Recovering Vote by Race in Primary Elections: Does Local Ecological Inference Provide Accurate Estimates of Voting Behavior Without Polling Data?

Ryan D. Enos
renos@gov.harvard.edu
Harvard University

Benjamin E. Lauderdale
blauderdale@gov.harvard.edu
Harvard University

March 25, 2011

ABSTRACT

Scholars know far less about voting behavior in primary elections than in general elections, yet in the many parts of the country where one party dominates, it is primary elections that determine who holds office. Because of the absence of the party cue, the roles of race, income and education in voting behavior are potentially more varied and consequential in the context of primaries. For example, elections contested by multiple black candidates—typically in majority-black electorates—can show striking patterns of racial voting despite the apparent lack of any overt cue from the race of the candidate. Unfortunately, studying primary elections is difficult because polling is sparse and of low quality. In this paper, we begin an assessment of whether the use of a local ecological inference method based on geographically-weighted kernel regressions can facilitate systematic study of these unpolled elections. Our local ecological inference technique weakens the assumptions made by standard ecological inference methods in an intuitive way that is appropriate to political geography. We demonstrate the relative robustness of our local ecological estimates of support by race in a difficult case where the correct answer is known and standard ecological inference methods give spectacularly wrong estimates: the state-level vote by race for Barack Obama and John McCain in the 2008 presidential election. We then apply these methods to a pilot study of the 2000 Illinois 1st Congressional District primary between Barack Obama and Bobby Rush.

1. INTRODUCTION

Most congressional elections in the United States are decided long before the general election. Of the 435 seats in the House of Representatives up for reelection in 2012, the well-regarded *Cook Political Report* lists 330 as "Solid" Democrat or Republican. Campbell (2010) reports that from 1984 to 2008, the Cook Report accurately forecasted 99.8% of Democratic and 99.7% of Republican seats classified as "Solid" prior to the Labor Day before the election. This means that for over 75% of the House elections, the party that will control the seat can be predicted with near certainty. Even if we limited the sample to open seat elections, a large fraction of seats are safe with respect to one party's control. As such, the party primary determines who represents many districts. But the winner of the party primary does not simply determine what type of personality will be in Washington – because of the heterogenous nature of political parties in the United States, considerable ideological variation exists within parties. This ideological variation means that party primaries potentially have important consequences for the policy output of Congress.

Despite the importance of primary elections for determining the composition of Congress, primaries remain understudied, both theoretically and empirically. While there is a healthy literature in Presidential primaries (e.g. Bartels 1988, Cohen, Karol, Noel & Zaller 2008), the literature on low-information Congressional primaries is thin. Perhaps it is the low-key nature of primaries that makes the scholarship rare. After all, Congressional primaries usually receive just as little media attention as voter attention. However, it is this same low-salience electoral environment that makes primaries different from general elections and, consequently calls for scholarly investigation. The low-information, non-partisan environment means that the most robust facts about voter behavior in general elections do not apply to primary elections. Candidate partisanship—by far the most important predictor of vote choice in a general election—is constant within primaries. Moreover, in many primaries—especially those in racially homogeneous, dense urban areas—there is also no variation in candidates' characteristics along urban-rural or racial dimensions. How then do voters decide who to support in these important, yet understudied, elections in which cues such as partisanship, race, and income may not work as they do in the contexts where we have better data?

As we discuss in Section 2, there is reason to believe that primary elections may frequently break down on racial lines even when all candidates are themselves of the same race, but there is little systematic evidence on this question.

The dominant method for studying voting behavior, especially when trying to look for associations between vote choice and individuals characteristics, such as partisanship, income, and race is to use survey data. However, in congressional primaries, scholarly survey data with enough respondents to be useful in studying a single primary campaign is virtually non-existent. The increasing ease with which large surveys can be fielded over the Internet means that nationwide survey data from congressional primary season is increasingly available. But national polls rarely have enough respondents in a given congressional district to be useful in generating even simple cross-tabulations of within-party support for candidates as a function of important factors like race, income, or education. Another way to study vote choice in primaries would be to use to use exit polls, but again, because of the low salience of these elections, exit polls are seldom conducted unless there is a high profile election occurring at the same time.

Because of this lack of individual-level data, the most promising avenue to systematically explore voter behavior in primaries by race involves the use of ecological inference techniques on aggregate vote totals and demographic variation across those aggregations. Such cross-level inference about individual-level behavior from aggregate outcomes and composition require untestable assumptions about the kinds of variation in voter behavior that exist across precincts, which are typically the smallest available aggregation. It is well-established that—because such inferences are identified only through potentially faulty assumptions—in practice ecological inferences can lead to seriously flawed attributions of behavior across groups (Robinson 1950, Achen & Shively 1995, Freedman 1999). However, as our examples later in the paper show, some of the most serious biases that can result from applying ecological inference techniques to voting behavior are mitigated by exploiting geographic information about the aggregate units. Using geography explicitly in estimation allows estimates to exclude comparisons between distant geographic units where the basic ecological inference assumption—that varying outcomes are attributable to variation in group composition—is likely to be confounded. In Section 3 of this paper, we describe a

new *local ecological regression* estimator that is a straightforward extension of the classic Goodman regression (Goodman 1953, Goodman 1959) using geographically localized kernel regressions in place of a simple linear regression. Our approach makes fewer assumptions than previous ecological inference models that allow for spatial heterogeneity (Calvo & Escolar 2003, Haneuse & Wakefield 2004).

We use two examples to illustrate how using geography mitigates severe biases that can result from the fact that variation in voting behavior within racial groups across places in the U.S. is often highly correlated with the racial composition of those places. The first of these is not a primary election, but rather the relatively familiar state-level vote in the 2008 presidential election between Barack Obama and John McCain. We use this example to help motivate the estimator as we develop in in Section 3. Then, in Section 4, we perform a pilot study on the kind of election where we believe these methods are likely to be most useful: the precinct-level vote in the 2000 Democratic primary election for the Illinois 1st congressional district between Barack Obama and Bobby Rush. In both the examples, there are contextual effects in voting behavior by race that prevent estimators that ignore geographic information from recovering accurate estimates for some or all groups. Despite the very different geographic scales in the two examples, using local ecological regression estimates which incorporate geographic information provides more credible estimates of overall voting behavior within racial groups. We believe that the ready availability of precinct-level election returns combined with precinct-level data on race and census block group-level data on race by income facilitates far more comprehensive study of voting in primary elections than has previously been completed.

2. VOTING BEHAVIOR IN PRIMARY ELECTIONS

There is a broad field of research on the question of how voters may respond to the racial demographics of their community. Although this research has many unsettled questions, there seems to be general agreement that voters change their voting behavior depending on the demographics of their general geographic location. The most famous example of this kind of effect was described

by Key (1949), who showed that white voter turnout and support for conservative Democrats in the South was positively related to the proportion African-American in their county. However, the "Racial Threat" literature that followed Key only addresses how racial animus might drive voting behavior when the choice is between two candidates from different racial groups. For example, how white voters in a racially heterogeneous community vote when choosing between a white and African-American candidate (e.g. Carsey 1995); or when one candidate is explicitly racialized, for example Giles & Buckner's (1993) examination of David Duke's 1992 candidacy for Louisiana Governor (see also Voss 1996). While this is an important field of research, it does not test how voters respond when the major candidates are of the same race, which is often the case in large diverse electoral districts that have a majority or large plurality of voters from one racial group. Far more common than the electoral situation in which a white and black candidate are in the same race is the situation when candidates of the same racial group compete.[1] More often than not, this means two white candidates competing. In majority or plurality African-American districts, it often happens that two black candidates compete in the Democratic primary. In these situations, the Racial Threat literature tells us to expect racialized voting, but it is not clear in what way racialized voting will take shape because, unlike candidates from different racial groups or campaigns with a clear racial message, the candidates in these elections do not offer a pre-determined racial cue. It is these types of elections where we focus our inquiry.

Several scholars have examined the incentive structures that result from preferences for descriptive representation. For example, Griffin & Flavin (2007) use survey data and spatial estimates of Members' of Congress ideology to argue that information disparities cause whites and African-Americans to hold incumbents to different levels of accountability – with African-Americans far less likely than whites to punish incumbents for ideological distance from constituents. This means that in the situation where a black incumbent is challenged, we might expect different behavior from white and black voters. Oliver (2001) studied individual political activity as a function of the racial diversity of cities and larger metropolitan neighborhoods. He argued that white and black participation will operate inversely to the other as the percentage of whites in a city increases, with

[1]Even in elections where there are candidates of different races, Highton (2004) argues that whites show little reluctance to vote for African-American candidates.

white political participation increasing and African-American participation decreasing. Also examining heterogeneity within the African-American polity and turnout, Griffin & Keane (2006) argue that liberal African-Americans are more likely to participate when descriptively represented, while "moderate and conservative" African-Americans are actually less likely to vote under descriptive representation.

The most widely cited work on how white and black voters behave in majority black districts is Canon (1999). Canon (1999) considers the fate of candidates in black majority districts after the 1990 Congressional redistricting. Relying on a model of spatial voting (Downs 1957), he argues that the inclusion of a politically diverse black electorate in a black majority district will result in the election of centrist African-American candidates that are able to attract black and white votes. He calls these "new-style black candidates" to differentiate them from the more left-wing "old-style black candidates" that appeal more exclusively to African-American voters. Canon examines the 17 districts in 1992 in which a new black candidate was elected to office. He notes that in seven of these districts in which no white candidate emerged, a "new-style" black candidate won the election. Canon asserts this was accomplished by splitting the black vote and capturing the white vote; however, there is no direct evidence of this from individual-level voting data because such data do not exist.

We suspect that the Canon (1999) model may not be generally applicable across the electoral environments in congressional or similar elections. There are several recent high profile cases where the African-American electorate did not divide as Canon thought it would. Even where they constitute a majority, African-American voters must divide their vote across candidates to enable a competitive elections – a candidate favored by a sufficient super-majority of an African-American majority will be victorious over a candidate attempting to form a biracial coalition. We suspect that the African-American electorate does not always divide its support in the manner hypothesized by Canon. As an example, take the recent mayoral election in Washington, DC – not a Congressional election, but an election with many similar features to heavily Democratic, urban Congressional elections: a majority Black electorate, overwhelming Democratic registration, and a seat historically held by an African-American. In Canon's framework, the incumbent Mayor,

Adrian Fenty, was a "new style" black candidate that should have appealed to a sizable proportion of the Black electorate, while dominating the white electorate and, thereby, winning reelection. His opponent, Vincent Gray, seemingly, purposefully injected race into the election, for example criticizing Fenty for not appointing African-Americans to key posts. In Canon's framework, he was an "old-style" black candidate that should have split the black electorate with Fenty. However, journalistic accounts of this election tell a very different story, for example, the *Washington Times* described a "yawning" racial gap in the electorate, with whites overwhelming favoring Fenty (as Canon would predict), but the African-American majority aligning nearly as overwhelmingly behind Gray.[2] This racial divide, in the heavily African-American city, ensured victory for Gray.

The Gray-Fenty contest in Washington D.C. is hardly the only case where "new-style" candidates have lost to "old-style" candidates. In the pilot study in Section 4 of this paper, we explore the 2000 Congressional Democratic Primary between incumbent Bobby Rush and Barack Obama. In that election, a black majority electorate re-elected an established "old-style" African-American incumbent rather than a candidate that was already noted for his biracial appeal. Anecdotally, we are able to identify several recent cases that have similar characteristics – a "new-style" Black candidate facing an "old-style" Black incumbent in a majority African-American district – and the incumbent comes away victorious, seemingly with wide support from the Black electorate. For example, the 2002 Newark Mayoral Election in which four-term incumbent Sharpe James defeated Cory Booker; or the 2010 New York Democratic Primary when incumbent Charles Rangel, saddled with a recent corruption scandal still defeated challenger Adam Clayton Powell IV. We can identify many such cases, however, to move beyond anecdotal treatments of these cases, we need a way to estimate support by race across full populations of mayoral or congressional primary elections.

[2]Deborah Simmons, "D.C. mayor race defined by race", *The Washington Times*, September 7, 2010. `http://www.washingtontimes.com/news/2010/sep/7/dc-mayor-race-defined-by-race/print/`

3. ESTIMATING GROUP SUPPORT FOR CANDIDATES BY LOCAL ECOLOGICAL INFERENCE

The goal of ecological inference is to infer group behavior from aggregated outcomes combined with information about the group composition of the aggregate units. In the simplest case, there are a set of aggregate units (e.g. U.S. states), a proportional outcome (e.g. Democratic two-party vote share in a presidential election), and two groups with known population fraction in each aggregate unit (e.g. black voters, non-black voters). The standard starting point for inferring group behavior with this kind of data is the Goodman regression (Goodman 1953, Goodman 1959). Goodman's insight was that if one fit a simple linear regression of the aggregate outcome y_i (e.g. candidate 1's vote share in state i) on the fractional composition of Group 2 (e.g. the population share of blacks in state i) in each aggregation x_i, subject to some assumptions one could read off the value of the estimated linear predictor at $x = 0$ and $x = 1$ as estimates of the outcomes for Groups 1 and Group 2.

Unfortunately, the assumptions required to interpret the estimates of Goodman's regression as estimates of the average outcomes in Groups 1 and 2 are often violated (Achen & Shively 1995). A range of methods exist for varying numbers of groups/outcomes, applying more robust statistical frameworks, and for assessing the sensitivity of results to key assumptions (see especially King 1997, Gelman, Park, Ansolabehere, Price & Minnite 2001, Wakefield 2004, Imai, Lu & Strauss 2008, Greiner & Quinn 2009, Glynn & Wakefield 2010). However, none of these methods—nor any as-yet-undiscovered methods—can solve the fundamental non-identification problem with inferring individual-level behavior from aggregate observation. It is precisely this fundamental non-identification that requires careful attention to finding the set of assumptions that are least problematic while still yielding estimates that can be used to address substantive questions like those we have described earlier in this paper.

The methods we explore in this paper are focused on addressing the problem of geographic heterogeneity in the voting behavior of groups (Anselin & Tam Cho 2002), which we believe is especially important for the study of voting behavior generally and primary election returns par-

ticularly. Here, again, there is a literature of previous methodological work to draw upon (Calvo & Escolar 2003, Haneuse & Wakefield 2004), although it is small relative to the literature on ecological inference in general. We draw primarily on the geographically weighted regression approach (Brunsdon, Fotheringham & Charlton 1998) used by Calvo & Escolar (2003), though we implement geographic weighting in a simpler and more intuitive way than Calvo and Escolar.

To understand which assumptions we want to make and which we want to avoid, it is useful to start with the basic Goodman regression. As a running example, we use a particularly difficult dataset for ecological inference: state-level data from the 2008 U.S. presidential election combined with data on the fraction of each state's population that was African-American according to the U.S. Census. While not a primary election, this example has the benefit that the national political geography of the 2008 election is familiar to most readers. It is also a case in which the approach we use vastly outperforms methods that do not use geographic weighting, with the latter giving egregiously wrong answers.

The linear regression coefficients at the core of Goodman's method can be expressed as a weighted average of all the pairwise slopes in the data set (Berman 1988).[3]

$$\hat{\beta}^{LS} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad (1)$$

$$= \frac{\sum_{i,j} (y_i - y_j)(x_i - x_j)}{\sum_{i,j} (x_i - x_j)^2} \quad (2)$$

$$= \frac{\sum_{i,j} (x_i - x_j)^2 \frac{y_i - y_j}{x_i - x_j}}{\sum_{i,j} (x_i - x_j)^2} \quad (3)$$

The important piece of this expression is $\hat{\beta}_{i,j} = \frac{y_i - y_j}{x_i - x_j}$, which is the aforementioned pairwise slope between units i and j. In the running example, one such pair is Mississippi and Alabama. In these two states, the fraction of blacks in the population according to the U.S. Census are $x_{MS} = 0.36$ and $x_{AL} = 0.26$). The Obama vote shares in these two states are $y_{MS} = 0.43$ and $y_{AL} = 0.39$.

[3]From the equation, we can see that the weights are simply the difference in x values for each pair of data points i and j, squared. Thus, the highest weight is put on the pairwise comparisons where there is the largest difference between the group composition of the aggregate units. These weights make sense, because for aggregations with very similar compositions, the pairwise slope is dominated by error rather than the true slope. As noted by King (1997), it is best to fit the Goodman regression weighting the aggregate units by population, though this is only an issue of estimator efficiency if the assumptions of the model are correct.

Thus, the pairwise slope is $\hat{\beta}_{MS,AL} = 0.41$. This is a very crude estimate of the difference between levels of support among blacks and among non-blacks for Obama in these two states. The logic of the estimate is that there are more blacks in the Mississippi population than the Alabama population, and so (by the ecological assumption) we attribute the higher overall level of Obama support in Mississippi to blacks voting for Obama at higher rates than everyone else. Thus, the $\hat{\beta}^{LS}$ in the Goodman EI model can be understood as an estimate of the overall difference between the two groups calculated as a weighted average of the group differences implied by each pairwise comparison.

The logic of averaging over many pairwise state comparisons is that individual comparisons will be imprecise due to other factors influencing voting outcomes. The estimated difference between black and non-black support in Mississippi and Alabama is certainly not very precise. For this election, we have good survey data on vote by race in each state. According to the exit polls done for the 2008 presidential election,[4] the true levels of black support for Obama in Mississippi and Alabama were around 98%, while the true levels of non-black support for Obama in the two states were around 11%. At least in this case the assumption that blacks and non-blacks supported Obama at the same rates in both states is correct. Still, using only one pair of states to form an estimate yields an imprecise estimate (0.41) of the true difference (0.87) between white and black support.

In general, the assumption that true levels of support within racial groups are constant for all aggregate units is not true for U.S. presidential elections examined at the state level. For example, consider the pairwise comparison of Mississippi and Vermont, $x_{VT} = 0.01$ and $y_{VT} = 0.69$. Computing the same pairwise slope as before, we get the dramatically erroneous estimate the rate of black support for Obama was *lower* than the rate of non-black support by 0.71! The problem is that the rate of non-black support for Obama was radically different in Mississippi and Vermont, which is a violation of the assumptions that allow us to use these comparisons to estimate levels of support within groups. The Goodman regression implicitly includes the Mississippi to Vermont comparison, the Mississippi to Alabama comparison, and all the other pairwise comparisons of

[4] http://www.cnn.com/ELECTION/2008/results/polls.main/

states. The errors in the pairwise comparison slopes need to roughly cancel out in order for the regression estimate to be accurate. However, if the racial compositions of states are correlated with the deviations of state-level group behavior from the overall group average, the errors do not cancel out and the Goodman estimate is biased (Achen & Shively 1995).

When run on all U.S. states plus D.C., the Goodman regression indicates that Obama support was higher among blacks by 23%, yielding an estimate of 73% support among blacks versus 49% among non-blacks. The true levels of support according to the exit polls were 95% among blacks and 47% among non-blacks. However, even this level of proximity between the estimates and reality is tenuous: without the 23rd Amendment to the U.S. Constitution, the estimates would be far worse. In the state-level Goodman regression, D.C. is a highly influential point because it has a far higher fraction of blacks than any state ($x_{DC} = 0.60$). Without D.C., the Goodman regression yields abysmal estimates of 42% Obama vote share among blacks and 52% among non-blacks. The problem is not with the assumption that black voters in all states supported Obama at about the same rate, the exit polls reveal that about 92-98% of black voters supported Obama across the entire country. The problem is that white support for Obama varies enormously, and does so in a way that is strongly correlated with black population share. In the states with the largest black population shares, Obama tended to perform poorly among white voters, receiving 11% of the white vote in Mississippi, 14% in Louisiana, 23% in Georgia, 47% in Maryland (the state he won), 26% in South Carolina, and 10% in Alabama. In contrast, in the six states with the smallest black population shares, Obama did much better among whites, though he still only won two of the six, 45% of the white vote in Montana, 68% in Vermont, 33% in Idaho, 58% in Maine, 42% in North Dakota, and 41% in South Dakota. But we only know this because exit polls exist for presidential elections, for primary elections we have no way know if this kind of contextual effect is present.

The previous paragraph highlights a major problem with applying most ecological inference methods to studying voting by race in the U.S.: white voting behavior is correlated with black population fraction at many geographic scales.[5] At the scale of the entire country, whites who

[5]Key (1949) found that Southern whites who lived in counties with larger black populations were more inclined to support the candidates who took the most hardline positions on racial issues. Key could make this inference confidently only because of black disenfranchisement: essentially all the voters in all the counties in his data were white. Had blacks been able to vote, these patterns would have been much harder to detect. In particular, Key

live in the same states as many blacks tend to vote very differently than whites who live in states with few blacks. This problem exists at the level of metropolitan areas, albeit possibly for different reasons: for example, whites who choose to live in integrated areas may vote differently from whites who live in all-white areas. Similarly, blacks who live in integrated areas may vote differently from blacks who live in all-black areas.

While we cannot directly account for these contextual effects, if these contextual effects occur at a larger geographic scale than the geographic units at which the aggregate outcomes and compositions are available, we can improve ecological estimates by making localized comparisons rather than the global comparisons implicit in Goodman's regression and most subsequent approaches. Concretely, we want to eliminate long-distance comparisons like Mississippi to Vermont and focus on nearby comparisons like Mississippi to Alabama and Vermont to Massachusetts. In the remainder of this section, we describe our approach for doing this, and show how it radically improves ecological estimates for the 2008 presidential election case.

Our approach is to fit geographically localized Goodman regressions and then sum up their local predictions about voting in each aggregate unit (in this case, state). Thus, rather than fitting a single unweighted linear regression across all aggregate units, instead we fit a separate weighted regression for each aggregate unit, yielding estimates of $\hat{\alpha}_i$ and $\hat{\beta}_i$. The kernel weights for each local regression are determined by distance from the unit i, using a normal kernel in distance $e^{-\frac{D_{i,j}^2}{2h^2}}$, where the bandwidth parameter h is determined by a cross-validation procedure that is described below. From these $\hat{\alpha}_i$ and $\hat{\beta}_i$, we have local estimates of the rate of support for the candidate treated as the positive outcome for group 1, $\hat{\alpha}_i$ and group 2, $\hat{\alpha}_i + \hat{\beta}_i$. We convert these into raw vote counts from each group, for each candidate, for each aggregate unit. We sum over these estimates for all units i, yielding estimated vote counts for each candidate from each group. These are the reported group-level outcomes.

We determine the optimal bandwidth parameter by leave-one-out cross-validation. This involves running the same regressions of the aggregate vote shares y_i on x_i, with the same weighting function

would have had to make the assumption that none of the black voters were supporting these hardline candidates. Such an assumption would have been entirely defensible in that case, but the key point is that this sort of additional assumption/information is required for ecological inference to work when voting behavior within a group is related to that group's local share of the population.

as described in the previous paragraph, but excluding the data point i when we estimate the local Goodman regression at unit i. We can then calculate a prediction error for unit i by subtracting the predicted vote share $\tilde{y}_i$ from actual vote share y_i. The mean square error (MSE) of these prediction errors is our cross-validation score $\hat{R}(h)$. We find the the value of the bandwidth h that minimizes this cross-validation score by standard optimization methods. Note that if the best predictive model for vote share as a function of group share in the population is the Goodman model—if there is no geographic heterogeneity in within-group behavior—cross-validation will find a very large value of h relative to the scale of distances $D_{i,j}$.

For the case of the 2008 U.S. presidential election, we have run this model using the geographic distance in miles between the centroids of state pairs i, j as our distance measure $D_{i,j}$.[6] We find an optimal bandwidth of $h = 229$ miles when we include D.C. and $h = 231$ miles when we exclude D.C. Note that these are short distances on the scale of U.S. states.[7] This bandwidth indicates that where a state gets a relative weight of 1 in its own local regression, states at a distance of 230 miles from that state receive a relative weight of 0.24 in the regression, states at a distance of 460 miles receive relative weight of 0.05, states at a distance of 690 miles receive relative weight of 0.04, and states at a distance of 920 miles receive an essentially negligible weight 0.0001. For Mississippi, these distances correspond approximately to the distances from the center of the state to the centers of Alabama, Tennessee, Kentucky, and Indiana, respectively. Thus, this bandwidth indicates that the optimal estimates in terms of predicting state-level Obama vote share involves using only very nearby comparisons. The relative weight placed on the Vermont comparison ($\approx 8.6h$) for the local estimates for Mississippi is on the order of 10^{-17}.

Because the optimal model is so localized, our estimator provides strikingly different results from the Goodman regression (i.e. $h = \infty$). Recall that the correct answer for the entire U.S. for this example (according to the exit polls) was 95% support among blacks and 47% support among non-blacks. Recall that the Goodman regression including D.C. indicates 73% support

[6]Similar results are found when we use adjacency distance. Good arguments could be made for using adjacency distance, distance between population-weighted state centers, or other measures. All of these will be highly correlated, so we do not expect to find very different estimates given different measures.

[7]Since state size varies across the country, using geographic distance means that our estimates in the western half of the U.S. draw on fewer data points than those in the eastern half of the country. This is not necessarily a problem, though it might constitute an argument for using adjacency distance rather than geographic distance.

among blacks and 49% among non-blacks, while without D.C. it indicates 52% support among non-blacks and 42% among blacks. Our estimates are both more accurate and more robust to omitting D.C., the most informative and influential data point. The local ecological regression approach estimates 94% support among blacks and 48% among non-blacks when D.C. is included while still estimating 82% support among blacks and 49% support among whites when D.C. is excluded. The former figure is nearly exactly correct, while the latter is still far closer to the true value than the Goodman estimator. The local estimator is less sensitive to the exclusion of D.C. because it only enters the estimates for a few nearby states. Other commonly employed approaches that do not use geographic proximity fail in ways similar to the Goodman estimator. For example, when D.C. is included, King's EI model (King 1997),[8] yields estimates comparable to the Goodman model fitted without D.C., recovering 0.49 for Obama's vote share among blacks and 0.54 for Obama's vote share among whites. Perversely, it recovers worse estimates than the Goodman regression because it is a more robust method in the sense that it is not as sensitive to the presence of an influential point like D.C.. Unfortunately, in this case, the influential point is actually a relatively good indication of the true relationship between race and voting in this data set.

Given these results, we are optimistic about the prospects for using this local estimator to study precinct-level voting returns from primary elections. Voting behavior varies much more slowly across distance on the scale of urban precincts than across even neighboring states. Using the local estimator allows us to avoid the dubious assumption that all the citizens in a particular race or income group vote the same way, regardless of where they have chosen to live within a metropolitan area. Given what is known about American political geography, the same problems that we observe in the state-level analysis of the 2008 presidential election are likely to occur because the politics of American citizens is deeply connected to where they choose to live, both within the country and within their locality. While our local estimation method can still yield mistaken estimates if there are very sharp or non-monotonic changes in voting behavior across neighboring precincts, it does allow us to weaken our untestable assumptions to nearly constant behavior within groups

[8] Using the default settings in its R implementation.

across neighboring or nearby precincts. Gradual change in true group behavior across neighboring precincts will induce only mild biases, but nowhere near the spectacular biases that result from the application of global ecological inference methods to voting behavior.[9]

4. PILOT STUDY: THE 2000 ILLINOIS 1ST CONGRESSIONAL DISTRICT

In 1999, the sitting House representative of the 1st congressional district in Illinois, Bobby Rush (D) attempted to unseat the incumbent mayor of Chicago, Richard M. Daley. Rush was soundly defeated, earning only 28% of the citywide vote against Daley's 72%. Sensing potential weakness in Rush's failure to even unify the African-American vote in Chicago (37% of the city population in 2000), State Senator Barack Obama challenged Bobby Rush for the Democratic nomination for the 1st congressional district in 2000. Rush ultimately won the primary, winning by 61% to 30% for Obama and 9% for minor candidates. We will focus on just the two-candidate vote share and the portion of the district that is within the Chicago city limits for the purposes of this pilot study. In the part of the district that we consider, the two-candidate vote was 68% to 32% in favor of Rush. According to the 2000 U.S. Census, the adult population for this area was 76% black, 15% white, 6% Hispanic, and 3% Other (primarily Asian). Figures 1 and 2 show the precinct-level support for Obama and the precinct-level racial composition of the district.

The pair of maps in Figure 1 show clearly that Obama received much higher vote share in the parts of the district where black population share was lowest. This occurred despite the fact that both candidates were black, both had experience in community organizing on the South Side of Chicago, and both had been representatives for part or all of the district in city, state and national government. However, the association of race with voting is unsurprising given the rhetoric of the campaign and the starkly different biographies of the candidates. As much as any pair of

[9]Especially in urban areas with substantial segregation as a function of race and income, the assumptions necessary to make inferences about voting behavior along these dimensions from aggregate (precinct) outcomes are relatively defensible. First, residential segregation limits the errors that EI techniques can make. Second, selective social networks may make race and class indicators like income important determinants of the vote in low salience elections. In this paper, we focus on race rather than income, primarily because no interpolation is required to recover racial composition of voting precincts from the U.S. Census. However, exploratory analysis that is not reported in this paper suggests that such an interpolation is viable for recovering race by income groups as well.

Figure 1: The distribution of support for Obama and of Black population share across Chicago precincts in the Illinois 1st Congressional District in 2000.

Figure 2: The distribution of Hispanic and Other population share across Chicago precincts in the Illinois 1st Congressional District in 2000.

candidates could, Rush and Obama represented the divide between "old-style" and "new-style" black representatives (Canon 1999). Rush was born in Georgia in 1946 to two black parents, was active in the civil rights movement in the 1960s, co-founded the Illinois Black Panthers in 1968, and has degrees from Roosevelt University and McCormick Theological Seminary in Chicago. Obama was born in Hawaii in 1961 to a white mother and black father, grew up in Hawaii and Indonesia, has degrees from Columbia and Harvard, and taught constitutional law at the University of Chicago.

Rush's argument against Obama's challenge revolved heavily around the question of whether Obama could provide descriptive and substantive representation for the black majority in the district. He put it more bluntly, "Barack Obama went to Harvard and became an educated fool. Barack is a person who read about the civil-rights protests and thinks he knows all about it."[10] Donne Trotter, one of the minor candidates in the race, was even harsher with respect to Obama's credibility with the black community, "Barack is viewed in part to be the white man in blackface in our community. You have only to look at his supporters. Who pushed him to get where he is so fast? It's these individuals in Hyde Park, who dont always have the best interest of the community in mind."[11] Hyde Park, the area where Obama lived, is easily seen in Figure 1: it is the set of racially integrated precincts in the northeast of the district along Lake Michigan. Given the politics of the election, there is good reason to suspect not only that Obama's vote share was quite different in different racial groups, but also in different parts of the district. The white and black voters in the integrated areas of Hyde Park and neighboring South Kenwood may have voted differently than those living in less affluent and more segregated precincts elsewhere on the South Side. There is also visible variation in vote share across the all-black neighborhoods in the district (see Figure 1), although in general support for Rush was very high across all of them. The high degree of precinct-level segregation guarantees that even the Goodman ecological regression will yield decent estimates of voting by race; however, this is still a case where we potentially learn more if we take local geography seriously.

There is no reliable polling data for this race.[12] What happens when we assess overall vot-

[10]David Remnick, "The Joshua Generation: Race and the Campaign of Barack Obama", The New Yorker, November 17, 2008. http://www.newyorker.com/reporting/2008/11/17/081117fa_fact_remnick

[11]Ibid.

[12]The Rush campaign released support by race from internal polling once during the campaign; however, the figures

ing by race using a Goodman regression versus the geographically localized ecological regression method described in the previous section? According to the Goodman regression, Obama's vote share among black, white, Hispanic, and other voters was 0.17, 0.93, 0.26, and 0.47, respectively. According to the local estimator, these vote shares were 0.18, 0.90, 0.42, and 0.27, respectively. The optimal bandwidth is estimated to be 0.83 miles (about 6.5 blocks on Chicago's street grid). While the differences for the white and black voters who make up most of the district are only slightly changed, the estimates for Hispanic and Other voters are quite different. Why does this happen and which estimate is more plausible?

In this district, most of the Hispanic population is in a single cluster in the northwest of the district, an area that also has a substantial numbers of white voters. The Goodman estimator assumes that these white voters behave like white voters elsewhere in the district, voting at a very high rate for Obama (93%). The implication of this is that Obama support among Hispanics must be very low because Obama did much worse in this area than in the other areas with substantial white populations (Hyde Park / Kenwood and Beverly / Mt. Greenwood). However, the local estimator does not compare the white voters who live in this area with majority Hispanic population to the white voters who live miles away in other corners of the district. Instead, it utilizes local variation in the fraction of whites and hispanics to determine that the vote share among whites for Obama in this area is much lower than elsewhere in the district, implying a higher level of Hispanic support for Obama.

The story with the changing estimate for the "other" category is similar. This group is concentrated in Hyde Park / Kenwood, with a few individuals on the outskirts of Chinatown in the northeast corner of the district. Because it involves so few individuals, it is extremely sensitive to the estimates for white and black voters in Hyde Park. Where the Goodman regression assumes that white and black voters in Hyde Park vote the same way as white and black voters elsewhere in the district, the local estimator does not. Given that this was Obama's primary base of support

---

they made available bear only a vague resemblance to the ultimate results. Four months before the election, Rush's campaign claimed that Rush was favored by 72% of black voters and 52% of white voters. Rush led with 68 % of the vote versus 10% for Donne Trotter and 9% for Obama. Clearly, given the time elapsed before the election and the underestimate of Obama's ultimate vote share, this poll has limited relevance to the ultimate results. See "Poll gives Rush big lead; Congressman's foes trail in own districts", Steve Neal, Chicago Sun-Times, November 21, 1999.

and the area in which he lived, further attention to how support by race varied between Hyde Park versus the rest of the district is warranted.

Because the local model relaxes the constancy assumption, we can check for the presence of contextual effects in the levels of support for each candidate. Figure 3 shows the local estimates of black support for Obama both on the map of the district and as a function of the precinct-level black population fraction. The map shows that some of the precinct to precinct variation that is evident in the raw data presented in Figure 1 is smoothed away by the local regression estimator as random variation in the small number of vote counts from precinct to precinct. Perhaps unsurprisingly, given the contours of the campaign, Obama support was somewhat higher among blacks who lived in integrated precincts, most of which were in the Hyde Park / Kenwood area where Obama lived. The plot of these local levels of support for Obama indicate the existence of the kinds of contextual effects that lead to bias in many ecological inference methods. The regression line shown in the figure, weighted by black population in the precinct, has an intercept of 0.256 (se=0.015) and a slope of -0.085 (se=0.016). Obama's support among black voters was lowest in the most segregated precincts, where on average he received 17% of the vote. Obama's support among black voters in a 50-50 white-black precinct was somewhat higher, estimated at 21%. To the extent that the local estimates are insufficiently fine grained, this could be an underestimate of the contextual effects,[13] however it does provide an explanation for which the local estimator finds a slightly smaller overall discrepancy between the white and black vote. If black support for Obama was actually higher in the precincts where there were more whites in the population, this implies lower support for Obama among whites in those places, and a lower difference in white and black vote share overall. The generally high level of segregation in this congressional district mean that these contextual effects do not create very large biases in the Goodman regression, but few congressional districts are as segregated as the 1st Congressional district in Illinois.

One area stands out as anomalous in both Figures 1 and 3, a square mile immediately to the right of the large cutout in the district that is a part of the district to the south. This square mile between Halsted (800 W), State (0 E/W), 71st and 79th, supported Obama far more than any

[13] Incorporating data on income might improve the estimates, as Obama's black support was probably coming from relatively affluent black voters.

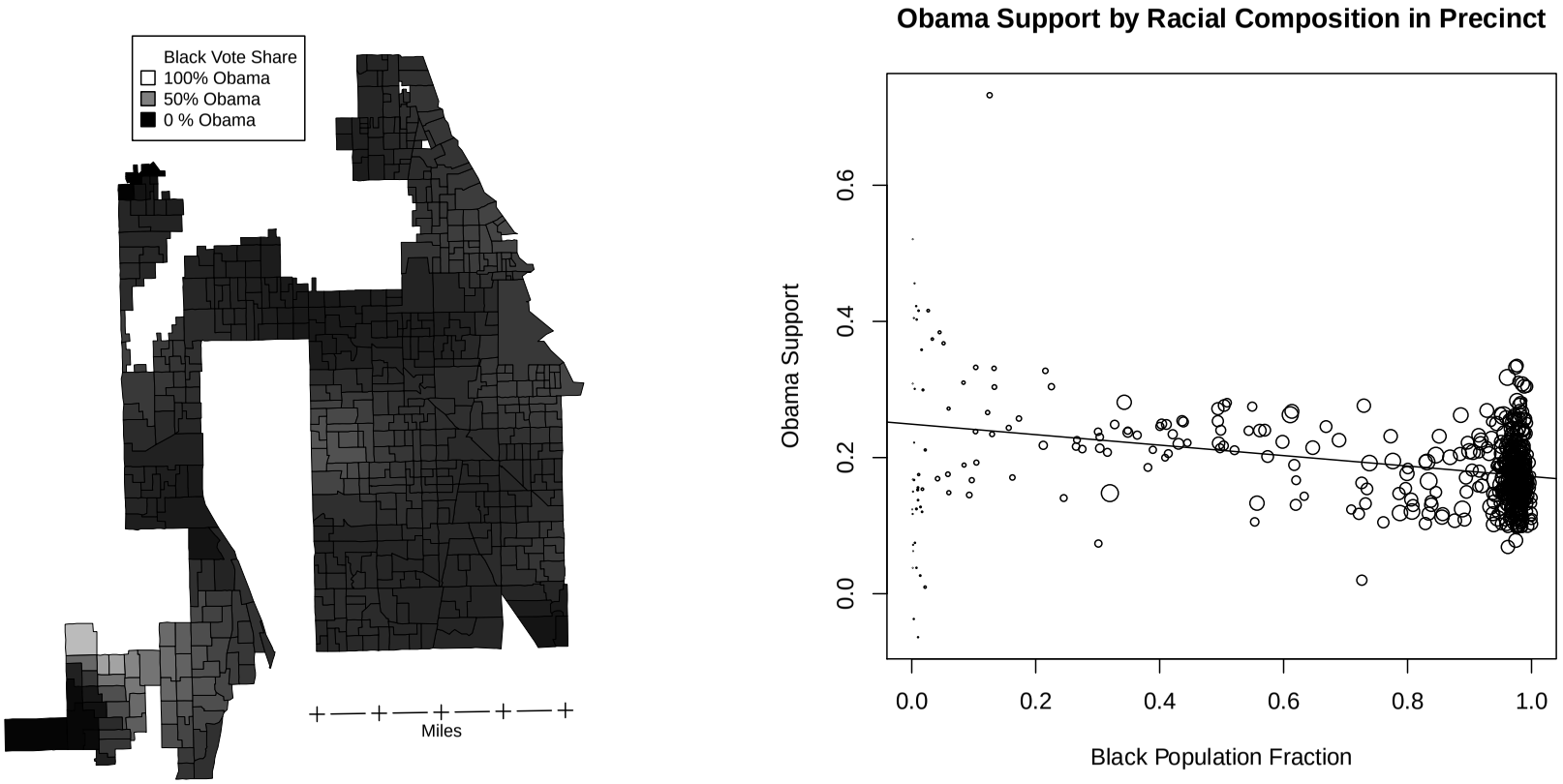

Figure 3: Left plot shows local ecological regression estimates of Obama vote share among blacks in each precinct. Right plot shows these same estimates as a function of local black population fraction. The regression line is based on regression weighted on black population in precinct, which is indicated by the size of the plotted point.

other area with a near 100% black population. It does not map cleanly onto Chicago's community areas, overlapping the border of Englewood and Gresham, to the west of Grand Crossing. This is not an area of particular affluence: in the middle of the area was, as of 2000, one of the worst elementary schools in the city. John Harvard Elementary School was subsequently handed over to a nonprofit organization, the Academy for Urban School Leadership, in 2007. While this area was part of Obama's state senate district, this alone does not explain why Obama's vote share was higher than in nearby areas with similar racial demographics that were also in his district. This is not some artifact of the estimation procedure, the fact that Obama had more support in this square mile is evident by visual inspection of the map of vote share before any ecological inferences are made (Figure 1). Regardless of why Obama received more support here, this example highlights the importance of being sensitive to local variation in voting behavior. Such unexpected patterns exist in many political contexts at many distance scales: it is important to use methods that are robust to such local variation.

5. CONCLUSION

In the case of the Obama versus Rush primary campaign, we see an example of a "new-style" black candidate successfully securing about 90% of the white vote, but losing handily in a super-majority black district to an "old-style" black candidate who secured about 82% of the black vote. Of course, this is only one case, and one that was chosen far from at random from the population of relevant cases. Given the lack of alternative options for studying voting by race across the entire population of congressional primaries, we believe that local ecological regression methods are a promising approach to furthering research in this area. Local methods are more responsive to varying local political geography and avoid some of the biases of methods that do not take the proximity of aggregate units into account. While this paper presents only a single case study of the kind of election that can be used to evaluate theories of voting behavior by race like that of Canon (1999), the approach we suggest could potentially be useful in assessing not only theories of voting by race, but also of voting by income and other factors that are available at very small

aggregations from the U.S. Census.

REFERENCES

Achen, Christopher H. & W. Phillips Shively. 1995. *Cross-Level Inference*. University of Chicago Press.

Anselin, Luc & Wendy K. Tam Cho. 2002. "Spatial Effects and Ecological Inference." *Political Analysis* 10(3):276–297.

Bartels, Larry M. 1988. *Presidential primaries and the dynamics of public choice*. Princeton University Press.

Berman, Mark. 1988. "A Theorem of Jacobi and its Generalization." *Biometrika* 75(4):779–783.

Brunsdon, Chris, Stewart Fotheringham & Martin Charlton. 1998. "Geographically Weighted Regression: Modeling Spatial Non-Stationarity." *Journal of the Royal Statistical Society: Series D (The Statistician)* 47(3):431–443.

Calvo, Ernesto & Marcelo Escolar. 2003. "The Local Voter: A Geographically Weighted Approach to Ecological Inference." *American Journal of Political Science* 47(1):189–204.

Campbell, James E. 2010. "The Seats in Trouble Forecast of the 2010 Elections to the U.S. House." *PS: Political Science & Politics* 43(4):627–630.

Canon, David T. 1999. *Race, Redistricting, and Representation*. University of Chicago Press.

Carsey, Thomas M. 1995. "The contextual effects of race on white voter behavior: The 1989 New York City Mayoral Election." *Journal of Politics* 57(1):221–28.

Cohen, Marty, David Karol, Hans Noel & John Zaller. 2008. *The party decides: presidential nominations before and after reform*. University of Chicago Press.

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

Freedman, David A. 1999. Ecological Inference and the Ecological Fallacy. Technical Report 549 International Encyclopedia of the Social and Behavioral Sciences.

Gelman, Andrew, David K. Park, Stephen Ansolabehere, Philip N. Price & Lorraine C. Minnite. 2001. "Models, Assumptions and Model Checking in Ecological Regressions." *Journal of the Royal Statistical Society (Series A)* 164(1):101–118.

Giles, Micheal W. & Melanie A. Buckner. 1993. "David Duke and black threat: An old hypothesis revisited." *Journal of Politics* 55:702–13.

Glynn, Adam N. & Jon Wakefield. 2010. "Ecological Inference in the Social Sciences." *Statistical Methodology* 7:307–322.

Goodman, Leo. 1953. "Ecological Regression and the Behavior of Individuals." *American Sociological Review* 18:663.

Goodman, Leo. 1959. "Some Alternatives to Ecological Correlation." *American Journal of Sociology* 64:610–625.

Greiner, D. James & Kevin M. Quinn. 2009. "R x C Ecological Inference: Bounds, Correlations, Flexibility, and Transparency of Assumptions." *Journal of the Royal Statistical Society (Series A)* 172(1):67–81.

Griffin, John D. & Michael Keane. 2006. "Descriptive Representation and the Composition of African American Turnout." *American Journal of Political Science* 50(4):998–1012.

Griffin, John D & Patrick Flavin. 2007. "Racial Differences in Information, Expectations, and Accountability." *The Journal of Politics* 69(1):220–236.

Haneuse, Sebastien & Jonathan Wakefield. 2004. Ecological Inference Incorporating Spatial Dependence. In *Ecological Inference: New Methodological Strategies*, ed. Gary King, Ori Rosen & Martin A. Tanner. Cambridge University Press chapter 12, pp. 266–302.

Highton, Benjamin. 2004. "White Voters and African American Candidates for Congress." *Political Behavior* 26:1–25.

Imai, Kosuke, Ying Lu & Aaron Strauss. 2008. "Bayesian and Likelihood Inference for 2 x 2 Ecological Tables: An Incomplete-Data Approach." *Political Analysis* 16:41–69.

Key, V. O. 1949. *Southern Politics in State and Nation*. University of Tennessee Press.

King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press.

Oliver, Eric. 2001. *Democracy in Suberbia*. Princeton, NJ: Princeton University Press.

Robinson, W. S. 1950. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 15:351–357.

Voss, D. Stephen. 1996. "Beyond racial threat: Failure of an old hypothesis in the new South." *The Journal of Politics* 58(4):1156–70.

Wakefield, Jon. 2004. "Ecological Inference For 2 x 2 Tables." *Journal of the Royal Statistical Society (Series A)* 167(3):385–426.