Enhancing Geographic Discontinuities Through Matching

Luke Keele, Rocío Titiunik & José Zubizarreta

Penn State, Michigan, & UPenn

April 13, 2012

Geographic Natural Experiments

- One form of natural experiment is based on geography.
- Here treatment is assigned as an accident of geography.
- Economics: Card and Krueger (1994): minimum wage.
- Political Science: Kranso and Green (2008): media markets and presidential advertising.

Geographic Natural Experiments

- In a geographic natural experiment, an analyst compares units in a *treated area*, A_t, to units in a *control area*, A_c.
- Assignment to these areas is thought to be as-if random or at least haphazard.
- Here, T_i is treatment status, where $T_i = 1$ denotes that unit i is within A_t and $T_i = 0$ denotes the converse.
- Each unit i has two potential outcomes, Y_{i1} and Y_{i0}, which correspond to levels of treatment.
- What design might an analyst use if assignment is not strictly as-if random?

Design I: Conditioning on Observables

Assumption (Conditional Geographic Treatment Ignorability)

For any unit, the potential outcomes are independent of treatment assignment once we condition on the treatment assignment mechanism. That is, $Y_{i1},\ Y_{i0} \perp T_i \mid \mathbf{X}$.

Identification comes from conditioning on a set of observed covariates.

Design II: Geographic Regression Discontinuity

- Another option is to exploit the geographic boundary between A_t and A_c.
- Treatment assignment jumps discontinuously along the boundary.
- Here it is thought that units near the border are more comparable than those farther away.

Design II: Geographic Regression Discontinuity Design

Assumption (Continuity in two-dimensional score)

The conditional regression functions are continuous in (S_1, S_2) *at all points* (c_1, c_2) *on the boundary:*

$$\lim_{(z_1,z_2)\to(c_1,c_2)} E\{Y_{i0}|(S_{i1}, S_{i2}) = (z_1, z_2)\} = E\{Y_{i0}|(S_{i1}, S_{i2}) = (c_1, c_2)\}$$

$$\lim_{(z_1,z_2)\to(c_1,c_2)} E\{Y_{i1}|(S_{i1}, S_{i2}) = (z_1, z_2)\} = E\{Y_{i1}|(S_{i1}, S_{i2}) = (c_1, c_2)\},$$

for all points (c_1, c_2) *on the boundary.*

Intuition Behind Identification

- In a GRD design, when units sort with error around the border or if border is drawn with error, a local treatment effect will be identified.
- Under this design, the only covariate we need is distance to the border.
- But we should expect that people will often be able to carefully select their place of residence based on features such as the quality of schools, crime rates, distance to public transportation, and the price of housing, which implies that the effect may not be identified.

Design III: A Combination

- One alternative would be to combine both designs into a single design.
- Here geographic distance between treated and control unit is minimized while balance in pre-treatment covariates is also enforced.
- Intuition: after conditioning on covariates, treatment assignment near the boundary is as-if random.
- This design may reduce sensitivity to bias from unobserved confounders over and above either design used in isolation.

Design III: A Combination

- Combining these designs within a matching framework would seem straightforward.
- Distance to the border simply becomes one additional covariate in $\mathbf{X}$, the matrix of observed covariates.
- However, a standard matching algorithm may produce suboptimal results in this context.

Design III: A Combination

- Consider unit i that is within A_t and units j and k that are within A_c.
- Let d_{tc} represent the distance between treated and control units, such that d_{ij} is the distance from unit i to unit j.
- Assume $d_{ij} < d_{ik}$.
- Perhaps balance on observables is better for a match between i and k than it is for a match between i and j.
- Assume that while $d_{ij} < d_{ik}$, $d_{ik} < \epsilon$, which is a pre-defined maximum distance outside of which units are not matched.
- One might reasonably wish to sacrifice small amounts of geographic distance for better overall covariate balance.

Design III: Matching

- We need to tailor the matching to allow accommodate this trade off.
- We use mixed integer programming (MIP) matching (Zubizarreta, 2012).
- We use it to minimize geographic distances subject to covariate balance constraints.
- MIP matching is a very flexible matching method for incorporating a variety of constraints such as exact matching, penalties, fine balance, and others.

Based on the decision variable

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

we solve the following matching problem

$$\underset{\mathbf{a}}{\text{minimize}} \quad \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} d_{t,c}a_{t,c} - \lambda \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} a_{t,c} \tag{1.1}$$

$$\text{subject to} \quad \sum_{c\in\mathcal{C}} a_{t,c} \leq 1, \ t \in \mathcal{T} \tag{1.2}$$

$$\sum_{t\in\mathcal{T}} a_{t,c} \leq 1, \ c \in \mathcal{C} \tag{1.3}$$

$$a_{t,c} \in \{0,1\}, \ t \in \mathcal{T}, c \in \mathcal{C} \tag{1.4}$$

$$\sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} \left| \mathbb{1}_{\{x_{t;e}=b_e\}} x_{t;e} - \mathbb{1}_{\{x_{c;e}=b_e\}} x_{c;e} \right| a_{t,c} = 0, \ e \in \mathcal{E} \tag{1.5}$$

$$\sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} a_{t,c}\mathbb{1}_{\{x_{t;f}=b_f\}} - \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} a_{t,c}\mathbb{1}_{\{x_{c;f}=b_f\}} = 0, \ b_f \in \mathcal{B}_f, f \in \mathcal{F} \tag{1.6}$$

$$\left| \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} a_{t,c}x_{t;m} - \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} a_{t,c}x_{c;m} \right| \leq \varepsilon_m \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} a_{t,c}, \ m \in \mathcal{M}, \tag{1.7}.$$

Application

- There is a large literature in political science which suggests that ballot initiatives increase turnout.
- In 2008, a coalition of local labor, educational and community organizations led by 9to5, the National Association of Working Women, placed the following initiative on the ballot in the city of Milwaukee:

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

We compare the turnout behavior of voters in the city of Milwaukee to voter suburbs within Milwaukee county.

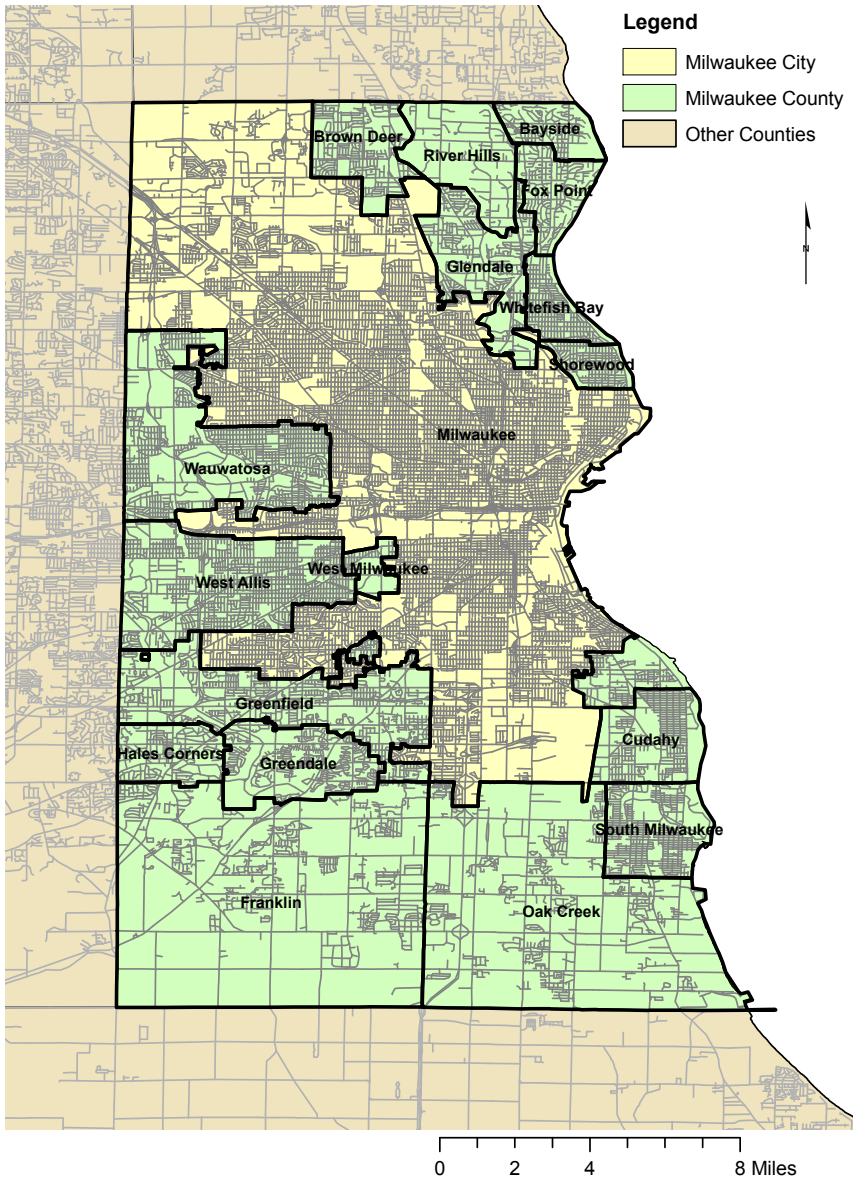

Legend
Milwaukee City
Milwaukee County
Other Counties
Brown Deer
River Hills
Bayside
Fox Point
Glendale
Whitefish Bay
Shorewood
Milwaukee
Wauwatosa
West Allis
West Milwaukee
Greenfield
Hales Corners
Greendale
Cudahy
South Milwaukee
Franklin
Oak Creek
0
2
4
8 Miles

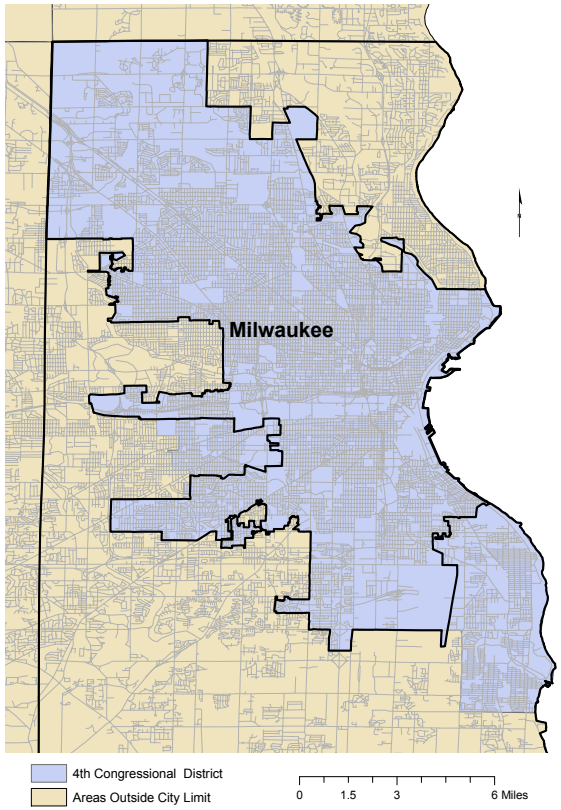
Figure: Milwaukee City Limit - Wauwatosa

Data and GIS Analysis

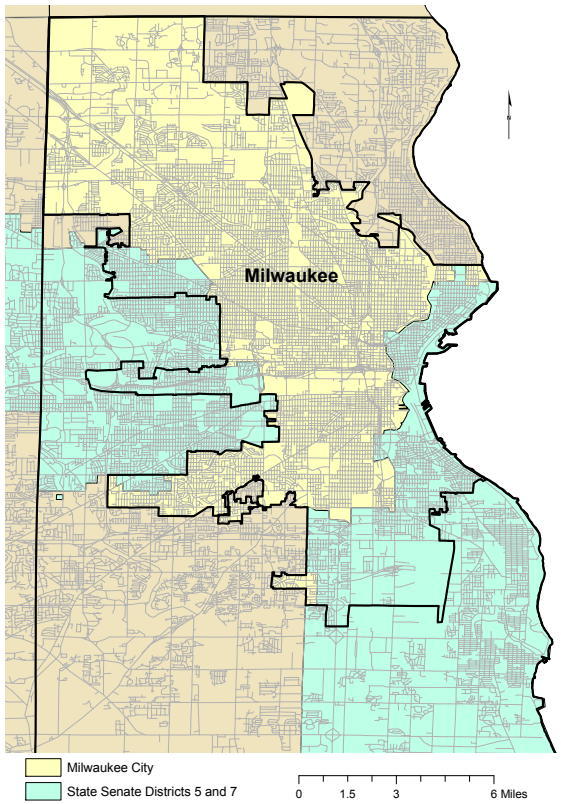
- Data is from Wisconsin Voter File and the Multiple Listing Service (MLS) database.
- Voters and house sales were geocoded to obtain latitude and longitude.
- Latitude and longitude allow us to calculate geographic distance in kilometers between voters.
- We estimated housing value for each voter based on housing sales around each voter.

Exact Matching on Legislative Districts

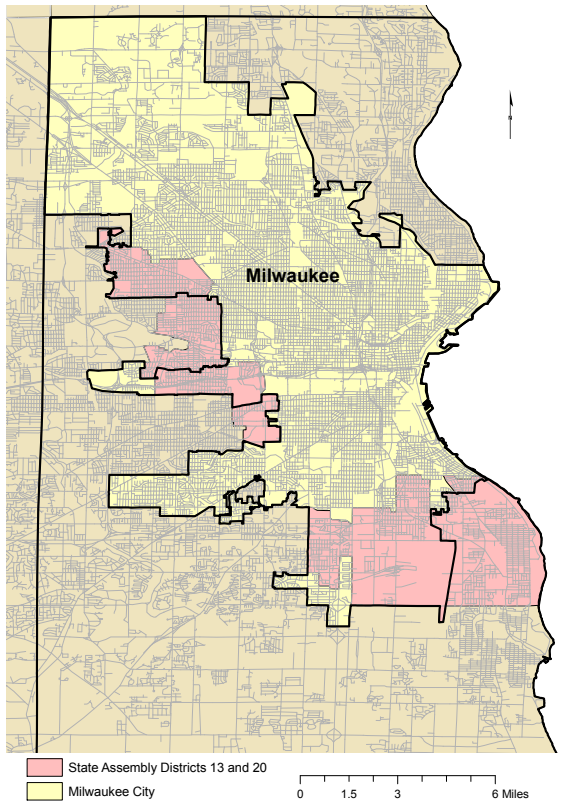
- Before applying the matching algorithm, we exact matched on U.S. Congressional, State Senate, and State Assembly districts.
- Given strategic redistricting, legislative districts are important covariates.
- Only three places where all three districts overlap with the Milwaukee city limit.
- These matches also help computationally.



(a) U.S. Congressional District

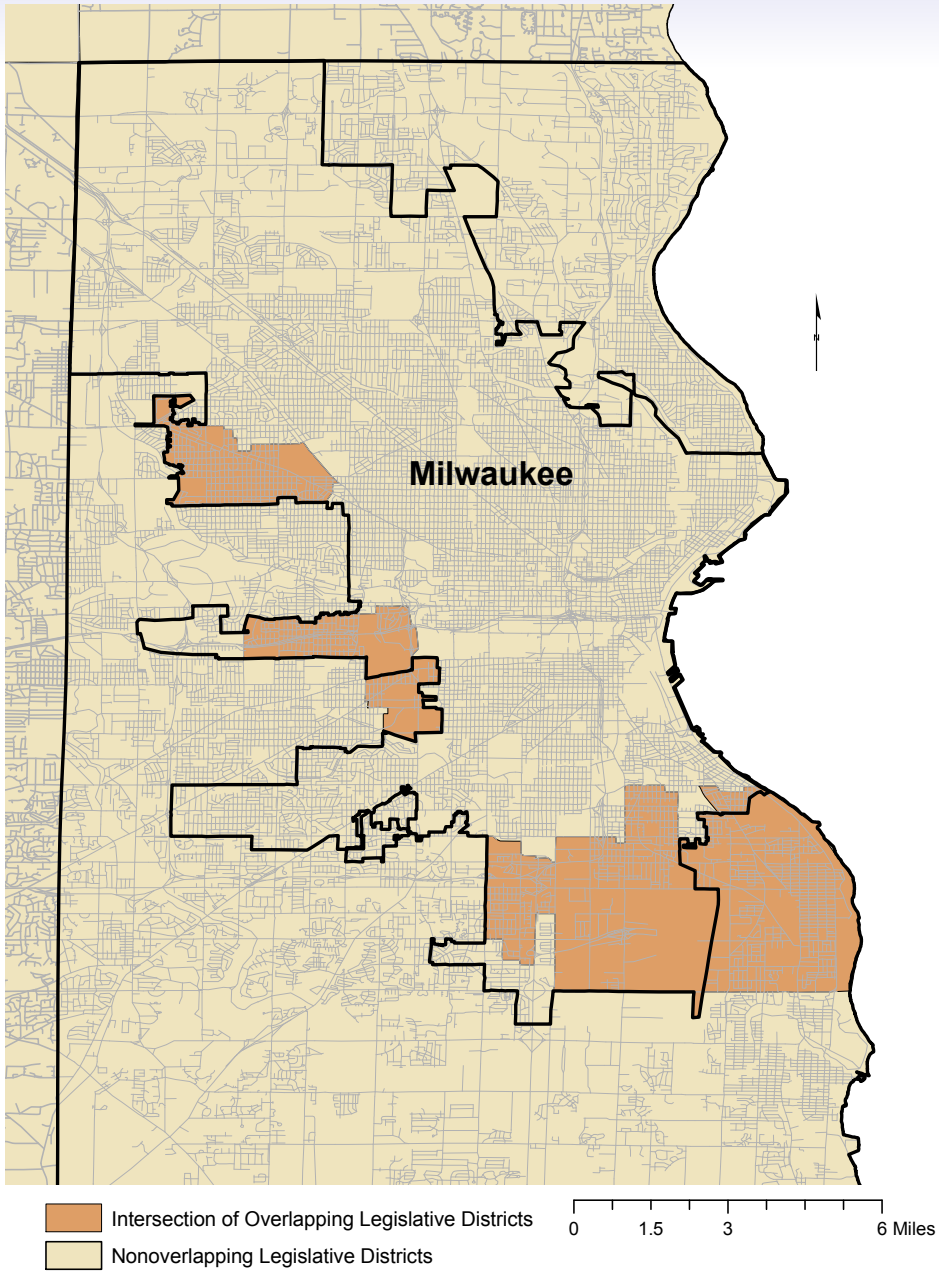


(b) State Senate Districts



(c) State Assembly Districts

Figure: Legislative Districts in Milwaukee Metro Area


Milwaukee
Intersection of Overlapping Legislative Districts
Nonoverlapping Legislative Districts
0
1.5
3
6 Miles

Design 1 - Match on Covariates

- Exact match on voting history and sex.
- Age constrained to be within 1 year.
- Housing value constrained to be within $1000
- We use rank-based Mahalanobis distance matrix.
- We enforce fine balance on seven categories of housing value.

Design 2 - Match on Distance

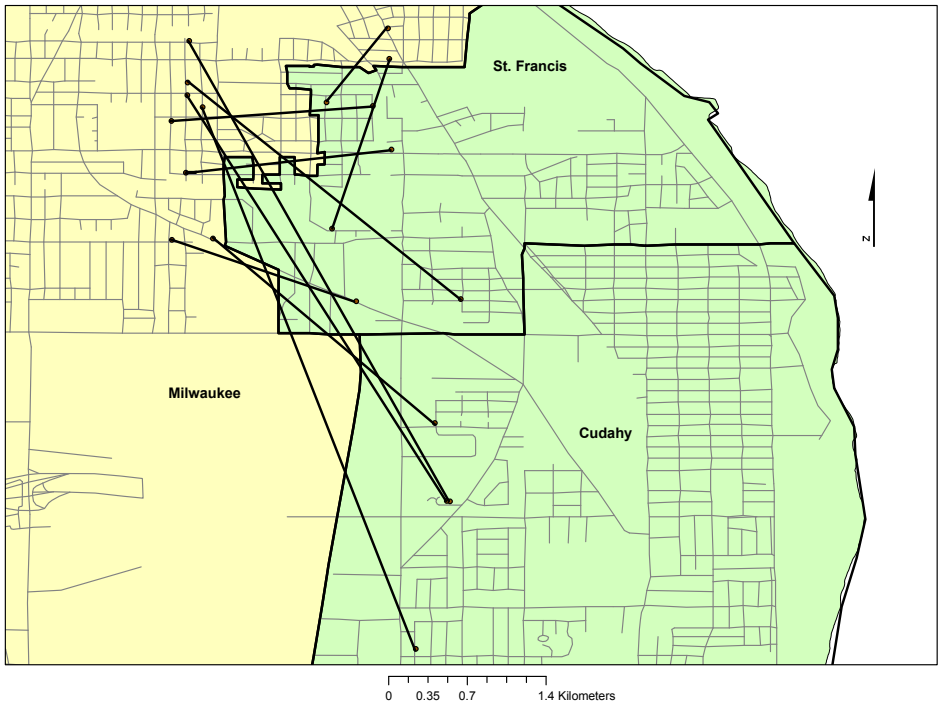
- Match on geographic distances between voters.
- Penalty enforced if match is greater than 2 km.
- Exact match on sex for computational purposes.

Design 3 - Match on Distance and Covariates

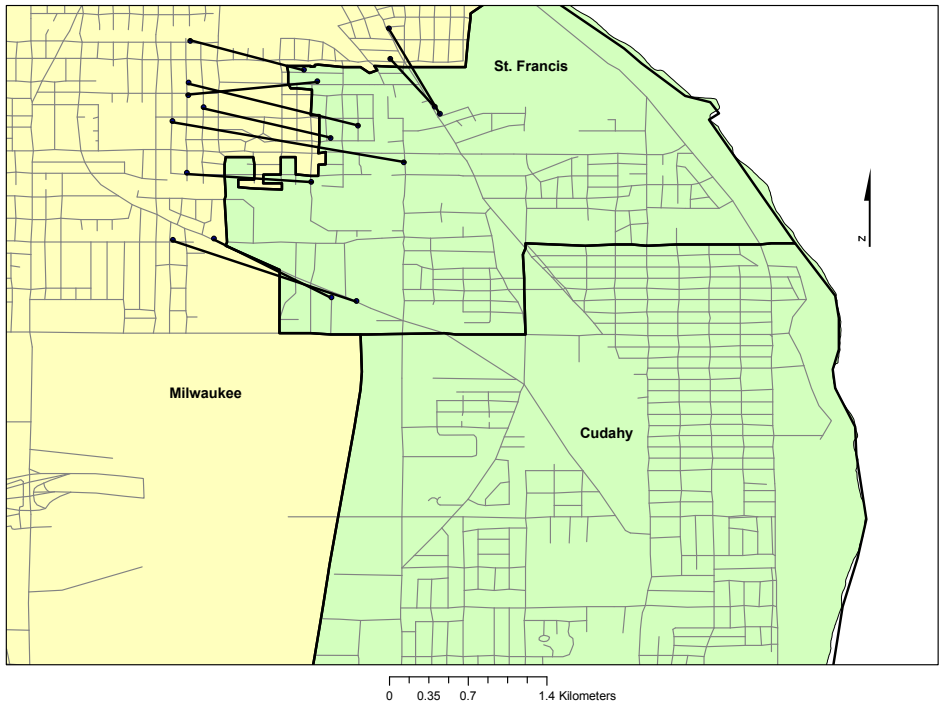
- Match on geographic distances between voters with 2 km penalty.
- Age constrained to be within 1 year.
- Housing value constrained to be within $1000
- We enforce fine balance on seven categories of housing value.
- Exact match on sex and voting history.

Table: Design Comparison

	House Value			Distance		
	Mean Treated	Mean Control	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



(a) Design 1 - Covariates Only



(b) Design 3 - Covariates and Distances

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

Outcomes and Sensitivity Analysis

- For outcome analysis, we combine both legislative district matches.
- Inference based on McNemar's test.
- We also calculate percentage of votes attributable to treatment.
- Based on algorithm from Rosenbaum (2002).
- We also calculate Rosenbaum's bounds to understand whether sensitivity to hidden bias changes across designs.

Table: Point Estimates and Inference Under Three Different Designs

	Design 1	Design 2	Design 3
% of Votes Attributable to Treatment	7.0	6.7	0
p-value	0.008	0.005	0.254
Γ	1.08	1.05	–

Note: The *p*-value is based on McNemar's test for paired binary data. 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.

Conclusion

- Natural experiments typically provide haphazard rather than random treatment assignment.
- As such, natural experiments often need a push from statistical methods.
- In the GRD, using distance alone may not remove all imbalance.
- Using covariates alone may miss important aspects of geography.
- MIP matching provides a way to combine designs in a principled fashion.
- Paper available at `http://www.personal.psu.edu/ljk20/GeoMatch.pdf`.