Improved Bayesian Ethnorace Prediction

Bertrand Wilden

Last updated on 10 March, 2021

Abstract

Quantitative social science research on race and ethnicity can be constrained by the lack of available individual-level data with these markers. In response to this issue, methods have been developed to predict individuals' race and ethnicity using Bayes' rule. Racial distributions from the US Census Surname List are combined with distributions from geolocations, such as Census blocks or zip codes, to yield predicted probabilities of individuals' race and ethnicity. I expand upon existing methods by incorporating information from a nationwide list of first names as well as from residence characteristics. Along with some other adjustments, these improvements lead to substantial gains in predictive performance when validated against official state voter files. The largest of these predictive gains are found for African Americans—a group which existing methods find difficult to predict accurately.

Introduction

Research on racial and ethnic disparities in the United States can be constrained by the lack of available individual-level data with these markers (Trounstine 2020; Kuk, Hajnal, and Lajevardi 2020; Burch 2013). This is particularly true when racial geography is a key aspect of the research design. In these contexts, researchers are often forced to use aggregate-level data on group proportions, such as county statistics, to draw inferences. But this method is susceptible to the modifiable areal unit problem, whereby statistical bias is introduced due to arbitrary and unequal geographic unit sizes (Fotheringham and Wong 1991). Furthermore, some research designs all but require individual-level data. Geographic regression discontinuity—an increasingly common causal inference technique—typically relies on individually geocoded addresses (Keele and Titiunik 2018; Velez and Newman 2019; Cantoni 2020). Therefore if race or ethnicity are central elements of the research question, individual-level identifiers for these categories are likely necessary.

To overcome some of these challenges, I describe a method for imputing individual ethnoraces directly. This method uses Bayes' rule to make predictions by combining information from nationwide distributions of six ethnorace categories over other characteristics, such as names, geographies, political party identification, and others. My implementation builds off existing methods which use a similar prediction algorithm (Elliott et al. 2009; Imai and Khanna 2016; Voicu 2018). The method described by Imai and Khanna (2016) in particular, and the associated `R` package `wru`, has become popular in recent studies on race and ethnicity. Its application has been used in work on racial protests and voting patterns (Enos, Kaufman, and Sands 2019), disparities in campaign financing (Grumbach and Sahn 2020; Grumbach, Sahn, and Staszak 2020), the impact of electoral institutions on local representation (Abott and Magazinnik 2020), and public health issues such as suicide rates (Studdert et al. 2020).

My ethnorace prediction method improves upon Imai and Khanna (2016) in several ways. Whereas their method only takes as inputs distributions over surnames, geolocation, political party, age, and gender, my implementation adds information from a nationwide list of first names (Tzioumis 2018) as well as address characteristics. Additionally, I incorporate insights from the machine learning literature to further improve predictive performance. The result of these modifications is a substantial increase in predictive power compared to Imai and Khanna (2016) in validation tests. These predictive gains are particularly strong in regards to correctly classifying African American and Hispanic individuals—especially in contexts where fine-grain geolocation data are unavailable. My method is available in an easy-to-use `R` package `bper`.[1]

In the next section I provide some background and specifics regarding the inputs and outputs of my ethnorace prediction method. Then I explain the methodology and compare my implementation with previous versions. Finally, I demonstrate the predictive performance of my method when validated against the combined North Carolina and Florida voter file (n = 21,164,503).

Data

[1] `bper`: Bayesian Prediction for Ethnicity and Race. https://github.com/bwilden/bper

Outputs

Before discussing the methodology further, I want to first define what "predicting ethnicity or race" means. These are categories which, although relatively immutable compared to other identities, do not have universally accepted delineations and meanings (Omi and Winant 2014). I follow the convention from previous ethnorace prediction methods by using the US Census Bureau categorizations (Elliott et al. 2009; Imai and Khanna 2016; Voicu 2018). In this framework, individuals can be classified as non-Hispanic White, non-Hispanic Black or African American, non-Hispanic Asian and Pacific Islander, non-Hispanic American Indian and Alaska Native, Hispanic or Latino alone, and non-Hispanic Other Race.[2] Because Hispanic identity is defined by the Census, and understood commonly, as an ethnicity, rather than a race, I use the term "ethnorace" in this paper to refer to any of the previously-mentioned groups.

There are a few benefits to using the Census ethnorace categorization. These definitions capture a common understanding of race and ethnicity in the US, and correspond to the groups found most frequently in social science research. The data sources of these groups' distributions that serve as inputs to the prediction formula also rely on the Census categorization. This also facilitates comparison of my method against previous ethnorace prediction methods.[3] One downside to using the Census categories, however, is that it obscures substantial heterogeneity that may exist within each group. Within Asian Americans and Latinos, for example, there is considerable variation in terms of national ancestry. Furthermore, the unfortunate necessity of a catch-all Other Race category ensures that important sources of diversity are lost.[4]

Inputs

First Names

The first names list I use comes from Tzioumis (2017). It is drawn from mortgage applications and contains ethnorace counts in each of the six groups across 4,250 first names. Unlike Census

[2] Non-Hispanic Other Race includes individuals who identify as belonging to two or more race/ethnicities, as well as those who may not identify with the other Census categories.

[3] Unlike my method, Imai and Khanna (2016) do not include a separate category for American Indian and Alaska Native. In order to create similar comparison groups, I recode all predicted American Indian and Alaska Native individuals as Other Race during the validation exercises. If desired, however, the `bper` package will produce predicted probabilities for the American Indian and Alaska Native category.

[4] This is hinted at empirically by the method's poor predictive performance for the Other Race category.

data, which form the basis for much of my other data sources, this list of first names may be unrepresentative of the larger US population. To the extent that first name distributions differ by ethnorace given employment status, for example, this may be a concern. But the predictive benefits from using first name data, as I will demonstrate, likely overwhelm these worries in most contexts.

Last Names

For my last names data, I use the 2010 Census Surnames List.[5] This list comes from the 2010 decennial Census and contains over 160,000 common US last names (those occurring 100 or more times in the population). Like the first names list, these data include counts of individuals in each of the six ethnorace categories across each last name.

Geolocations

My ethnorace distributions by geographies come from the 2010 decennial Census, accessed via IPUMS NHGIS.[6] In decreasing order of mean population, these geographies include *state*, *county*, *Census place*, *ZIP code*, and *Census block*. Predictions tend to improve with more precise levels of geography. With this in mind, my implementation automatically matches each individual to the most fine-grain level of geography available.

Party Identification

My party identification data come from a 2012 Gallup poll.[7] The three categories of political party I include are Republican, Democrat, and Other (including Independents and "don't knows"). The Gallup report tells me both the probability that an individual with a given ethnorace belongs to a particular political party, and the probability that an individual with a given political party identifies with a particular ethnorace.

[5] https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

[6] Steven Manson, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 15.0 [dataset]. Minneapolis, MN: IPUMS. 2020. http://doi.org/10.18128/D050.V15.0

[7] https://news.gallup.com/poll/160373/democrats-racially-diverse-republicans-mostly-white.aspx

Age and Gender

Like my geolocation data, age and gender distributions come from the 2010 decennial Census, accessed via IPUMS NHGIS. These variables do not contain much predictive power in terms of ethnoracial classification, but nevertheless, I find that their inclusion in the algorithm helps slightly.

Multi-unit Address

These data refer to ethnorace distributions over multi-unit housing occupancy. Individuals are matched to these probabilities if their address contains "Apt", "Unit", "#", or other such identifier. Unfortunately, I was not able to find these distributions for the 2010 decennial Census, so instead I use those from the year 2000.

Data Structure

The raw data sources I describe above, with the exception of party ID,[8] all contain counts of individuals with a particular attribute (i.e. the first name JOHN, or the ZIP Code "92092") per ethnorace category. Taking proportions by cell across a given attribute tells us $Pr(Ethnorace|Attribute)$, and taking proportions by cell across a given ethnorace group tells us $Pr(Attribute|Ethnorace)$. These two conditional probabilities form the building blocks of the classification algorithm described later.

If any cell in the input data is empty (i.e. if there are no individuals of a particular ethnorace with some attribute), then the conditional probabilities $Pr(Ethnorace|Attribute)$ and $Pr(Attribute|Ethnorace)$ will be zero. As will become clear in the Methodology section, if either of those two probabilities for an individual equal zero for a given ethnorace, the algorithm will predict a zero percent probability that the individual belongs to that ethnorace. This will occur even if some other attributes about that individual predict a high probability of belonging to the ethnorace. For example, an individual could have first and last names that are highly predictive of being Hispanic, but reside in a Census block which had zero Hispanic occupants at the time of the 2010 decennial Census. Blocks typically contain only around 400 individuals—so this is a real possibility. For this hypothetical person the input data claims that $Pr(Hispanic|CensusBlock) = 0$, which yields $Pr(Hispanic) = 0$ due to the structure of the prediction algorithm. To resolve this

[8] Party ID percentages by ethnorace are directly available in the Gallup report.

issue, I apply a technique from the machine learning literature known as Laplace smoothing to my input data. This works by adding some constant, or psuedocount, to number of individuals in every cell in the input data, then calculating the conditional probabilities $Pr(Ethnorace|Attribute)$ and $Pr(Attribute|Ethnorace)$.[9] Through validation tests, I found that Laplace smoothing led to significant gains in predictive performance for Asian individuals in particular. I also conjecture that, absent this smoothing technique, the algorithm's predictions are too beholden to the specifics of the 2010 decennial Census. The predictions will generalize better to other time periods without the rigid assumptions of zero conditional probability for some attribute/ethnorace combinations.

Methodology

In general terms, the method computes predicted probabilities for each of the six aforementioned ethnorace categories for each individual. Then, each individual is classified into the category corresponding to the highest predicted probability. These predicted probabilities can be stated more formally as the conditional probability of identifying as a particular ethnorace for an individual with a particular profile of first name, last name, geolocation, party ID, age, gender, and address type. Bayes' rule provides a template for how to answer this sort of conditional probability problem.

$$Pr(R = r|X) = \frac{Pr(X|R = r)Pr(R = r)}{Pr(X)} \qquad (1)$$

Where R is an individual's true ethnorace, r is one of six possible ethnorace categories (White, Black, Asian, Native American, Hispanic, or Other race), and X is the joint probability of an individual having a particular profile of attributes (first name, last name, geolocation, party ID, age, gender, and address type). Unfortunately, the joint probability X in Equation is intractable due to both data constraints and the astronomically large number of combinations of possible attribute profiles. If however, we assume conditional independence of ethnorace among each attribute in X, we can rewrite Equation (1) in terms of less complex conditional probabilities:

[9] Missing counts are only a problem in the first names, last names, and geolocation data so psuedocounts are only added for those inputs. The exact value for the Laplace smoothing psuedocount could be any number greater than zero, but through out-of-sample validation I found 5 to be the optimal value.

$$Pr(R = r|X) = \frac{Pr(R = r|x') \prod_{j=1}^{6} Pr(x_j|R = r)}{\sum_{i=1}^{6} Pr(R = r_i|x') \prod_{j=1}^{6} Pr(x_j|R = r_i)} \tag{2}$$

Where x is the vector of individual attributes indexed by j. The particular attribute x' comes from using the chain rule to decompose the joint probability $Pr(R = r, X)$. The choice of which attribute to use for x' is atheoretical, but all previous prediction methods have used last names (Elliott et al. 2009; Imai and Khanna 2016; Voicu 2018). During my validation exercises, I found that the choice of x' has potentially large consequences for predictive performance. For example, using last names for x' appears to help predictions of Whites—but to the detriment of non-Whites. In light of these trade-offs, my method cycles through every attribute as the choice of x' and computes $Pr(R = r|X)$ for each. These posterior probabilities are then averaged within each ethnoracial category to generate final predicted probabilities that an individual belongs to a particular ethnorace. The result is more balanced predictions across each ethnorace.

The conditional independence assumption necessary for transforming equation (1) to (2) says that knowing both a particular attribute of an individual, and that individual's ethnorace, should give us no extra knowledge of any other attribute for that individual. Stated formally, $Pr(x_j|R = r, X) = Pr(x_j|R = r)$ for all x_j. This assumption is almost certainly violated in the present context. One example that has been demonstrated empirically is that last name distributions by race vary across regions in the US (Crabtree and Chykina 2018).

Violations of the conditional independence assumption are commonplace in most applications of similar classification algorithms. Nevertheless, these prediction methods perform unreasonably well in many contexts (Lewis 1998; Domingos and Pazzani 1997; Rish 2001). This is likely because of the decision rule governing the final classifications—the posterior probabilities of the true class do not have to necessarily be statistically valid, they only need to be higher than those of every other class to be accurately classified. As is true with my method, when more attribute inputs are added to the model (first names, multi-unit occupancy) researchers should be cautious trying to interpret the posterior probabilities $Pr(R = r|X)$ directly. Rather, the maximum a posteriori ethnorace classifications should be used alone.

Validation

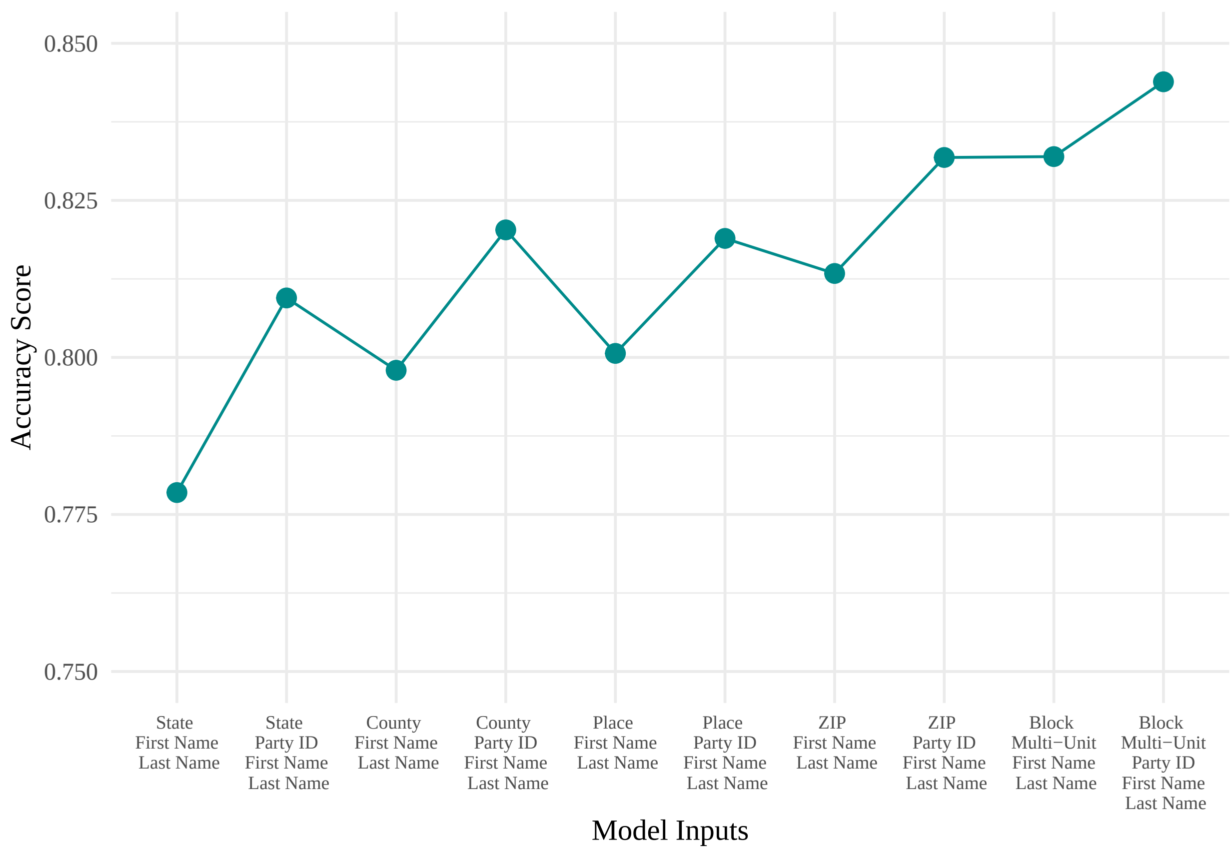
To test the performance of the model, I apply the predictions to the combined North Carolina and Florida State voter file. These files contain snapshots of the registered voters in their respective states and provide individual-level data for first names, last names, address, political party, age, gender, and crucially self-identified ethnorace. After combining the two voter files, I then geocoded each unique address in the sample. This allowed me to match individual observations to Census places and blocks, and ZIP codes. Then I applied the prediction algorithm described above using the `bper` package and calculated each individuals' predicted ethnorace. In order to compare my method against an existing benchmark, I also used the `wru` package (Imai and Khanna 2016) to calculate ethnorace predictions for the same individuals.

Unlike typical machine learning techniques, my prediction method does not fit the model on some subsample, or training set, from these data and then compare predictions against a held-out test set. Instead, the conditional probabilities for each attribute and ethnorace described above are merged into voter file from the input data sources. This allows the entire voter file to be used for validation. And because the input data come from nationally representative samples, the risk of overfitting due to any peculiarities of the North Carolina/Florida electorate are minimized. Together, these two states represent 21,164,503 individuals. Compared to nationwide percentages, Florida has a higher proportion of Hispanics and North Carolina has a higher proportion of African Americans. When combined they form a reasonably ethnoracially diverse population—2% Asian, 16% Black, 11% Hispanic, 6% Other Race, 65% White.

The most straightforward metric for assessing individual-level predictive performance is the Accuracy score, or Overall Error Rate. This number is the proportion of correctly classified individuals in the sample. I ran models separately for different combinations of input variables to mimic data availability constraints in real-world applications, and then calculated the Accuracy score for each. Figure 1 displays a summary of the results of these different models.

As expected, the model with the greatest number of input data sources, and at the most precise geographic level, on the far right of the figure performs the best in terms of overall Accuracy. Using Census blocks, multi-unit address, party ID, first names, and last names, this model correctly identifies the ethnorace of 84.4% of individuals in the sample. The upward trend in model accuracy

Figure 1: Accuracy Scores by Input Data



seen in the figure corresponds to shrinking the size of the geolocation variable used. Moving from state to county, from county to place, from place to ZIP code, and from ZIP code to block all improve the overall predictive performance in terms of overall Accuracy. I include the input for multi-unit occupancy only for the Census block models because I believe this reflects the practical contexts where `bper` might be used. If a researcher has access to individual-level addresses, they should be able to geocode these to find the matching Census blocks and should also be able to parse the residency type (multi-unit or stand-alone). But researchers relying on more aggregate geographies likely do not have access to individual addresses, and hence the residency characteristics, for their sample.[10] In Figure 1 I pair each geolocation variable with a model missing the party ID input. The gains to predictive performance across all geographies appears roughly uniform with the addition of party ID.

Accuracy scores, however, are an incomplete metric for assessing predictive performance. In contexts where the true distribution of classes is highly imbalanced, Accuracy can provide overly-optimistic results. For example, if we were to simply classify every individual as White in the North Carolina/Florida voter file, we would achieve 65% Accuracy without even trying! We can evaluate the models in a more rigorous way by looking at each ethnorace category separately.

Two better metrics to assess group-level predictions are Precision and Recall. Precision is the percentage of correctly classified individuals among all individuals predicted to belong to a specific ethnorace. It answers the question of how likely an individual's predicted ethnorace in our sample matches their true, or self-identified, ethnorace. Recall, also known as Sensitivity or the True Positive Rate, is the percentage of all individuals who belong to a given ethnorace group which the model correctly classifies.

Precision and Recall reflect substantively important concerns for real-world applications of the method, and the inherent trade-offs between optimizing for either metric provide a balanced assessment of the predictive performance. On the one hand, Precision rewards very conservative classification procedures. We could, for example, only classify individuals as White if their posterior probability of being White was greater than 99%. This would ensure a very high Precision score for Whites because we are only capturing the low-hanging fruit. A conservative classification procedure

[10]In the event that data is available, adding multi-unit occupancy inputs to aggregate geographies, such as ZIP codes, places, counties, or states *greatly* enhances the predictive accuracy of the model.

like this, however, would likely result in extremely low Recall for Whites. If we only capture the low-hanging fruit, a greater share of White individuals will be mis-classified as non-White. Likewise, optimizing the algorithm for perfect Recall for Whites is trivial. By classifying every individual as White, we ensure 100% of Whites are correctly classified. Of course this procedure would result in extremely low Precision for Whites because every non-White individual would be classified as White as well. Achieving both high Precision and high Recall, therefore, is a difficult task.

Figure 2: Precision/Recall Scores by Ethnorace and Input Data

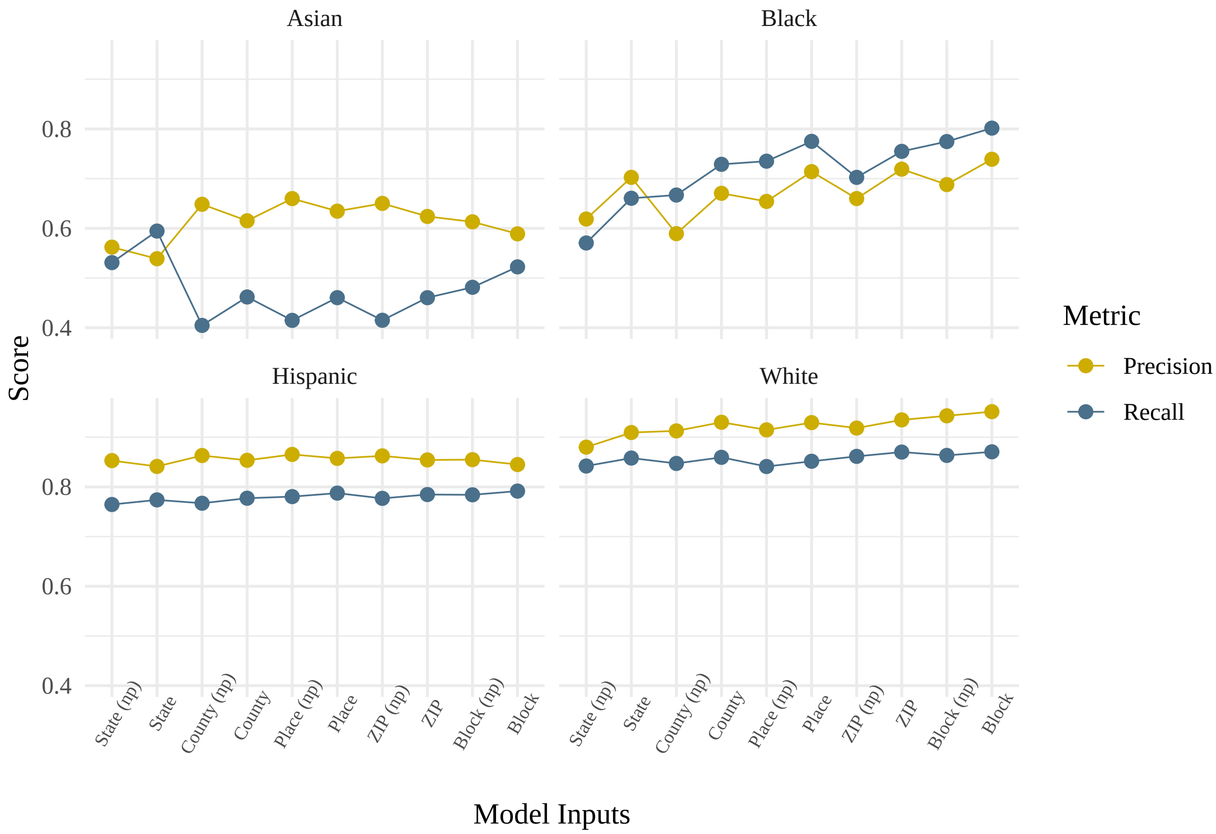


Figure 2 displays the Precision and Recall scores broken down by ethnorace for the same models used in Figure 1.[11] For both Whites and Hispanics, Precision and Recall remain high across all models. Using the most input data available (the models on the far right of the figure), White Precision and Recall are 0.951 and 0.871, respectively. Hispanic Precision and Recall in this model are 0.845 and 0.792, respectively. Hispanic predictive performance across all models likely remains high due to the distinctiveness of Spanish surnames. These metrics are uniformly lower for Asians—

[11]The model names are abbreviated by the level of geography, and all use first name and last name inputs. Models with *(np)* do not use the party ID inputs.

likely reflecting the relative rarity of Asian individuals in the sample. When using Census blocks as the unit of geography, predictions for African Americans are quite strong (Precision: 0.739, Recall: 0.802). But predictive performance falls with broader geolocations. At the state level, without party ID, Black Precision is 0.619 and Recall is 0.571. This difference in predictive performance between geographies could be explained by the legacy of Black segregation in at least the North Carolina and Florida sample.

The classification metrics detailed above provide some information about the predictive shortcomings of my method. These limitations notwithstanding, however, the ethnorace predictions made by `bper` are nearly uniformly better than those from `wru`. Figure 3 displays the difference in Accuracy between `bper` and `wru` using the same input data discussed previously.[12] The baseline of zero in the figure represents the Accuracy scores from each model using `wru`, and the height of the bars display the change in Accuracy using `bper`. Across all model types, `bper` scores roughly between 1.7 and 4.3 percentage points higher on this metric. Regardless of input data, `bper` classifies a higher proportion of individuals correctly in the North Carolina/Florida voter file.

Figures 4 through 7 show the same comparisons between `wru` and `bper` for Precision and Recall across different ethnoraces. For Asians in Figure 4, the comparisons shows absolute gains for `bper` in both metrics at block-level geographies. However, when using more aggregate geographies there is some loss to Recall. The large gains in Precision for Asians should generally compensate for these concerns, but this may depend on the particular substantive application of the method.

The comparison among predictions for Black individuals in Figure 5 are striking. My method shows dramatic gains in Precision for all models while also modestly increasing Recall. Without party ID, `bper` improves upon `wru` by nearly 30 percentage points in Precision using place or county geolocations. This means that individuals predicted to be Black by `bper` are nearly 30 percentage points more likely to self-identify as Black relative to the predictions generated by `wru`. The modest gains in Recall across all models signifies that `bper` is correctly predicting a greater share of self-identified African Americans in the sample as well.

Figure 6 shows the comparison in predictions among Hispanics. Again, here `bper` out-performs `wru` in both Precision and Recall for every model. The magnitudes of these differences, however, are

[12] The `wru` package does not provide ZIP code or state level predictions so those models are excluded from the comparison.

Figure 3: Accuracy Score Comparison to `wru`

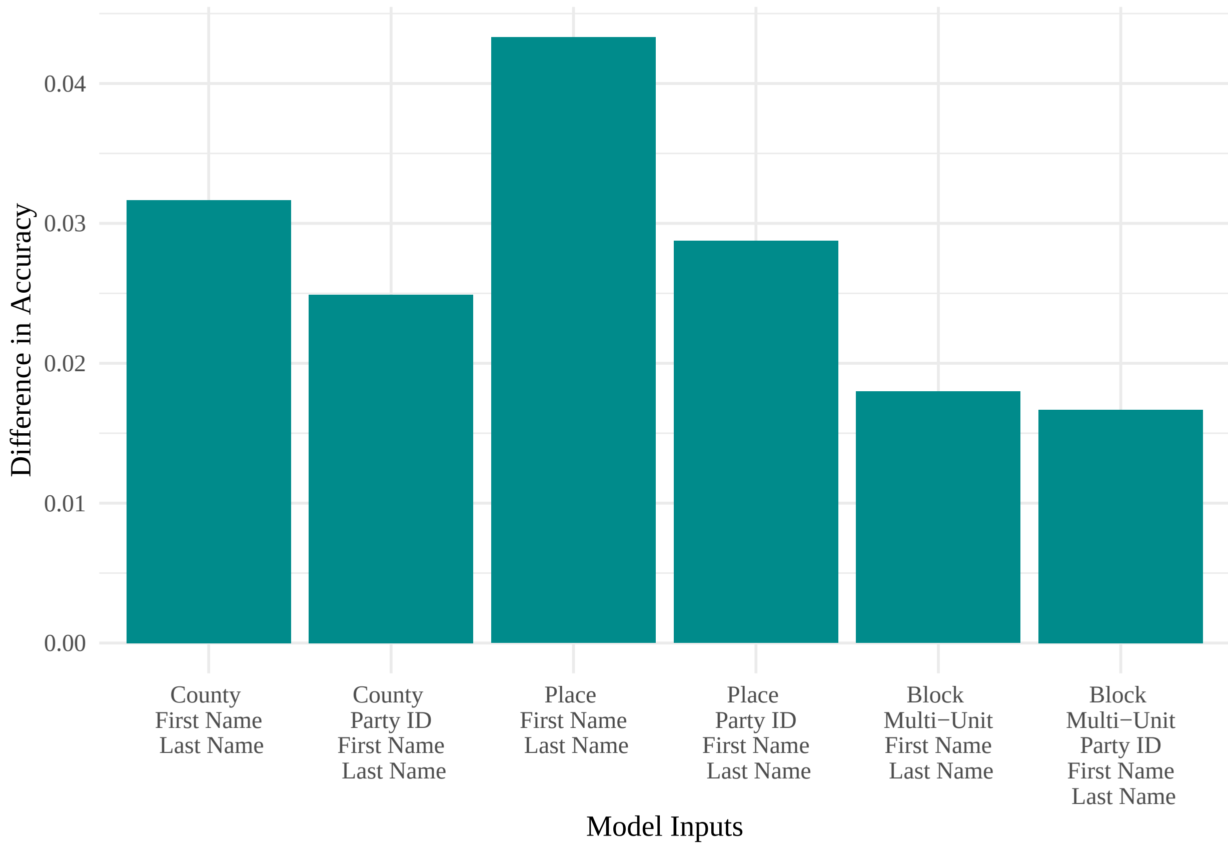


Figure 4: Precision and Recall Score Comparison to `wru`: Asian

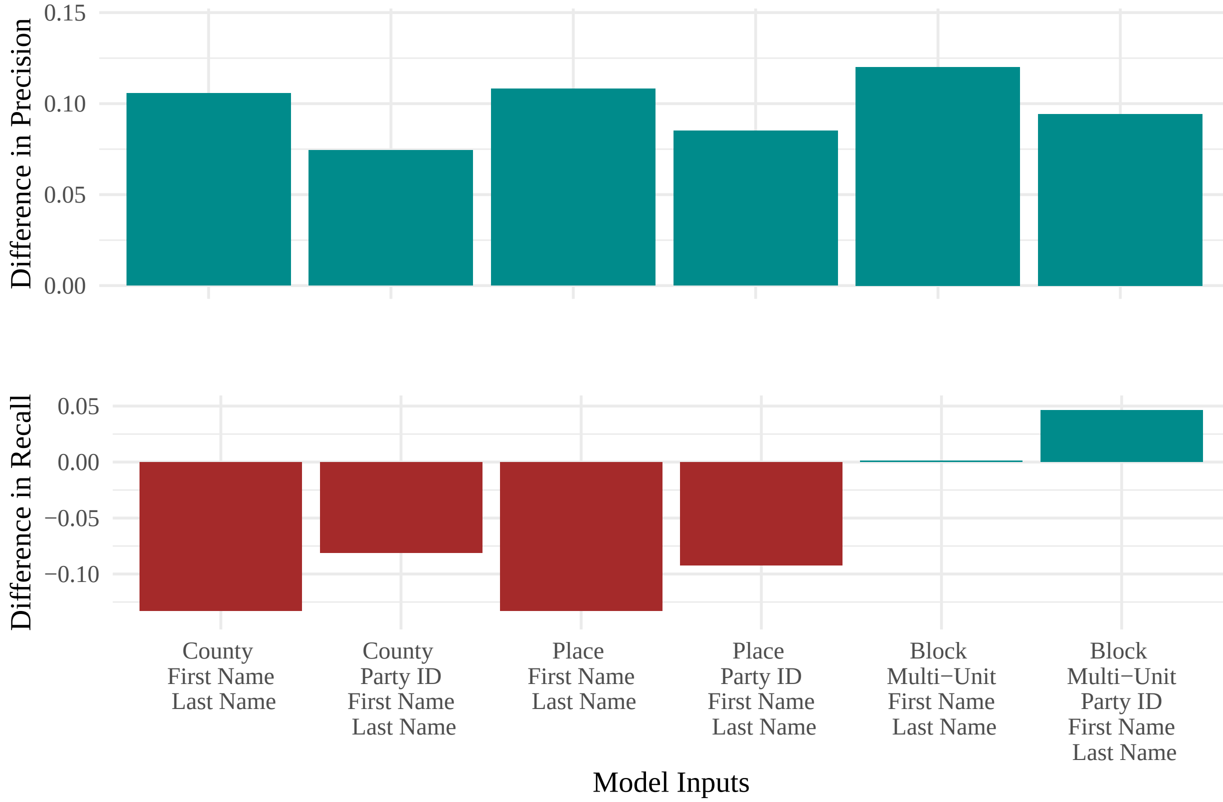
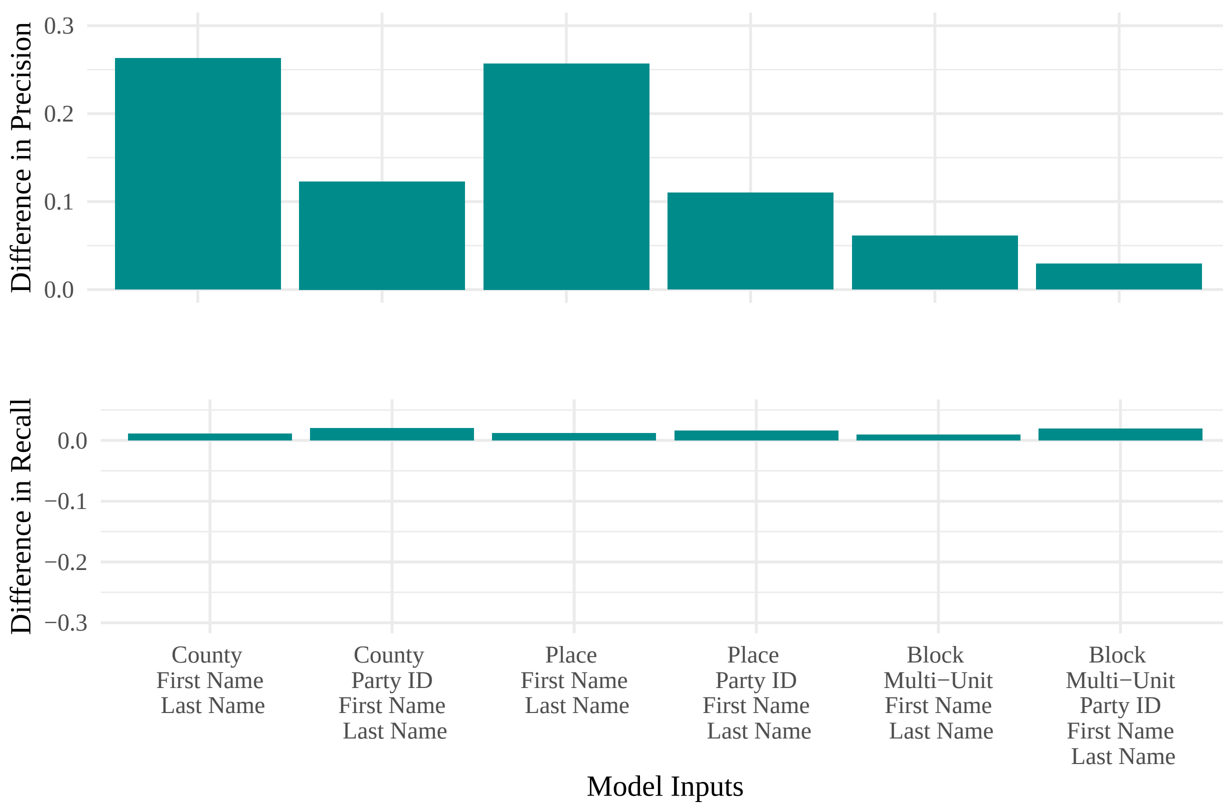
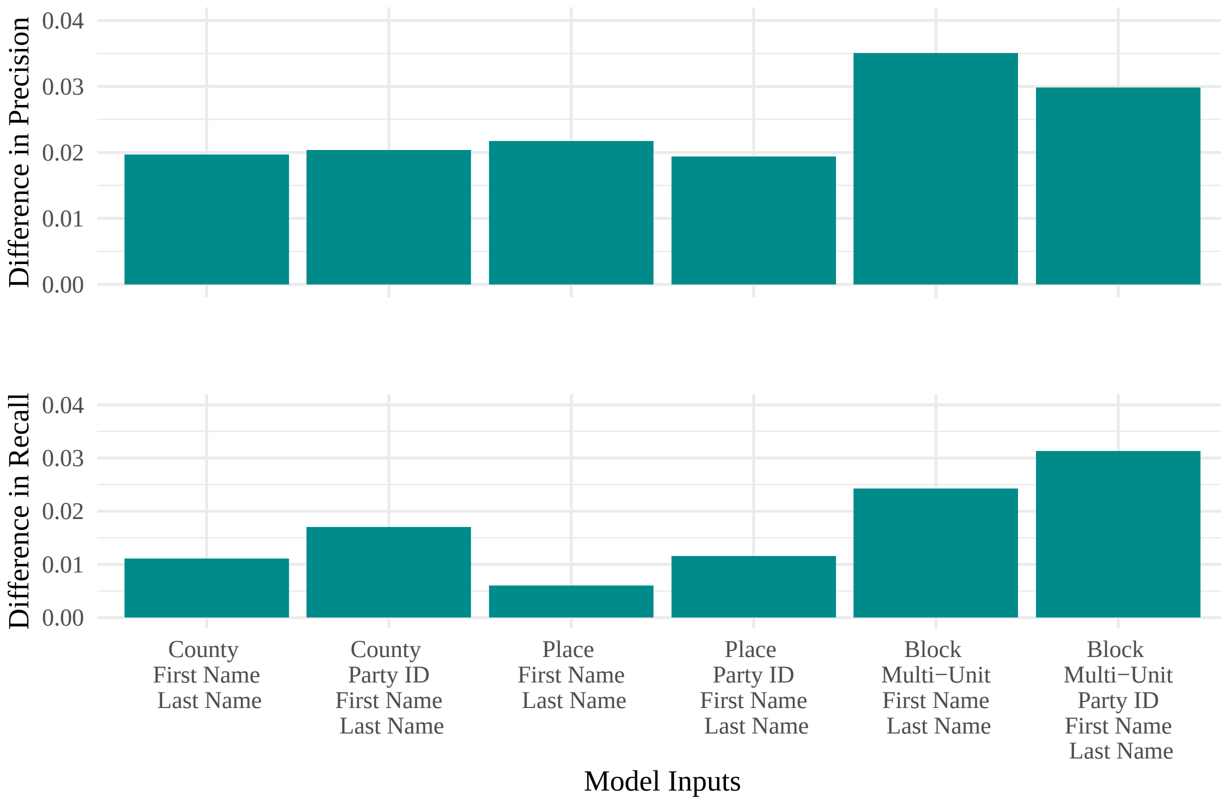


Figure 5: Precision and Recall Score Comparison to `wru`: Black



smaller than those for Black or Asian predictions. This is likely due to distinct Spanish last names, which already provide a lot of predictive information for Hispanics. Therefore, the additional predictive improvements in `bper`, such as first name information and smoothing, have less to contribute.

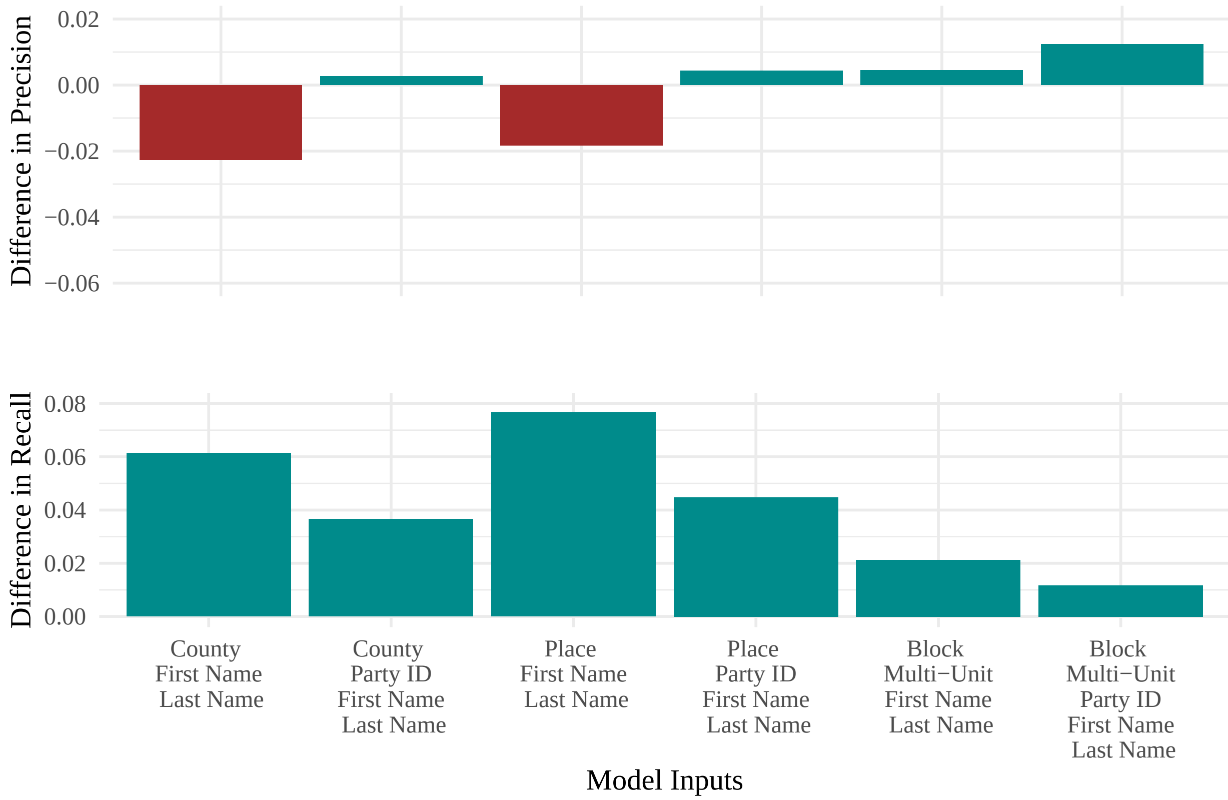
Figure 6: Precision and Recall Score Comparison to `wru`: Hispanic



Lastly, Figure 7 displays the results for the White predictions. In all models except county and place geolocations (without party ID), `bper` out-performs `wru` on both Precision and Recall for White individuals. In the two exceptions, however, losses to Precision are compensated with gains in Recall. This means that, while `bper` is doing a better job classifying a higher proportion of Whites, it is mis-classifying more non-Whites as White compared to `wru` for county and place geolocations without party ID.

In sum, when use block-level geography (with or without party ID), `bper` outperforms `wru` on both Precision and Recall for each group. When we transition to more aggregate geographies, `bper` shows dramatic gains in Precision for Asians and African Americans. Without party ID, there is some loss to Precision for Whites in aggregate geographies, but gains in Recall are more than triple

Figure 7: Precision and Recall Score Comparison to `wru`: White



the magnitude for these models.

References

Abott, Carolyn, and Asya Magazinnik. 2020. "At-Large Elections and Minority Representation in Local Government." *American Journal of Political Science* 64 (3): 717–33. https://doi.org/10.1111/ajps.12512.

Burch, Traci. 2013. *Trading Democracy for Justice: Criminal Convictions and the Decline of Neighborhood Political Participation*.

Cantoni, Enrico. 2020. "A Precinct Too Far: Turnout and Voting Costs." *American Economic Journal: Applied Economics* 12 (1): 61–85. https://doi.org/10.1257/app.20180306.

Crabtree, Charles, and Volha Chykina. 2018. "Last Name Selection in Audit Studies." *Sociological Science* 5: 21–28. https://doi.org/10.15195/v5.a2.

Domingos, Pedro, and Michael Pazzani. 1997. "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier." *Machine Learning* 29: 103–30.

Elliott, Marc N., Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantoja, and Nicole 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. https://doi.org/10.1007/s10742-009-0047-1.

Enos, Ryan D., Aaron R. Kaufman, and Melissa L. Sands. 2019. "Can Violent Protest Change Local Policy Support? Evidence from the Aftermath of the 1992 Los Angeles Riot." *American Political Science Review* 113 (4): 1012–28. https://doi.org/10.1017/S0003055419000340.

Fotheringham, A S, and D W S Wong. 1991. "The Modifiable Areal Unit Problem in Multivariate Statistical Analysis." *Environment and Planning A: Economy and Space* 23 (7): 1025–44. https://doi.org/10.1068/a231025.

Grumbach, Jacob M., and Alexander Sahn. 2020. "Race and Representation in Campaign Finance." *American Political Science Review* 114 (1): 206–21. https://doi.org/10.1017/S0003055419000637.

Grumbach, Jacob M., Alexander Sahn, and Sarah Staszak. 2020. "Gender, Race, and Intersectionality in Campaign Finance." *Political Behavior*, June. https://doi.org/10.1007/s11109-020-09619-0.

Imai, Kosuke, and Kabir Khanna. 2016. "Improving Ecological Inference by Predicting Individual

Ethnicity from Voter Registration Records." *Political Analysis* 24 (2): 263–72. https://doi.org/10.1093/pan/mpw001.

Keele, Luke, and Rocío Titiunik. 2018. "Geographic Natural Experiments with Interference: The Effect of All-Mail Voting on Turnout in Colorado." *CESifo Economic Studies* 64 (2): 127–49. https://doi.org/10.1093/cesifo/ify004.

Kuk, John, Zoltan Hajnal, and Nazita Lajevardi. 2020. "A Disproportionate Burden: Strict Voter Identification Laws and Minority Turnout." *Politics, Groups, and Identities*, June, 1–9. https://doi.org/10.1080/21565503.2020.1773280.

Lewis, David D. 1998. "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval." In *Machine Learning: ECML-98*, edited by Claire Nédellec and Céline Rouveirol, 1398:4–15. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/BFb0026666.

Omi, Michael, and Howard Winant. 2014. *Racial Formation in the United States*. Routledge.

Rish, Irina. 2001. "An Empirical Study of the Naive Bayes Classifier." *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence* 3 (22): 41–46.

Studdert, David M., Yifan Zhang, Sonja A. Swanson, Lea Prince, Jonathan A. Rodden, Erin E. Holsinger, Matthew J. Spittal, Garen J. Wintemute, and Matthew Miller. 2020. "Handgun Ownership and Suicide in California." *New England Journal of Medicine* 382 (23): 2220–9. https://doi.org/10.1056/NEJMsa1916744.

Trounstine, Jessica. 2020. "The Geography of Inequality: How Land Use Regulation Produces Segregation." *American Political Science Review* 114 (2): 443–55. https://doi.org/10.1017/S0003055419000844.

Tzioumis, Konstantinos. 2018. "Demographic Aspects of First Names." *Scientific Data* 5 (1): 180025. https://doi.org/10.1038/sdata.2018.25.

Velez, Yamil Ricardo, and Benjamin J. Newman. 2019. "Tuning in, Not Turning Out: Evaluating the Impact of Ethnic Television on Political Participation." *American Journal of Political Science* 63 (4): 808–23. https://doi.org/10.1111/ajps.12427.

Voicu, Ioan. 2018. "Using First Name Information to Improve Race and Ethnicity Classification." *Statistics and Public Policy* 5 (1): 1–13. https://doi.org/10.1080/2330443X.2018.1427012.