BENCHMARKING BAYESIAN IMPROVED SURNAME GEOCODING
AGAINST MACHINE LEARNING METHODS

A PREPRINT

Ari Decter-Frain
Jeb E. Brooks School of Public Policy
Cornell University
Ithaca, NY, USA
agd75@cornell.edu

June 30, 2022

ABSTRACT

Bayesian Improved Surname Geocoding (BISG) is the most popular method for proxying race/ethnicity in voter registration files that do not contain it. This paper benchmarks BISG against a range of previously untested machine learning alternatives, using voter files with self-reported race/ethnicity from California, Florida, North Carolina, and Georgia. This analysis yields three key findings. First, when given the exact same inputs, BISG and machine learning perform similarly for estimating aggregate racial/ethnic composition. Second, machine learning outperforms BISG at individual classification of race/ethnicity. Third, the performance of all methods varies substantially across states. These results suggest that pre-trained machine learning models are preferable to BISG for individual classification. Furthermore, mixed results at the precinct level and across states underscore the need for researchers to empirically validate their chosen race/ethnicity proxy in their populations of interest.

Keywords  Imputation Methods · Machine Learning · Computational social science · Bayesian imputation

1 Introduction

Research in political science and social science often requires constructing a race/ethnicity proxy variable. Constructing such a proxy is an important step for conducting ecological inference in voting rights litigation (Barreto et al. [2019], Imai and Khanna [2016]), redistricting (DeLuca and Curiel [2022], Kenny et al. [2021]), and substantive research on the role of race in politics (Enos [2016], Enos et al. [2019], Grumbach and Sahn [2020]). The most common method for imputing race is Bayesian Improved Surname Geocoding (BISG), which uses Bayes’ rule to compute a probability distribution over race-ethnicity categories conditional on a voter’s surname and where they live (Elliott et al. [2008, 2009]). BISG and its variations have attained widespread popularity for race-ethnicity imputation due to their parsimony, computational efficiency, and superior performance when compared to existing alternatives, namely spatial interpolation of Census racial-ethnic composition from Census geographies (Imai and Khanna [2016], Clark et al. [2021], Shah and Davis [2017], Decter-Frain et al. [2023]).

While BISG performs well compared to commonly used alternatives, it has not yet been benchmarked against machine learning (ML) models that can produce race/ethnicity predictions from more flexible and potentially more accurate models. In this paper I present the results of such an exercise. I train a range of machine learning models using voter registration data from Florida, Georgia, North Carolina, and a portion of California where voters self-report their race/ethnicity upon registration. The registries in these states contain over 26 million labelled observations, which roughly equates to a five percent non-representative sample of the United States electorate. I then compare BISG against predictions from these models made out-of-state. I find that machine learning models consistently outperform BISG at
individual classification. However, for estimating precinct racial composition – a task crucial for ecological inference and redistricting work, BISG often performs as well or better than any of the ML alternative.

Machine learning leverages abundant data and computational power to identify high-dimensional relationships between predictors and minimize predictive loss (Hastie et al. [2001]). These methods have become increasingly common for prediction problems in the social sciences, including in sociology (Molina and Garip [2019]) and economics (Varian [2014]). The standard suite of ML models range in complexity from straightforward multinomial regressions to gradient-boosted decision trees.

Training an ML model involves learning a set of unknown parameters to optimize some objective error function given a) the structure of the model and b) a set of observations for which the outcome is observed. In the case of proxying for race/ethnicity, there exist many millions of observations from states where voters do self-report their race/ethnicity, which enables training models with high complexity and many unknown parameters. In contrast, BISG uses a straightforward (and reasonable) model for combining inputs to generating predicted probabilities without leveraging existing labelled data.

However, using ML to proxy race/ethnicity does carry a major risk of over-fitting to the characteristics of states for which labelled data is available. Existing research on ML for race-ethnicity classification has achieved exceptional performance when models are trained in the same geographic regions where they are tested (Sood and Laohaprapanon [2018], Lee et al. [2017]), but political scientists and practitioners do not need to impute the race/ethnicity of voters in the regions where race/ethnicity is already self-reported – they need to impute it in all other geographic regions of the United States.

To ensure a fair comparison between BISG and ML while accounting for the risk of over-fitting, I apply a maximally conservative out-of-sample evaluation strategy. Throughout this paper, all supervised models are trained using data from three states and used to generate predictions in the fourth. Thus, I apply these models in a different context with a different cultural and institutional history than where they were trained.

The potential for error due to over-fitting in this setting is further amplified by the relatively extreme racial and ethnic diversity of these four states. Georgia and North Carolina have among the largest Black population shares in the country, while California has the largest Asian share. Furthermore, although Florida and California are among the states with the largest Hispanic population share, their Hispanic populations are quite different. The largest inflows of Hispanics to Florida come from Cuba (41 percent), while Californian Hispanics are predominantly of Mexican origin (84 percent). To the extent that surname frequency differs across nationalities, this could pose challenges for models fit using surnames to predict race/ethnicity.

When compared against ML models tested out of state, BISG performs well but is not always the clear best option. For individual classification, ML models consistently outperform it in all four states for most race/ethnicity groups. When incorporating more information into the model (first and middle names), the ML models ubiquitously outperform BISG in terms of errors, and generally appear better-calibrated. In contrast, I find more mixed results for performance when estimating aggregate racial composition, with the preferred model varying depending on the race/ethnicity and state of interest.

The results of these analyses indicate that BISG is still an effective tool for developing a race/ethnicity proxy variable. However, it is not always the optimal approach, especially when proxying at the individual level. Furthermore, this paper presents the largest multi-state evaluation of BISG thus far, and reveals that the performance of BISG and ML alternatives varies between geographic regions. Ultimately, researchers should always to validate and compare methods using a labelled sample from their local population of interest.

---

1 Raw voter data used in this paper cannot be shared directly. An R package enabling researchers to produce race/ethnicity predictions from pre-trained machine learning models is in development and will be available should the article be accepted for publication.

2 With the exception of Hawaii. Typically, people self-reporting in the Census category Native Hawaiian/Pacific Islanders are included under “Asian” (Imai and Khaman[2016]).

3 https://www.pewresearch.org/hispanic/2004/03/19/latinos-in-california-texas-new-york-florida-and-new-jersey/
2 Methods for imputing race/ethnicity

2.1 BISG and extensions

BISG computes the probability of race given a voter’s surname and geographic location through an application of Bayes’ theorem. Assuming \( P(G|R) \parallel P(S|R) \),

\[
P(R = r | S = s, G = g) \propto P(G = g | R = r) P(R = r | S = s)
\]  

(1)

The probability \( P(G = g | R = r) \) can be obtained from Census summary tables by taking the number of people of race \( R = r \) in neighborhood \( G = g \) divided by the total number of people of race \( r \). Recent research has shown that the imputations become more accurate as the geographic areas used to define \( G \) becomes smaller. Most performance gains are achieved by matching people to geographic areas by zip code, although the best results are still obtained by geocoding the data and matching voters to blocks ([Clark et al., 2021]). In this paper I use block-level geographic information from 2020 Decennial Census throughout. The probability of race given surname, \( P(R = r | S = s) \), comes directly from the Census Bureau’s surname lists which contain the proportion of all Decennial Census respondents with each surname in each racial-ethnic category. I use the `merge_surnames` function from the wrus package to match voters with the correct \( P(R|S) \) ([Khanna and Imai, 2021]). I then compute posterior probability \( P(R|G, S) \) by multiplying the two quantities together for each race category and dividing by the sum across all race categories to normalize. For comparison, I also use the wrus package to perform BISG, and find that the results of my calculations match closely to the package (see appendices A and B).

Previous work to extend BISG has involved layering on additional assumptions that enable the inclusion of more information within the framework of the Bayes’ rule formula. Given a series of conditional independence assumptions, various extensions can be made by altering equation 1,

\[
P(R = r | S = s, G = g, X = x, Y = y, ...) \propto P(G = g | R = r) (X = x | R = r) (Y = y | R = r) ... P(R = r | S = s)
\]  

(2)

Where \( X, Y, \) and so on are any voter characteristics whose probability conditional on race is computed from available data. Previous work has taken this approach to add first names ([Voicu, 2018]), age, gender, and party affiliation ([Imai and Khanna, 2016]), each yielding better performance than when using surname and geography alone.

Here, in addition to BISG, I also consider an extended method that combines information about geographic, surname, middle name and first name using [2]. For each, I construct four separate lookup tables containing their probability given race-ethnicity. Each of the four tables uses data from three of the four states. Then, to compute estimates in any given state, I use the probabilities from the lookup table that does not contain that state. Therefore, extensions of BISG make use of data from out-of-state.

2.2 Machine learning methods

In equations [1] and [2] every term is known and no parameters must be estimated prior to computing a prediction. This is an advantage of BISG – predictions follow directly from an application of Bayes Theorem and so require no information beyond indirect data from the Census Bureau. However, data do exist with which to estimate unknown parameters in a more flexible or detailed model of race/ethnicity. This fact motivates experimenting with a series of more complex models, as described below. Here, I provide a brief overview of each of the methods I apply in the paper, in rough order of complexity.

2.2.1 Multinomial Logistic Regression

A straightforward way to model probability of race/ethnicity given a series of observed characteristics is with a multinomial logistic regression (MLR). Keeping with the previous notation, we can assume that

\[
\log \frac{Pr(R = r | S = s, G = g)}{Pr(R = \text{not } r | S = s, G = g)} = \sum_{i=1}^{5} \left( \alpha_{ri} P(G = g | R = i) + \beta_{ri} P(R = i | S = s) \right)
\]  

(3)

[Imai and Khanna, 2016] make a slightly different assumption that is weaker and adds the need for an expectation-maximization step to the computation of the posterior distribution. They note that in their empirical investigations the choice of their approach over the one described here does not make a major difference for performance.

[For a full treatment of supervised learning methods, see Hastie et al., 2001]
Where \( r \) and \( i \) separately index the five race/ethnicity categories. This is a basic multinomial regression specification with ten total predictors, which are exactly the same ten probabilities used to generate predictions from BISG. Instead of combining two of these probabilities to generate predicted probabilities for each race/ethnicity (as in equation 1), we input all ten probabilities as predictors of all five possible outcomes. Equation 3 contains two sets of 25 unknown parameters, \( \alpha_{ri} \) and \( \beta_{ri} \). These parameters are learned by fitting the model with labelled data from Florida, Georgia, North Carolina, and California. Training involves finding the set of parameters that minimizes a likelihood function, \( L \), for the inputted data. After training, one can use the learned parameters generate predicted probabilities of each race/ethnicity given the corresponding inputs.

Clearly, the MLR framework can accommodate additional predictors beyond the ten used for standard BISG, including probabilities based on other characteristics like first and middle name, along entirely distinct predictors like ‘bag-of-letters’ counts of the number of each letter that appears in a voter’s full name, or additional characteristics of a voter’s neighborhood, like its median income, average education level, and so on. For brevity and to match previous work on BISG, I only consider geographic and name-based information here. Much more complex sets of features are possible, though, and can be implemented in the future.

### 2.2.2 Regularization

Fitting an MLR model with many predictors for race/ethnicity imputation may lead to poor performance when these models are applied out of state, because highly parameterized models can fit too closely to the characteristics of the training states. Regularization ‘encourages’ weights placed on each predictor to tend towards zero. It works by modifying the likelihood equation, \( L \), minimized to solve MLR models. Here I use elasticnet regularization. The likelihood gains additional terms,

\[
L + \lambda \left[ (1 - \delta) \sum \Psi^2 + \delta \sum |\Psi| \right],
\]

where

\[
\Psi = \alpha_{1,1}, \beta_{1,1}, \ldots, \alpha_{5,5}, \beta_{5,5}
\]

Alongside the typical multinomial likelihood function, an additional term adds to its output a magnitude proportional to the magnitude of all the estimated coefficients. This has the effect of depressing coefficients towards zero. The squared penalty and absolute value penalty on \( \Omega \) are called the Ridge and Lasso penalties. In elasticnet, both are used and a tuning parameter \( \delta \) determines the weight placed on each. An additional tuning parameter, \( \lambda \), determines the weight placed on the overall penalty term. I use an implementation of elasticnet that performs cross-validation during training to optimize these two hyperparameters.

### 2.3 Random Forests

Whereas the previous two methods involve weighted summations of predictors, the next two are based on classification trees. A classification tree successively splits observations into buckets according to cutoffs in the parameter space, and assigns every observation to the majority class in their final bucket. Splits are made successively in a greedy manner, with each made to minimize classification error in the training set. Random forests are an ensemble of multiple classification trees, where each tree is constructed using a subset of all predictors and training observations.

### 2.4 Gradient Boosting

Gradient tree boosting implements a greedy construction of the forest, wherein for each training iteration \( t \), a number of candidate decision-trees, \( f_t(x_i) \), are constructed and one tree is selected to minimize the loss, \( \mathcal{L}^{(t)} \), from the previous iteration,

\[
\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t),
\]

where \( \hat{y}_i^{(t-1)} \) is the predicted value from the previous iteration \[Chen and Guestrin\,2016\]. In other words, each subsequent tree is selected to minimize the residuals from the ensemble in the previous iteration. \( \Omega(f_t) \) is an additional regularization term.
3 Data and methods

The data used for this analysis are voter registration records from Florida, Georgia, North Carolina, and California. The Florida and Georgia data were collected after the 2018 national election, North Carolina after the 2016 national election, and California after the 2020 election. For Florida, North Carolina, and Georgia, the data contain all registered voters and had already been geocoded as part of previous projects. For California, only roughly twenty percent of the voter file contains self-reported race-ethnicity. These data were obtained in raw format and roughly eighty percent of them were successfully geocoded using the Census Bureau’s geocoding API. The incompleteness of the California data should not meaningfully impact the conclusions in this paper, since my focus is on comparing methods for race-ethnicity imputation rather than downstream applications.

|                  | Florida | Georgia | N. Carolina | California |
|------------------|---------|---------|-------------|------------|
| N. obs           | 8,008,186 | 6,302,094 | 6,347,235   | 5,289,420  |
| Prp. White       | .687    | .527    | .684        | .442       |
| Prp. Black       | .134    | .301    | .209        | .052       |
| Prp. Hispanic    | .139    | .034    | .028        | .232       |
| Prp. Asian       | .017    | .024    | .013        | .159       |
| Prp. Other       | .023    | .114    | .066        | .115       |
| Mean Age         | 55.2    | 45.6    | 49.0        | 41.8       |
| Prp. Male        | .446    | .465    | .442        | .094       |
| Prp. Female      | .541    | .533    | .525        | .083       |
| Prp. Unknown     | .013    | .001    | .032        | .822       |

Table 1: Self-reported demographic characteristics of voters in Florida, Georgia, North Carolina, and a subset of voters from California.

Table 1 presents the aggregate demographic characteristics of the voters in each state after pre-processing. For all four states, the data contain voters’ first, middle, and last names, along with their gender, age, and self-reported race-ethnicity. The options for self-reporting race-ethnicity in all four states are White, Black, Hispanic, Asian, Other, and Unknown. For each dataset, voters without unique IDs and who could not be successfully geocoded were removed. All voters whose race was categorized as missing or ‘Unknown’ were also removed. I identified voters’ Census blocks of residence by using the sf package in R [Pebesma, 2018] to conduct spatial joins to 2020 block geographic regions using their geolocation.

I trained all model types described in section 2.2. I used the tidymodels framework in R to tune, train, and evaluate the models [Kuhn and Wickham, 2020]. For each model type, I tuned and trained four separate models, each completely withholding data from one state. The model tuning procedure involved constructing a Latin hypercube grid of possible hyperparameter combinations and iterating over the grid using five-fold cross-validation to identify the best-performing combination. For this tuning procedure I used a random sample of the 100 thousand voters from the data. Then, using the best-performing set of hyperparameters from the tuning procedure, I trained each model on a one million-voter sample. I repeated this procedure twice using two different sets of predictors. First, I used only the ten probabilistic inputs to a basic BISG-based prediction. Second, I used an expanded set that also includes probabilistic inputs based on voters’ first and middle names. Finally, I used the trained models to generate predictions for all voters in each held-out state. I present an evaluation of these models compared to BISG in the next section.

4 Out-of-state validation

This section compares the performance of different race/ethnicity proxy methods at the individual and aggregate level based on an out-of-state validation exercise. I tuned and trained each model four times, each time leaving out data from one of the four states. Using each of these models I generated predictions in the fourth, held-out state, and evaluated these predictions against self-reported race/ethnicity. I then compared the performance of each method against BISG.

6The author received all data from Dr. Matt Barreto and Dr. Loren Collingwood. The data had previously been geocoded for use in consulting and other research projects, including Barreto et al. [2019] and Decter-Frain et al. [2022].
all supervised models are evaluated out-of-state. For example, wherever performance is reported for a supervised learning model in Florida, that model was trained using voter data from California, Georgia, and North Carolina, before being used to make predictions in Florida.

In the following subsections I present figures which summarize performance differences between BISG and ML models in general. For a more detailed comparison between models, appendices A and B contain tables with the specific values underlying these figures. From these appendices, two results are worth mentioning here. First, results using my implementation of BISG are similar to results when using the popular wru package.

Second, both BISG and ML approaches struggle with proxying the ‘other’ race/ethnicity category. For this class, area under the curve for individual classification ranges between 0.5 and 0.7 for most models, compared to .8 to 1 for the other race/ethnicities. The ‘other’ category may be difficult to model because it contains quite a diverse group of individuals, including those who identify with more than one race. Researchers typically tolerate poor performance on the ‘other’ category because the group represents less than one percent of the population and because most research and applied questions typically pertain to the more well-defined groups.

4.1 Classifying individual voters

I first examined the overall performance of each method at classifying voter race/ethnicity. Figure 1 presents the area under the receiver operation curve (AUC) for all models by state, race, and set of inputs. Using a minimal set of inputs \( P(G|R) \) and \( P(R|S) \), BISG’s performance compared to ML appears mixed, sometimes outperforming ML and sometimes not. When using the larger pool of inputs, ML models consistently outperform BISG.

![Figure 1: Comparison of area under the receiver operating curve (AUC) between BISG and a range of ML methods. Each x represents a different ML method, while the vertical line represents BISG. Left panels use probabilistic information about block and surname only. Right panels use these, plus information about first name and middle name. Machine learning models are trained on three states and used to generate predictions in the fourth.](image)

Notably, when using this larger set of inputs all ML methods cluster together and attain similar performance. This suggests that simple models should suffice to obtain these performance improvements over BISG for individual classification.

Next, I examined the calibration of BISG and ML approaches. Figure 2 plots calibration curves for each method and each combination of state and race. A well-calibrated model will fall close the identity line. Being under the identity line means that too much probability is assigned to a given race/ethnicity, and being above it means not enough probability has been assigned.

Insufficient calibration has fairness implications [Corbett-Davies and Goel, 2018]. Assigning excess probability to one race/ethnicity may lead to over-weighting that group in downstream analyses, and vice versa for under-calibration.
For a given application, researchers may seek to achieve the best calibration possible, or may seek the method that minimizes over- or under-calibration for particular groups of interest.

In most cases, BISG appears equally or worse-calibrated than ML models, particularly for Asian and Hispanic voters. The one case where BISG appears better calibrated is for Asian voters in California. Both models do not assign enough probability to Californian voters being Asian, but the supervised model is the more poorly-calibrated of the two. This poor calibration likely results from the relatively extreme size of the Asian population in California. Supervised models trained on other states are partially informed by the base rates of each race/ethnicity in each state, and California has the largest Asian share of any state in the country.

![Calibration curves](image)

Figure 2: Calibration curves plotting the observed proportion of voters self-reporting with each race-ethnicity as a function of the predicted probabilities from each model. Both models use the full set of input data. The diagonal line indicates perfect calibration such that for any predicted probability, the actual proportion of voters with self-identifying with that race-ethnicity equals the predicted probability. All points calculated by taking the mean within bins equal to one tenth of the probability space.

In general, the calibration curves for the two methods appear on the same side of the identity line. The exception is for Hispanics in Florida, where BISG is over-calibrated and the supervised method is under-calibrated. It is unclear why BISG appears over-calibrated, though the under-calibration of the supervised method is likely again due to differences in the base-rate share of Hispanics in Florida compared to the other states under study.

Lastly, both methods perform very poorly at predicting when individuals will self-report in the ‘other’ race/ethnicity group. Figure 2 suggests that regardless of what probability either model assigns to an individual self-reporting as ‘other’, the true probability remains close to zero. This group is challenging because it combines small minority
race/ethnicities (eg. Native American), voters who identify with multiple race/ethnicities, and any additional voters who choose not to identify as one of the main four race/ethnicities. Given the diverse range of reasons why a voter may select this category, it is perhaps unsurprising that no clear signal emerges to identify them.

4.2 Group racial composition

Next, I evaluated the performance of supervised methods at estimating the racial composition of a geographic area. Although many applications require analyzing voter data at the precinct level, I follow [Kenny et al., 2021] in conducting composition evaluations at the tract level. Tracts are advantageous over precincts from an analytical perspective because they are of roughly equal size and are always nested inside counties and states. Furthermore, results at the tract level should be highly consistent with those at the precinct level.

Figure 3 presents root mean-squared error (RMSE) for estimating the tract-level share of each race/ethnicity in each state. In Georgia and North Carolina, results are consistent with those for individual classification. For Florida and California, however, results appear more mixed. After including first and last name information, BISG remains the best or one of the best methods for estimating racial composition in Florida, and in certain race/ethnic categories in California.

Figure 4 presents the average bias for estimating the share of each race/ethnicity in each state. Again, in Georgia and North Carolina, ML nearly always improves on BISG once first and middle names are included. BISG is substantially less biased than ML for estimating white population share in Florida and Asian and Black population share in California. Notably, the biases of both BISG and ML methods differ in magnitude and direction across states. In Georgia, BISG overestimates the white share of tracts by, on average, almost seven percent. Meanwhile in the same bias in North Carolina is only four percent and near zero in Florida.

5 Discussion

In this paper, I conducted the first direct comparison between BISG and ML methods for constructing a race/ethnicity proxy variable in voter registration data, using data from four states and evaluating the ML models on out-of-state performance. The exercise yielded three novel results. First, BISG and machine learning perform similarly for estimating aggregate racial/ethnic composition. Second, machine learning outperforms BISG at individual classification of race/ethnicity. Third, the performance of all methods varies substantially across states.
Figure 4: Bias values for various ML methods compared against BISG. Bias was computed weighting each tract by its population size. Each x represents a different ML method, while the vertical line represents BISG. Left panels use probabilistic information about block and surname only. Right panels use these, plus information about first name and middle name. Machine learning models are trained on three states and used to generate predictions in the fourth.

The first two points have clear implications for future research. When a researcher’s goal is to estimating the racial composition of a neighborhood or precinct, the evidence presented here suggests BISG is among the best tools to do so. While it does not perform best for every race/ethnicity in all of the tested states, it performs comparably or better in many instances. For individual prediction, provided researchers have access to voters’ first and middle names, they should use ML. In this study, ML consistently produced more accurate imputations than BISG in every state for every race/ethnicity. Furthermore, ML methods were better-calibrated in every case except for Asians in California.

The result that all methods perform differently across states is equally important. To the author’s knowledge, the four states used here represent the largest evaluation set against which BISG has been validated thus far. Notable differences emerge between states. For instance, in Florida, estimates of the white share of the electorate appear mostly unbiased (bias = -.005). Meanwhile, BISG overestimates the white share of the North Carolina electorate by four percent on average, and by 7.5 and 7.9 percent in California and Georgia, respectively. As another illustration, the RMSE for estimating the Hispanic share of the electorate in California is more that twice what it is in any other state.

Thresholds for tolerating bias and error in race/ethnicity proxying will vary across applications. Nonetheless, it is notable that published validations of BISG typically use data from one or two states, while substantive applications apply it in other regions (Enos [2016]) or even nationwide (Grumbach and Sahn [2020]). Ideally, researchers applying any method for proxying race/ethnicity should first validate their chosen method using a small sample from the local population to ensure it is accurate and well-calibrated.
References

Matt Barreto, Loren Collingwood, Sergio Garcia-Rios, and Kassra AR Oskooii. Estimating Candidate Support in Voting Rights Act Cases: Comparing Iterative EI and EI-RxC Methods. Sociological Methods & Research, page 004912411985239, July 2019. ISSN 0049-1241, 1552-8294. doi:10.1177/0049124119852394 URL https://journals.sagepub.com/doi/10.1177/0049124119852394

Kosuke Imai and Kabir Khanna. Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records. Political Analysis, 24(2):263–272, 2016. ISSN 1047-1987, 1476-4989. doi:10.1093/pan/mpw001 URL https://www.cambridge.org/core/product/identifier/S1047198700010962/type/journal_article

Kevin DeLuca and John A. Curiel. Validating the Applicability of Bayesian Inference with Surname and Geocoding to Congressional Redistricting. Political Analysis, pages 1–7, 2022.

Christopher T. Kenny, Shiro Kuriwaki, Cory McCartan, Evan T. R. Rosenman, Tyler Simko, and Kosuke Imai. The use of differential privacy for census data and its impact on redistricting: The case of the 2020 U.S. Census. Science Advances, 7(41):eabk3283, October 2021. ISSN 2375-2548. doi:10.1126/sciadv.abk3283 URL https://www.science.org/doi/10.1126/sciadv.abk3283

Ryan D. Enos. What the Demolition of Public Housing Teaches Us about the Impact of Racial Threat on Political Behavior. American Journal of Political Science, 60(1):123–142, January 2016. ISSN 0092-5853, 1540-5907. doi:10.1111/ajps.12156 URL https://onlinelibrary.wiley.com/doi/10.1111/ajps.12156

Ryan D. Enos, Aaron R. Kaufman, and Melissa L. Sands. Can Violent Protest Change Local Policy Support? Evidence from the Aftermath of the 1992 Los Angeles Riot. American Political Science Review, 113(4):1012–1028, November 2019. ISSN 0003-0554, 1537-5943. doi:10.1017/S0003055419000340 URL https://www.cambridge.org/core/product/identifier/S00030554190000340/type/journal_article

Jacob M. Grumbach and Alexander Sahn. Race and Representation in Campaign Finance. American Political Science Review, 114(1):206–221, February 2020. ISSN 0003-0554, 1537-5943. doi:10.1017/S0003055419000607 URL https://www.cambridge.org/core/product/identifier/S0003055419000607/type/journal_article

Marc N. Elliott, Allen Fremont, Peter A. Morrison, Philip Pantajo, and Nicole Lurie. A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity: New Method for Estimating Race/Ethnicity and Associated Disparities. Health Services Research, 43(5p1):1722–1736, May 2008. ISSN 0017-9124, 1475-6773. doi:10.1111/j.1475-6773.2008.00854.x URL http://link.springer.com/10.1111/j.1475-6773.2008.00854.x

Marc N. Elliott, Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantajo, and Nicole Lurie. 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, June 2009. ISSN 1387-3741, 1572-9400. doi:10.1007/s10742-009-0047-1 URL http://link.springer.com/10.1007/s10742-009-0047-1

Jesse T. Clark, John A. Curiel, and Tyler S. Steelman. Minmaxing of Bayesian Improved Surname Geocoding and Geography Level Ups in Predicting Race. Political Analysis, pages 1–7, November 2021. ISSN 1047-1987, 1476-4989. doi:10.1017/pan.2021.31 URL https://www.cambridge.org/core/product/identifier/S1047198721000310/type/journal_article

Paru R. Shah and Nicholas R. Davis. Comparing Three Methods of Measuring Race/Ethnicity. The Journal of Race, Ethnicity, and Politics, 2(1):124–139, March 2017. ISSN 2056-6085. doi:10.1017/rep.2016.27 URL https://www.cambridge.org/core/product/identifier/S2056608516000271/type/journal_article

Ari Decter-Frain, Pratik Sachdeva, Loren Collingwood, Juandalyn Burke, Hickari Murayama, Matt Barreto, Scott Henderson, Spencer A Wood, and Joshua Zingher. Comparing Methods for Estimating Demographics in Racially Polarized Voting Analyses. preprint, SocArXiv, April 2022. URL https://osf.io/e854z

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Mario Molina and Filiz Garip. Machine Learning for Sociology. Annual Review of Sociology, 45(1):22, 2019.

Hal R. Varian. Big Data: New Tricks for Econometrics. Journal of Economic Perspectives, 28(2):3–28, May 2014. ISSN 0895-3309. doi:10.1257/jep.28.2.3 URL https://pubs.aeaweb.org/doi/10.1257/jep.28.2.3

Gaurav Sood and Suriyan Laohaprapanon. Predicting Race and Ethnicity From the Sequence of Characters in a Name. arXiv:1805.02109 [stat], May 2018. URL http://arxiv.org/abs/1805.02109 arXiv: 1805.02109

Jinhyuk Lee, Hyunjae Kim, Miyoung Ko, Donghee Choi, Jaehoon Choi, and Jaewoo Kang. Name Nationality Classification with Recurrent Neural Networks. In Proceedings of the Twenty-Sixth International Joint Conference
Kabir Khanna and Kosuke Imai. *wru: Who are You? Bayesian Prediction of Racial Category Using Surname and Geolocation*, 2021. URL [https://CRAN.R-project.org/package=wru](https://CRAN.R-project.org/package=wru). R package version 0.1-12.

Ioan Voicu. Using First Name Information to Improve Race and Ethnicity Classification. *Statistics and Public Policy*, 5(1):1–13, January 2018. ISSN 2330-443X. doi:10.1080/2330443X.2018.1427012 URL [https://www.tandfonline.com/doi/full/10.1080/2330443X.2018.1427012](https://www.tandfonline.com/doi/full/10.1080/2330443X.2018.1427012).

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco California USA, August 2016. ACM. ISBN 978-1-4503-4232-2. doi:10.1145/2939672.2939785 URL [https://dl.acm.org/doi/10.1145/2939672.2939785](https://dl.acm.org/doi/10.1145/2939672.2939785).

Edzer Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446, 2018. doi:10.32614/RJ-2018-009 URL [https://doi.org/10.32614/RJ-2018-009](https://doi.org/10.32614/RJ-2018-009).

Max Kuhn and Hadley Wickham. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*, 2020. URL [https://www.tidymodels.org](https://www.tidymodels.org).

Sam Corbett-Davies and Sharad Goel. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning, August 2018. URL [http://arxiv.org/abs/1808.00023](http://arxiv.org/abs/1808.00023). Number: arXiv:1808.00023 arXiv:1808.00023 [cs].
A  Full individual-level results
| State       | Race/Eth | WRU | BISG | Block + Surname | Logit | Logit2 | ELNET | RF | GB | BISG | Logit | Logit2 | ELNET | RF | GB |
|-------------|----------|-----|------|----------------|-------|--------|-------|----|----|------|-------|--------|-------|----|----|
| California  | White    | 0.893 | 0.897 | 0.875 | 0.873 | 0.882 | 0.887 | 0.892 | 0.911 | 0.921 | 0.921 | 0.921 | 0.920 | 0.917 |
|             | Black    | 0.878 | 0.891 | 0.905 | 0.895 | 0.909 | 0.920 | 0.928 | 0.900 | 0.942 | 0.945 | 0.941 | 0.943 | 0.946 |
|             | Hispanic | 0.915 | 0.921 | 0.903 | 0.905 | 0.905 | 0.911 | 0.916 | 0.929 | 0.940 | 0.941 | 0.938 | 0.937 | 0.938 |
|             | Asian    | 0.916 | 0.931 | 0.918 | 0.910 | 0.922 | 0.917 | 0.936 | 0.911 | 0.949 | 0.953 | 0.951 | 0.951 | 0.942 |
|             | Other    | 0.527 | 0.524 | 0.508 | 0.512 | 0.513 | 0.508 | 0.512 | 0.547 | 0.556 | 0.538 | 0.554 | 0.537 | 0.546 |
| Florida     | White    | 0.899 | 0.903 | 0.877 | 0.862 | 0.883 | 0.903 | 0.904 | 0.926 | 0.942 | 0.942 | 0.942 | 0.941 | 0.939 |
|             | Black    | 0.909 | 0.911 | 0.896 | 0.877 | 0.902 | 0.933 | 0.932 | 0.918 | 0.954 | 0.957 | 0.955 | 0.957 | 0.954 |
|             | Hispanic | 0.932 | 0.940 | 0.932 | 0.930 | 0.936 | 0.939 | 0.939 | 0.947 | 0.972 | 0.972 | 0.972 | 0.971 | 0.968 |
|             | Asian    | 0.788 | 0.784 | 0.832 | 0.844 | 0.844 | 0.857 | 0.858 | 0.809 | 0.916 | 0.929 | 0.921 | 0.932 | 0.923 |
|             | Other    | 0.575 | 0.576 | 0.591 | 0.593 | 0.605 | 0.580 | 0.587 | 0.591 | 0.662 | 0.646 | 0.665 | 0.644 | 0.633 |
| Georgia     | White    | 0.867 | 0.871 | 0.830 | 0.793 | 0.837 | 0.869 | 0.873 | 0.895 | 0.906 | 0.907 | 0.906 | 0.907 | 0.907 |
|             | Black    | 0.904 | 0.908 | 0.859 | 0.815 | 0.869 | 0.906 | 0.909 | 0.920 | 0.933 | 0.934 | 0.933 | 0.934 | 0.934 |
|             | Hispanic | 0.873 | 0.895 | 0.921 | 0.929 | 0.922 | 0.927 | 0.929 | 0.915 | 0.962 | 0.964 | 0.962 | 0.961 | 0.963 |
|             | Asian    | 0.878 | 0.899 | 0.915 | 0.927 | 0.920 | 0.926 | 0.929 | 0.908 | 0.960 | 0.964 | 0.958 | 0.962 | 0.963 |
|             | Other    | 0.531 | 0.525 | 0.542 | 0.555 | 0.544 | 0.550 | 0.550 | 0.540 | 0.566 | 0.577 | 0.561 | 0.572 | 0.569 |
| N. Carolina | White    | 0.853 | 0.855 | 0.822 | 0.790 | 0.831 | 0.859 | 0.865 | 0.887 | 0.905 | 0.906 | 0.906 | 0.906 | 0.905 |
|             | Black    | 0.889 | 0.887 | 0.852 | 0.821 | 0.862 | 0.901 | 0.905 | 0.908 | 0.935 | 0.940 | 0.935 | 0.939 | 0.940 |
|             | Hispanic | 0.833 | 0.853 | 0.910 | 0.912 | 0.911 | 0.913 | 0.920 | 0.875 | 0.949 | 0.952 | 0.949 | 0.953 | 0.952 |
|             | Asian    | 0.817 | 0.840 | 0.911 | 0.915 | 0.911 | 0.919 | 0.927 | 0.857 | 0.954 | 0.961 | 0.954 | 0.963 | 0.962 |
|             | Other    | 0.571 | 0.555 | 0.587 | 0.594 | 0.596 | 0.595 | 0.605 | 0.569 | 0.600 | 0.627 | 0.599 | 0.643 | 0.634 |

Table 2: Area under the curve (AUC) for each model for each race in each state.
B Full precinct-level results
| State      | Race/Eth | BISG | WRU | ELNET | Logit | Logit2 | RF | GB | + First + Middle |
|------------|----------|------|-----|-------|-------|--------|----|----|------------------|
|            |          |      |     |       |       |        |    |    |                  |
| California | White    | .087 | .1  | .121  | .121  | .136   | .107| .112| .088             |
|            | Black    | .036 | .042| .066  | .064  | .063   | .053| .051| .029             |
|            | Hispanic | .070 | .072| .055  | .051  | .056   | .050| .050| .064             |
|            | Asian    | .039 | .044| .117  | .102  | .106   | .103| .091| .041             |
|            | Other    | .113 | .113| .083  | .082  | .083   | .084| .088| .105             |
|            |          |      |     |       |       |        |    |    | .037             |
| Florida    | White    | .045 | .049| .143  | .148  | .162   | .094| .075| .037             |
|            | Black    | .038 | .046| .101  | .106  | .125   | .039| .035| .035             |
|            | Hispanic | .032 | .029| .074  | .077  | .078   | .066| .048| .029             |
|            | Asian    | .018 | .017| .024  | .021  | .024   | .021| .023| .010             |
|            | Other    | .016 | .012| .075  | .075  | .075   | .090| .063| .015             |
| Georgia    | White    | .097 | .115| .130  | .141  | .184   | .070| .078| .087             |
|            | Black    | .059 | .077| .120  | .132  | .177   | .056| .052| .034             |
|            | Hispanic | .046 | .045| .030  | .039  | .023   | .021| .023| .026             |
|            | Asian    | .025 | .024| .012  | .014  | .013   | .009| .011| .017             |
|            | Other    | .092 | .092| .067  | .065  | .063   | .056| .068| .092             |
| N. Carolina| White    | .044 | .057| .109  | .120  | .149   | .069| .058| .051             |
|            | Black    | .041 | .051| .093  | .104  | .129   | .038| .033| .034             |
|            | Hispanic | .039 | .035| .022  | .022  | .020   | .019| .017| .017             |
|            | Asian    | .024 | .023| .016  | .017  | .015   | .013| .012| .014             |
|            | Other    | .046 | .049| .050  | .052  | .051   | .050| .047| .044             |

Table 3: RMSE values for every model-race-state combination. All calculations weight by tract size.
| State     | Race/Eth | Block + Surname | + First + Middle |
|-----------|----------|-----------------|-----------------|
|           | BISG     | WRU  | ELNET | Logit | Logit2 | RF | GB | BISG | ELNET | Logit | Logit2 | RF | GB |
| California | White    | 0.070 | 0.079 | 0.085 | 0.079 | 0.093 | 0.095 | 0.100 | 0.075 | 0.072 | 0.062 | 0.059 | 0.061 | 0.078 |
|           | Black    | 0.018 | 0.016 | 0.058 | 0.050 | 0.043 | 0.042 | 0.039 | 0.013 | 0.030 | 0.028 | 0.021 | 0.035 | 0.026 |
|           | Hispanic | 0.031 | 0.027 | -0.015 | -0.012 | -0.016 | -0.014 | -0.012 | 0.025 | -0.002 | 0.000 | 0.004 | 0.000 | 0.007 |
|           | Asian    | -0.029 | -0.032 | -0.082 | -0.072 | -0.075 | -0.076 | -0.070 | -0.032 | -0.065 | -0.057 | -0.048 | -0.066 | -0.047 |
|           | Other    | -0.091 | -0.090 | -0.046 | -0.045 | -0.045 | -0.047 | -0.057 | -0.081 | -0.036 | -0.033 | -0.035 | -0.031 | -0.064 |
| Florida   | White    | -0.013 | 0.002 | -0.084 | -0.081 | -0.081 | -0.079 | -0.061 | -0.005 | -0.051 | -0.049 | -0.046 | -0.064 | -0.032 |
|           | Black    | -0.013 | -0.018 | 0.014 | 0.019 | 0.014 | 0.002 | -0.004 | -0.015 | -0.008 | -0.009 | -0.012 | -0.011 | -0.016 |
|           | Hispanic | 0.014 | 0.008 | -0.025 | -0.027 | -0.026 | -0.026 | -0.015 | 0.015 | -0.017 | -0.017 | -0.013 | -0.021 | -0.010 |
|           | Asian    | 0.009 | 0.009 | 0.021 | 0.018 | 0.020 | 0.018 | 0.020 | 0.003 | 0.006 | 0.005 | 0.004 | 0.010 | 0.006 |
|           | Other    | 0.002 | -0.001 | 0.073 | 0.072 | 0.073 | 0.086 | 0.060 | 0.002 | 0.070 | 0.070 | 0.067 | 0.086 | 0.051 |
| Georgia   | White    | 0.086 | 0.100 | 0.082 | 0.085 | 0.078 | 0.056 | 0.068 | 0.079 | 0.033 | 0.033 | 0.034 | 0.018 | 0.039 |
|           | Black    | -0.041 | -0.049 | -0.060 | -0.069 | -0.058 | -0.034 | -0.035 | -0.016 | 0.017 | 0.014 | 0.013 | 0.009 | 0.012 |
|           | Hispanic | 0.030 | 0.025 | 0.023 | 0.024 | 0.022 | 0.018 | 0.019 | 0.016 | 0.010 | 0.009 | 0.006 | 0.010 | 0.009 |
|           | Asian    | 0.010 | 0.009 | 0.010 | 0.010 | 0.012 | 0.005 | 0.009 | 0.006 | 0.002 | 0.004 | 0.002 | 0.004 | 0.005 |
|           | Other    | -0.085 | -0.085 | -0.056 | -0.051 | -0.053 | -0.046 | -0.060 | -0.085 | -0.062 | -0.060 | -0.055 | -0.041 | -0.065 |
| N. Carolina | White   | 0.023 | 0.035 | -0.042 | -0.045 | -0.045 | -0.050 | -0.038 | 0.040 | 0.001 | 0.001 | 0.003 | 0.016 | 0.001 |
|           | Black    | -0.025 | -0.028 | -0.005 | -0.004 | 0.000 | 0.000 | 0.001 | -0.021 | -0.022 | -0.022 | -0.018 | -0.018 | -0.016 |
|           | Hispanic | 0.028 | 0.023 | 0.021 | 0.022 | 0.020 | 0.018 | 0.015 | 0.012 | 0.009 | 0.009 | 0.005 | 0.009 | 0.004 |
|           | Asian    | 0.010 | 0.010 | 0.015 | 0.016 | 0.014 | 0.012 | 0.009 | 0.004 | 0.007 | 0.007 | 0.004 | 0.008 | 0.004 |
|           | Other    | -0.035 | -0.040 | 0.011 | 0.012 | 0.011 | 0.021 | 0.013 | -0.035 | 0.005 | 0.005 | 0.006 | 0.017 | 0.009 |

Table 4: Bias values for every model-race-state combination. All calculations weight by tract size.