**, 469–489.

Everson, D. (1981) The effects of initiatives on voter turnout: a comparative state analysis. *West. Polit. Q.*, **34**, 415–425.

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

Hodges, J. L. and Lehmann, E. (1963) Estimates of location based on ranks. *Ann. Math. Statist.*, **34**, 598–611.

Holland, P. W. (1986) Statistics and causal inference. *J. Am. Statist. Ass.*, **81**, 945–960.

Imbens, G. W. and Lemieux, T. (2008) Regression discontinuity designs: a guide to practice. *J. Econmetr.*, **142**, 615–635.

Keele, L. J. and Titiunik, R. (2013a) Natural experiments based on geography. *Manuscript*.

Keele, L. J. and Titiunik, R. (2013b) Geographic boundaries as regression discontinuities. *Manuscript*.

Lee, D. S. (2008) Randomized experiments from non-random selection in u.s. house elections. *J. Econmetr.*, **142**, 675–697.

Lee, D. S. and Lemieux, T. (2010) Regression discontinuity designs in economics *J. Econ. Lit.*, **48**, 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 *Housing Economics and Public Policy: Essays in Honour of Duncan Maclennan* (eds K. Gibb and A. O'Sullivan), pp. 67–89. Oxford: Blackwell Publishing.

Neyman, J. (1990) On the application of probability theory to agricultural experiments: essay on principles, section 9 (Engl. transl.). *Statist. Sci.*, **5**, 465–472.

Rosenbaum, P. R. (2002a) *Observational Studies*, 2nd edn. New York: Springer.

Rosenbaum, P. R. (2002b) Attributing effects to treatment in matched observational studies. *J. Am. Statist. Ass.*, **97**, 1–10.
Rosenbaum, P. R. (2010) *Design of Observational Studies*. New York: Springer.
Rosenbaum, P. R. (2012). Optimal matching of an optimally chosen subset in observational studies. *J. Computnl Graph. Statist.*, **21**, 57–71.
Rosenbaum, P. R., Ross, R. N. and Silber, J. H. (2007) Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Am. Statist. Ass.*, **102**, 75–83.
Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.
Rubin, D. B. (1986) Which ifs have causal answers. *J. Am. Statist. Ass.*, **81**, 961–962.
Sinclair, B., McConnell, M. and Green, D. P. (2012) Detecting spillover in social networks: design and analysis of multilevel experiments. *Am. J. Polit. Sci.*, **56**, 1055–1069.
Smith, D. A. and Tolbert, C. J. (2004) *Educated by Initiative: the Effects of Direct Democracy on Citizens and Political Organizations in the American States*. Ann Arbor: University of Michigan Press.
Thistlethwaite, D. L. and Campbell, D. T. (1960) Regression-discontinuity analysis: an alternative to the ex post facto experiment. *J. Educ. Psychol.*, **51**, 309–317.
Tolbert, C. J., Grummel, J. A. and Smith, D. A. (2001) The effects of ballot initiatives on voter turnout in the american states. *Am. Polit. Res.*, **29**, 625–648.
Zubizarreta, J. R. (2012) Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Am. Statist. Ass.*, **107**, 1360–1371.
Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S. A. and Rosenbaum, P. R. (2013) Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *Ann. Appl. Statist.*, **7**, 25–50.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary materials'.