Advance Access publication March 17, 2016

Political Analysis (2016) 24:263–272
doi:10.1093/pan/mpw001

Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records

Kosuke Imai
Department of Politics and Center for Statistics and Machine Learning, Princeton University, Princeton, NJ 08544
e-mail: kimai@princeton.edu; URL: http://imai.princeton.edu (corresponding author)

Kabir Khanna
Department of Politics, Princeton University, Princeton, NJ 08544

Edited by Justin Grimmer

In both political behavior research and voting rights litigation, turnout and vote choice for different racial groups are often inferred using aggregate election results and racial composition. Over the past several decades, many statistical methods have been proposed to address this ecological inference problem. We propose an alternative method to reduce aggregation bias by predicting individual-level ethnicity from voter registration records. Building on the existing methodological literature, we use Bayes's rule to combine the Census Bureau's Surname List with various information from geocoded voter registration records. We evaluate the performance of the proposed methodology using approximately nine million voter registration records from Florida, where self-reported ethnicity is available. We find that it is possible to reduce the false positive rate among Black and Latino voters to 6% and 3%, respectively, while maintaining the true positive rate above 80%. Moreover, we use our predictions to estimate turnout by race and find that our estimates yields substantially less amounts of bias and root mean squared error than standard ecological inference estimates. We provide open-source software to implement the proposed methodology.

1 Introduction

In political behavior research as well as voting rights litigation, it is often of interest to infer turnout and vote choice among different racial groups. For instance, political scientists estimate turnout by race in order to study disparities in political participation (e.g., Gay 2001; Hajnal and Trounstine 2005), mobilization efforts (e.g., Barreto 2007), and the effects of co-ethnic candidates and representatives (e.g., Herron and Sekhon 2005). In voting rights cases, litigants wish to estimate turnout and vote choice among ethnic groups to build empirical evidence for the existence of racial polarization (e.g., Greiner 2007).

However, such efforts face a well-known methodological obstacle, known as the ecological inference problem. Since the race of individual voters is typically unknown, one must infer turnout by race from aggregate data. A number of statistical methods have been developed to address this problem (e.g., Goodman 1953; King 1997; King, Rosen, and Tanner 2004; Wakefield 2004; Greiner and Quinn 2008; Imai, Lu, and Strauss 2008). Nevertheless, all of these methods suffer from a fundamental problem of indeterminacy, and as a result, in recent years, methodologists have turned to the idea of combining aggregate data with individual-level data (e.g., Wakefield 2004; Imai Lu, and Strauss 2008; Greiner and Quinn 2010).

Authors' note: We thank Bruce Willsie, the CEO of L2, for the data and answering numerous questions, and the participants of "Building the Evidence to Win Voting Rights Cases" conference at the American Constitutional Society for Law and Policy for their helpful comments. Two anonymous reviewers provided helpful suggestions. The **R** package, wru: Who Are You? Bayesian Prediction of Racial Category Using Surname and Geolocation, is freely available for download at https://cran.r-project.org/package=wru. Replication files for this study are available on the *Political Analysis* Dataverse at http://dx.doi.org/10.7910/DVN/SVY5VF. Supplementary materials for this article are available on the *Political Analysis* Web site.

In this article, we propose to improve upon ecological inference by predicting individual race from voter registration records. Building on the existing methodological literature in public health (Fiscella and Fremont 2006; Elliott et al. 2008, 2009), we use Bayes's rule to combine the Census Bureau's Surname List with information in geocoded voter registration records. By incorporating additional information such as party registration, this methodological framework offers improvements over the common practice of using surname alone or surname and geolocation to predict individual ethnicity (e.g., Michelson 2003; Barreto, Segura, and Woods 2004; Tam Cho, Gimpel, and Dyck 2006; Fieldhouse and Cutts 2008; Henderson, Sekhon, and Titiunik 2014; Enos 2015; Harris 2015). We also explicate and probe the assumptions that underlie the existing and proposed methods. Although some scholars have turned to proprietary methods of estimating voter race (e.g., Ansolabehere and Hersh 2003; Fraga 2013, 2016), we believe that methodological transparency is important for academic research, and these assumptions reveal the promise and limitations of the methods discussed here.[1] To implement the proposed methodology, the **R** package, `wru: Who Are You? Bayesian Prediction of Racial Category Using Surname and Geolocation`, is freely available for download at https://cran.r-project.org/package=wru.

Finally, this article reports the results of a large-scale empirical validation study. We examine the performance of various methods of estimating individual-level race, as well as turnout by race at the precinct and district levels. Specifically, we use the Florida voter file, predicting the race of over nine million voters and validating our predictions using self-reported race data.[2] We choose Florida because self-reported race is collected on voter registration cards by law.[3] Florida also has a relatively large number of Blacks and Latinos, enabling us to empirically validate the accuracy of the proposed method and other methods at the individual level among these minority groups. We show that the proposed method reduces the false positive rate among Black and Latino voters to 6% and 3%, respectively, while maintaining the true positive rate at above 80%. Moreover, we find that the bias and root mean squared error (RMSE) of our estimated turnout by racial groups are substantially less than those of the standard ecological inference estimates.

2 The Methodology

We begin by describing the existing Bayesian method in public health that combines the surname list with the geocoded location of individual residence. We then describe our extension, which allows researchers to incorporate the various information in voter registration records.

2.1 *The Bayesian Prediction*

Researchers interested in measuring racial disparities in healthcare have developed a methodology to combine surname analysis and geocoded data to estimate individual race via Bayes's rule (Fiscella and Fremont 2006; Elliott et al. 2008, 2009). We begin by describing the Bayesian method developed by Elliott et al. (2009). Let the surname and geolocation of voter i be denoted by S_i and G_i, respectively. We use R_i to represent an unobserved variable indicating the racial group voter i belongs to. Let $\mathcal{R}$, $\mathcal{G}$, and $\mathcal{S}$ represent the set of all racial groups, all geolocations, and all surnames, respectively.

We are interested in estimating $\Pr(R_i = r | S_i = s, G_i = g)$, or the conditional probability that voter i belongs to racial group r given his/her surname s and geolocation g. Using the data from the Census Bureau, we have the racial composition of frequently occurring surnames, that is,

[1] In addition, unlike the Bayesian methods, the Catalist's race prediction method does not offer a formal probabilistic prediction and instead utilizes an informal scheme of "Highly Likely," "Likely," and "Possibly."

[2] Fraga (2016) conducts an empirical validation of Catalist's proprietary race prediction method. There are several differences between the current validation and that of Fraga (2016). For example, Catalist bases its predictions on self-reported race in the voter file whenever it is available. In contrast, our goal is to predict individual race when such information is not available. To do this, we utilize other available information in the voter file, such as surname, geolocation, and party registration.

[3] Voter registration cards in Alabama, Florida, Georgia, Louisiana, Mississippi, North Carolina, and South Carolina ask voters to identify their race/ethnicity. Pennsylvania and Tennessee provide an optional blank field for race.

$\Pr(R_i = r|S_i = s)$, the racial composition of each geolocation (e.g., Census blocks and voting precincts), that is, $\Pr(R_i = r|G_i = g)$, and the population proportion of each geolocation, that is, $\Pr(G_i = g)$.

The method assumes that geolocation and surname are statistically independent conditional on race. That is, once we know a voter's race, her surname is not informative about where she lives.[4] We formalize this assumption as follows:

$$G_i \perp\!\!\!\perp S_i|R_i. \tag{1}$$

Assuming equation (1) holds, Bayes's rule implies

$$\Pr(R_i = r|S_i = s, G_i = g) = \frac{\Pr(G_i = g|R_i = r)\Pr(R_i = r|S_i = s)}{\sum_{r' \in \mathcal{R}} \Pr(G_i = g|R_i = r')\Pr(R_i = r'|S_i = s)}, \tag{2}$$

where using Bayes's rule again we can calculate $\Pr(G_i = g|R_i = r)$ as $\Pr(R_i = r|G_i = g)\Pr(G_i = g)/\sum_{g' \in \mathcal{R}} \Pr(R_i = r|G_i = g')\Pr(G_i = g')$. Thus, the method provides a probabilistic prediction of individual ethnicity.

2.2 *The Proposed Extension*

We propose to extend the above Bayesian prediction method by incorporating a set of individual-level covariates available in the voter files. In this article, we focus on age, gender, and party registration, which are often available in voter files. However, under the proposed framework, other information can be incorporated in a similar manner. Let X_i represent our two demographic variables, that is, age and gender. Furthermore, let P_i represent the party registration of voter i.

To incorporate the demographic variables X_i, we replace the assumption given in equation (1) with the following:

$$\{G_i, X_i\} \perp\!\!\!\perp S_i|R_i. \tag{3}$$

This assumption states that given a voter's race, his/her surname does not contain any information about his/her geolocation and demographics. It could be violated, for example, if the rate of interracial marriage is correlated with surname and geolocation through age or gender within each racial category.[5] As with equation (1), we view the validity of this assumption as an empirical question.

If equation (3) holds, it is straightforward to predict individual race using Bayes's rule,

$$\Pr(R_i = r|S_i = s, G_i = g, X_i = x) = \frac{\Pr(G_i = g, X_i = x|R_i = r)\Pr(R_i = r|S_i = s)}{\sum_{r' \in \mathcal{R}} \Pr(G_i = g, X_i = x|R_i = r')\Pr(R_i = r'|S_i = s)}, \tag{4}$$

where $\Pr(G_i = g, X_i = x|R_i = r)$ can be obtained from the Census Summary File.

We further extend this method to incorporate party registration as well as demographics by considering two possibilities. The first approach requires that researchers have information about the population distribution of party registration given each racial category, that is, $\Pr(P_i = p|R_i = r)$ for all $p \in \mathcal{P}$ and $r \in \mathcal{R}$, where $\mathcal{P}$ is the set of all parties.[6] For example, we may obtain an estimate of this quantity from a national survey. This approach is based on the following conditional

[4]There are different ways in which this assumption could be violated. For example, surnames may be associated with wealth, which may be predictive of where people live, even within a racial group. Another scenario is that within racial groups, families cluster together in neighborhoods. While recognizing these possibilities, ultimately, we view the validity of this assumption as an empirical question. Our analysis shows that by conditioning on race, we can account for much of the association between surname and geolocation (see Supplementary Appendix A.3). We also find that our predictions of race are quite accurate, suggesting that equation (1) is reasonable.

[5]We thank an anonymous reviewer for pointing out this possibility.

[6]We classify voters as Democrats, Republicans, or Other. Other includes Independents and members of minor parties. Knowing that a voter is not registered with a major party is informative, because the racial composition of this group differs from the racial composition of registered Democrats and Republicans.

independence assumptions:

$$\{G_i, P_i, X_i\} \perp\!\!\!\perp S_i | R_i \tag{5}$$

$$\{G_i, X_i\} \perp\!\!\!\perp P_i \,|\, R_i. \tag{6}$$

Equation (5) implies that once we know a voter's race, his/her surname is not informative about his/her geolocation, party registration, and demographics. Similarly, the second assumption in equation (6) states that given a voter's race, his/her party registration does not provide any additional information about his/her geolocation and demographics. Under these assumptions, we can apply Bayes's rule to predict individual ethnicity:

$$\begin{aligned} &\Pr\,(R_i = r | S_i = s, G_i = g, X_i = x, P_i = p) \\ &= \frac{\Pr\,(G_i = g, X_i = x | R_i = r)\Pr\,(P_i = p | R_i = r)\Pr\,(R_i = r | S_i = s)}{\sum_{r' \in \mathcal{R}} \Pr\,(G_i = g, X_i = x | R_i = r')\Pr\,(P_i = p | R_i = r')\Pr\,(R_i = r' | S_i = s)}. \end{aligned} \tag{7}$$

Unlike the first approach, the second approach for incorporating party registration allows one to predict race without additional information. This alternative strategy is based on the following independence assumption as well as the assumption given in equation (1):[7]

$$\{X_i, P_i\} \perp\!\!\!\perp S_i | G_i, R_i, \tag{8}$$

which implies that given a voter's geolocation and race, her surname has no predictive power for her demographics and party registration. Under these assumptions, the application of Bayes's rule yields:

$$\begin{aligned} &\Pr\,(R_i = r | S_i = s, G_i = g, P_i = p, X_i = x) \\ &= \frac{\Pr\,(P_i = p, X_i = x | G_i = g, R_i = r)\Pr\,(G_i = g | R_i = r)\Pr\,(R_i = r | S_i = s)}{\sum_{r' \in \mathcal{R}} \Pr\,(P_i = p, X_i = x | G_i = g, R_i = r')\Pr\,(G_i = g | R_i = r')\Pr\,(R_i = r' | S_i = s)}, \end{aligned} \tag{9}$$

where we model the first term in the numerator and denominator as:

$$\begin{aligned} &\Pr\,(P_i = p, X_i = x | G_i = g, R_i = r) \\ &= \Pr\,(P_i = p | X_i = x, G_i = g, R_i = r)\Pr\,(X_i = x | G_i = g, R_i = r). \end{aligned} \tag{10}$$

The second term of this equation can be calculated directly from the Census data as $\Pr\,(X_i = x | G_i = g, R_i = r) = \Pr\,(X_i = x, R_i = r | G_i = g) / \sum_{x' \in \mathcal{X}} \Pr\,(X_i = x', R_i = r | G_i = g)$. The first term is unknown but models the party registration as a function of demographics, geolocation, and race. To estimate this model and obtain a maximum likelihood estimate of individual race via equation (9), we use the standard Expectation-Maximization algorithm by treating race as missing data (Dempster, Laird, and Rubin 1977) (see Supplementary Appendix A.2 for details).

3 Empirical Validation

In this section, we present an empirical validation study of the methods described above and assess the accuracy of their prediction relative to that of the existing methods.

[7] Technically, the assumption given in equation (8) can be slightly relaxed using the following set of sequential independence assumptions, although in our empirical study they do not appear to make substantial differences:

$$X_i \perp\!\!\!\perp S_i \,|\, R_i, G_i$$

$$P_i \perp\!\!\!\perp S_i \,|\, R_i, G_i, X_i.$$

3.1 *Data*

We analyze voter registration data from Florida, which include approximately ten million individual records. Our data are based on statewide voter files and come from L2 (formerly Labels & Lists, Inc.), a leading nonpartisan firm and the oldest organization in the United States that supplies voter data and related technology to candidates, political parties, pollsters, and consultants for use in campaigns. For every active registered voter in the state, we have gender, birth date, original registration date, address, district, precinct, party registration, and turnout history.[8]

We also use the 2010 U.S. Census Summary File for Florida, which contains the joint distribution of individual characteristics, including age, gender, and race, at the levels of various geographical units, including blocks, tracts, and precincts. The summary file contains raw counts of individuals, which we aggregate by various geographical units and then use to calculate $\Pr(G_i|R_i = r)$ and $\Pr(G_i, X_i|R_i = r)$. As explained in more detail in Barber and Imai (2013), we geocode voters in the L2 data using their addresses so that we know the geographical unit to which each voter belongs. We also verify that the Census data accurately reflect the racial composition of voting precincts in the L2 data (see Figure 3 in Supplementary Appendix A.4).

The Census Bureau also provides data on the racial distribution of surnames in the United States. In 2007, the Census Bureau released the percent of individuals who are White, Black, Latino, Asian, and so on for each surname occurring at least 100 times in the 2000 Census. The list contains a total of 151,671 names, capturing 90% of the population enumerated in the 2000 Census. We supplement this list with Census's Spanish Surname List, which contains 12,500 common Latino surnames, about half of which are on the 2007 Census Surname List. From this data, we can calculate $\Pr(R_i = r|S_i)$ for well over 150,000 surnames in the U.S. See Supplementary Appendix A.1 for details.

We divide race into five categories: White, Black, Latino, Asian, and Other. These are similar to the racial groups used in the Census data and self-reported race in the voter files. The major difference is that we do group American Indian/Alaska Native with Other, because American Indians and Alaska Natives jointly constitute less than 1% of records in the Florida voter file. Moreover, we find that our misclassification rate is approximately equal among the American Indian/Alaska Native and Other groups.[9]

3.2 *Validation of Race Predictions*

To validate the proposed methodology, we compare the race predictions from each method with voters' self-reported race, which is available for approximately nine million voters in Florida. For each voter, we find the race with the greatest predicted probability and classify the voter as belonging to that racial group. The goal of this validation exercise is to examine whether and how additional information, such as geolocation and party registration, improves the race predictions.

We assess the performance of each method by calculating the overall error rate, which simply represents the proportion of voters whose racial group it incorrectly classifies. We also compute the two types of group-specific error rates: false positives (Type I errors) and false negatives (Type II errors). For example, with respect to Latinos, classifying a non-Latino voter as Latino would be a false positive, whereas classifying a Latino voter as non-Latino would be a false negative. Although the goal is to minimize both types of error, there is a clear trade-off between the two.

Table 1 displays the error rates for five sets of predictions based on different sets of information. We begin with a name-only prediction that classifies race on the basis of the Census Surname List. We then enhance the prediction by incorporating voters' geolocation, testing both voting precinct

[8]The data contain all active registrants as of July 2012. L2 removes voters who were classified as inactive by the Secretary of State's Office. Inactive voters are those who did not vote in the past several elections or respond to an official request to confirm their address and registration. See Barber and Imai (2013) for details. Replication files are available at http://dx.doi.org/10.7910/DVN/SVY5VF.

[9]We combine the Census Mixed Race category with Other, because our voter files do not have a separate mixed-race category. However, in theory, researchers may use Census data to identify the growing mixed-race population, which is over nine million or 2.9% of the U.S. population in 2010.

Table 1 Empirical validation of individual-level race classification using the Florida registration records

		Name	*Name Precinct*	*Name Block*	*Name Precinct Party*	*Name Block Party*
Overall error rate		0.215	0.158	0.152	0.151	0.145
White (68%)	False negative	0.047	0.060	0.059	0.065	0.061
	False positive	0.523	0.294	0.266	0.257	0.237
Black (13%)	False negative	0.839	0.381	0.320	0.290	0.249
	False positive	0.011	0.027	0.026	0.033	0.029
Latino (13%)	False negative	0.193	0.150	0.155	0.158	0.162
	False positive	0.037	0.039	0.038	0.038	0.037
Asian (2%)	False negative	0.540	0.519	0.533	0.520	0.532
	False positive	0.006	0.007	0.007	0.007	0.007
Other (4%)	False negative	0.991	0.989	0.969	0.989	0.968
	False positive	0.001	0.001	0.002	0.001	0.002

Notes: The table displays the overall classification error rate as well as false negative (Type I error) and false positive (Type II error) rates for White, Black, Latino, Asian, and Other voters using our proposed prediction method. We classify each registered voter to the racial category with the greatest predicted probability. Each column corresponds to the results based on different sets of information. We start with the information based on the Census Surname List only and then add the voter's geolocation and party registration. The total sample size is 9,247,810.

and Census block. Finally, we include voters' party registration as an individual-level covariate. We use publicly available Gallup polling data to obtain the distribution of partisanship by race, that is, $\Pr(P_i = p|R_i = r)$ (Newport 2013).

The first row of Table 1 displays each prediction method's overall classification error rate, measuring the accuracy of each prediction across all voters. We find that the additional information reduces the overall error rate from approximately 22%, which is obtained when only voters' names are used, to 15% when their geolocation and party registration are incorporated. In particular, the prediction based on voters' name, block, and party registration performs best according to this measure. We also find that using demographics does not substantially change our predictions. In addition, our second method of incorporating party registration, which does not require external data on the distribution of partisanship by race, performs slightly worse than the ones presented here (see Table 4 in Supplementary Appendix A.5 for a full set of results).

We further examine the performance of the proposed methodology for each racial category. Among Whites, the name-only prediction results in a substantially high false positive rate of over 50%. Incorporating voters' geolocation and party registration, we are able to reduce this to approximately 25% without substantially increasing the false negative rate. Among Blacks, the false negative rate for the name-only prediction exceeds 80%, while incorporating additional information reduces this by more than half. In both cases, adding party registration as well as geolocation appears to be beneficial.[10]

For Latinos and Asians, the improvement in accuracy due to the additional information appears to be minimal. Among Latinos, the name-only prediction already has a relatively low false negative rate of about 19%.[11] Indeed, incorporating voters' geolocation and party registration further decreases the false negative rate, but only by three to four percentage points. Among Asians, who consist of only 2% of Florida registered voters, there is little performance difference across the

[10]We examined the self-reported race of voters we incorrectly classified as Whites (i.e., false positives). We find that voters misclassified as Whites are 50% Blacks, 18% Latinos, 7% Asians, and 25% Others. Among Black voters who are misclassified (i.e., false negatives), 94% are misclassified as Whites.

[11]We examined whether using the Spanish Surname List helps identify Latinos. We find that whether or not we use this list in conjunction with the full Census Surname List, our accuracy among Latinos remains nearly identical. We recomputed the Name and Precinct and the Name, Precinct, and Party predictions without using the Spanish Surname List and obtained the almost same overall error rate and false negative and positive rates as we report in Table 1. We suspect that this is because the Census Surname List contains many prominent Spanish surnames.

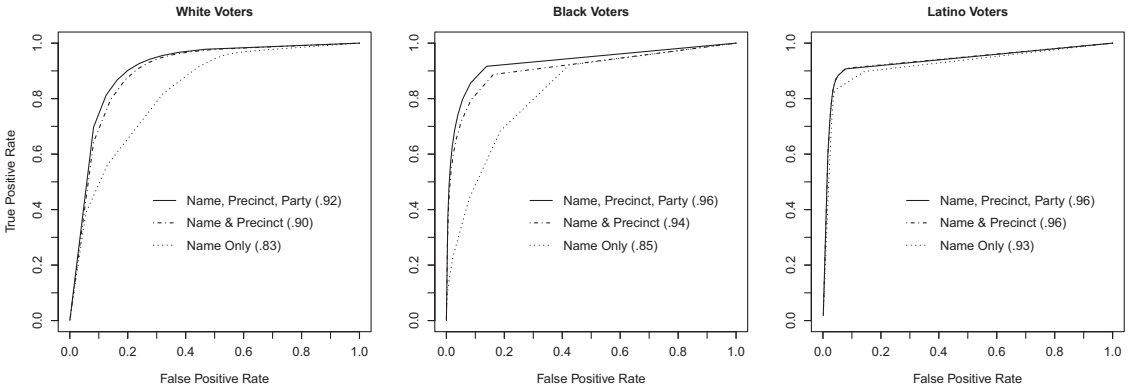

Fig. 1 ROC curves for the proposed race prediction methods. ROC curves plot true positive rate (vertical axis) against false positive rate (horizontal axis) for all possible thresholds used for classification. The area under the ROC curves, given in the legend, summarizes the overall classification success. Among White and Black voters, using voter precinct (denoted as "Precinct") in addition to surname ("Name") substantially improves classification accuracy. Adding voter party registration ("Party") results in further improvements. Among Latino voters, surname alone yields a high success rate and adding other information produces minor improvements.

methods. All methods have a high false negative rate, suggesting that it is difficult to identify Asian voters from the set of information considered in this article alone.[12]

A more comprehensive comparison of predictions, while recognizing the trade-off between false negatives and false positives, is to examine the receiver operating characteristic (ROC) curve for each prediction method. Rather than classifying voters on the basis of the greatest predicted probability, ROC curves display the true positive rate (sensitivity) against the false positive rate (specificity) for a variety of classification thresholds. Since it is desirable to have a higher true positive rate given a false negative rate (or a lower false negative rate given a true positive rate), the area under the ROC curve can be used to evaluate the performance.

In Figure 1, we plot ROC curves for three predictions among White, Black, and Latino voters. Among Whites and Blacks, we observe that the information about voters' geolocation significantly improves the accuracy of race prediction while adding the party registration yields only a modest improvement. Among Latinos, as we saw earlier, the name-only prediction performs relatively well. The figure shows that it is possible to reduce the false negative rate among Blacks and Latinos to 0.06 and 0.03, respectively, while maintaining the true positive rate above 0.8. This means that our method correctly classifies over 80% of Blacks and Latinos, while only misclassifying 6% of non-Blacks as Black and 3% of non-Latinos as Latino.

3.3 *Validation of the Turnout Estimates*

We now estimate voter turnout by racial category and validate our estimates against actual turnout by race at the precinct and congressional district levels in Florida. The goal is to investigate whether individual-level racial predictions improve the race-specific turnout rates obtained from the standard ecological inference techniques widely used in academia and elsewhere (i.e., Goodman 1953; King 1997).

We focus on turnout among White, Black, Latino, Asian, and Other registered voters in the 2008 presidential election. We estimate aggregate turnout for each racial group using the predicted probabilities directly. Specifically, we calculate the aggregate turnout for each race as the weighted average of turnout, where the predicted probabilities serve as weights. Formally, for each racial group r, we compute $\sum_{i=1}^{n} \Pr(R_i = r|S_i, G_i, P_i) Y_i / \sum_{i=1}^{n} \Pr(R_i = r|S_i, G_i, P_i)$, where Y_i is the binary

[12] As was the case with Black voters, the vast majority of false negatives among Latinos (86%) and Asians (75%) are misclassified as Whites.

Table 2 Bias and RMSE of predicted turnout by race across 8,828 precincts and 25 congressional districts in Florida

	Goodman's regression		*King's EI*		*Name-only prediction*		*Bayesian prediction*	
	Bias	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*
Precincts								
Whites	0.003	0.069	0.041	0.062	−0.003	0.015	−0.003	0.012
Blacks	−0.102	0.162	−0.133	0.217	−0.009	0.043	−0.007	0.039
Latinos	−0.114	0.251	−0.163	0.250	0.016	0.042	0.011	0.035
Asians	0.017	0.713	−0.470	0.550	0.041	0.116	0.040	0.111
Others	−0.214	0.499	−0.338	0.450	0.068	0.109	0.048	0.094
Districts								
Whites	0.008	0.037	0.047	0.058	−0.007	0.012	−0.001	0.004
Blacks	−0.147	0.197	−0.215	0.267	0.009	0.020	−0.006	0.010
Latinos	−0.272	0.463	−0.300	0.354	0.045	0.052	0.017	0.021
Asians	0.072	0.808	−0.459	0.530	0.055	0.058	0.043	0.046
Others	−0.229	0.527	−0.342	0.448	0.073	0.078	0.042	0.053

Notes: Goodman's regression, King's EI, name-only prediction (based on the Census Surname List), and our proposed Bayesian prediction method. Although Goodman's regression and King's EI use precinct-level turnout and racial composition data only, the proposed Bayesian methodology uses the name, residence location, and party registration of voters. Precinct-level bias and RMSE are weighted by the number of voters for each precinct. Generally, the proposed Bayesian method performs best, though the name-only prediction also yields a reasonable performance.

turnout variable for voter i. For the purpose of comparison, we also compute the prediction based on the Census surname alone and compute $\sum_{i=1}^{n} \Pr(R_i = r|S_i) Y_i / \sum_{i=1}^{n} \Pr(R_i = r|S_i)$.

We validate our estimates against true precinct-level and district-level turnout, which can be computed using the self-reported race for each voter. In addition to the name-only prediction, we compare the performance of our methodology against the two standard ecological inference techniques, that is, Goodman's ecological regression (Goodman 1953), and the King's EI (King 1997). Goodman's method regresses overall turnout on the proportion of voters of a particular race to estimate turnout for that race. The method assumes that the average turnout rate for each racial group does not depend on racial composition. We fit Goodman's ecological regression using precinct-level data in each congressional district. We fit a separate univariate model for each of the five racial groups.[13] This yields the estimates of turnout by race that can be used at both the precinct and district levels. The second standard technique is King's EI, which yields precinct-level turnout estimates (King and Roberts 2012). We fit a separate 2×2 EI model for each racial group, one district at a time. We then aggregate the estimated turnout among precincts within a district to estimate district-level turnout.

Table 2 reports the bias and RMSE of turnout estimates at both the precinct and district levels for each method. We begin by considering the two standard techniques. Goodman's regression does not perform well, underestimating turnout among Blacks, Latinos, and Others by over ten percentage points at the precinct level on average. The bias increases at the district level. Moreover, the RMSE is large for all groups but Whites. King's EI also performs poorly at the precinct and district levels, yielding large bias and RMSE. It is particularly biased for Others, underestimating turnout by over thirty percentage points on average.[14]

The name-only prediction and the proposed Bayesian approach significantly improve the results of the aforementioned standard methods. Both have much smaller bias and RMSE. In general, the proposed Bayesian methodology performs best, providing essentially unbiased estimates for Whites, Blacks, and Latinos. The magnitude of bias is somewhat larger for Asians and Others,

[13] We also fit a multivariate linear regression, regressing the overall turnout on the proportions of all racial groups. These results are substantively similar to the univariate results presented here (see Table 5 in Supplementary Appendix A.5).

[14] We also examine the performance of these methods in racially homogeneous precincts (defined as having over 90% of one race). In our data, the vast majority (92%) of such precincts are homogeneously White. The Bayesian predictions significantly outperform the other methods (see Table 6 in Supplementary Appendix A.5).

but is still less than five percentage points. In Table 5 provided in Supplementary Appendix A.5, we also present the results based on the name-only and Bayesian classifications, which classify each voter to a racial group and then aggregate turnout. As expected, these methods, which do not incorporate the uncertainty in the predictions, perform slightly worse than the corresponding methods presented here.

The name-only prediction does surprisingly well despite the fact that its classification error rate is greater than that of the Bayesian method. Indeed, the performance of the name-only prediction method is roughly comparable to that of the Bayesian method. This apparent inconsistency can occur because the turnout rate is approximately equal among false negative and false positive voters. That is, the classification error based on the Census Surname List is roughly independent of turnout (see Table 7 in Supplementary Appendix A.5). However, in other settings, such independence may not hold. As such, we recommend that applied researchers and litigators use the proposed Bayesian methodology.

4 Concluding Remarks

This article reviews and extends the methodology for predicting the race of an individual by incorporating name, geocoded residence, and other information from voter files. Our validation study has shown that the proposed Bayesian methodology provides accurate individual-level predictions and significantly improves the estimation of aggregate-level turnout for each racial group relative to the standard ecological inference methods. We believe that this methodology enables academic researchers and litigators to conduct more reliable ecological inference in states where registered voters are not asked to report their race. A straightforward and yet useful extension of the proposed methodology is to incorporate vote choice from survey data for predicting candidate choice as well as turnout by racial groups.

References

Ansolabehere, S., and E. Hersh. 2003. Gender, age, race, and voting: A research note. *Politics and Governance* 1(2):132–37.

Barber, M., and K. Imai. 2013. Estimating neighborhood effects on turnout from geocoded voter registration records. Working Paper available at http://imai.princeton.edu/research/neighbor.html (accessed February 24, 2016).

Barreto, M. A. 2007. Si Se Puede! Latino candidates and the mobilization of Latino voters. *American Political Science Review* 101(3):425–41.

Barreto, M. A., G. M. Segura, and N. D. Woods. 2004. Mobilizing effect of majority-minority districts. *American Political Science Review* 98(1):65–75.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* 39(1):1–37.

Elliott, M. N., A. Fremont, P. A. Morrison, P. Pantoja, and N. Lurie. 2008. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Services Research* 43(5p1):1772–36.

Elliott, M. N., P. A. Morrison, A. Fremont, D. F. McCaffrey, P. Pantoja, and N. Lurie. 2009. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology* 9(2):69–83.

Enos, R. D. 2015. Testing the elusive: A field experiment on intergroup competition and voting. Working Paper, Department of Government, Harvard University.

Fieldhouse, E., and D. Cutts. 2008. Diversity, density and turnout: The effect of neighbourhood ethno-religious composition on voter turnout in Britain. *Political Geography* 27(5):530–48.

Fiscella, K., and A. M. Fremont. 2006. Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research* 41(4p1):1482–500.

Fraga, B. 2013. Winning the race, losing the base? Demobilization, competitiveness, and electoral influence. Working Paper, Department of Political Science, Indiana University.

———. 2016. Candidates or districts? Reevaluating the role of race in voter turnout. *American Journal of Political Science* 60(1):97–122.

Gay, C. 2001. The effect of black congressional representation on political participation. *American Political Science Review* 95(3):589–602.

Goodman, L. 1953. Ecological regressions and behavior of individuals. *American Sociological Review* 18:663–66.

Greiner, J. D. 2007. Ecological inference in voting rights act disputes: Where are we now, and where do we want to be? *Jurimetrics* 47(2):115–67.

Greiner, D. J., and K. M. Quinn. 2008. R × c ecological inference: bounds, correlations, flexibility and transparency of assumptions. *Journal of the Royal Statistical Society, Series A* 172(1):67–81.

———. 2010. Exit polling and racial bloc voting: Combining individual level and ecological data. *Annals of Applied Statistics* 4(4):1774–96.

Hajnal, Z. L., and J. Trounstine. 2005. Where turnout matters: The consequences of uneven turnout in city politics. *Journal of Politics* 67(2):515–35.

Harris, J. A. 2015. What's in a name? A method for extracting information about ethnicity from names. *Political Analysis* 23(2):212–24.

Henderson, J. A., J. S. Sekhon, and R. Titiunik. 2014. Cause or effect? Turnout in Hispanic majority-minority districts. Department of Political Science, University of California, Berkeley.

Herron, M. C., and J. S. Sekhon. 2005. Black candidates and black voters: Assessing the impact of candidate race on uncounted vote rates. *Journal of Politics* 67(1):154–77.

Imai, K., Y. Lu, and A. Strauss. 2008. Bayesian and likelihood inference for 2 × 2 ecological tables: An incomplete data approach. *Political Analysis* 16(1):41–69.

Khanna, K., and K. Imai. 2016. Replication data for: Improving ecological inference by predicting individual ethnicity from voter registration records. http://dx.doi.org/10.7910/DVN/SVY5VF, Harvard Dataverse, V1 (accessed February 24, 2016).

King, G., 2004. *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton, NJ: Princeton University Press.

King, G., and M. Roberts. 2012. Ei: A(n r) program for ecological inference.

King, G., O. Rosen, and M. Tanner, eds. 2004. *Ecological inference: New methodological strategies*. Cambridge: Cambridge University Press.

Michelson, M. R. 2003. Getting out the Latino vote: How door-to-door canvassing influences voter turnout in rural central California. *Political Behavior* 25(3):247–63.

Newport, F. 2013. Democrats racially diverse; Republicans mostly white—Gallup. http://www.gallup.com/poll/160373/democrats-racially-diverse-republicans-mostly-white.aspx (accessed January 24, 2015).

Tam Cho, W. K., J. G. Gimpel, and J. J. Dyck. 2006. Residential concentration, political socialization, and voter turnout. *Journal of Politics* 68(1):156–67.

Wakefield, J. 2004. Ecological inference for 2 × 2 tables (with discussion). *Journal of the Royal Statistical Society, Series A* 167:385–445.