Improved Bayesian Ethnorace Prediction

Bertrand Wilden

Last updated on 07 February, 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 and Hispanics—groups which existing methods find difficult to predict accurately without fine-grain geolocation information.

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 (Cascio and Washington 2012; Kuk, Hajnal, and Lajevardi 2020). 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 designs—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 is a central element of the research question, individual-level identifiers for these categories are likely necessary.

To overcome some of these challenges, in this paper I describe a method for imputing individual ethnorace categories directly. This method uses Bayes' rule to make predictions by combining information from nationwide distributions of six ethnoracial categories over other characteristics, such as names, geographies, political party identification, among others. My implementation builds off existing methods which use a similar formula (Elliott et al. 2009; Imai and Khanna 2016; Voicu 2018). The method described in 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), 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 version 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 on the Naive Bayes algorithm. The result of these modifications is a substantial increase in predictive performance 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 will provide some background on the inputs and outputs of my prediction method. Then I will explain the methodology and compare my implementation with previous versions. Finally, I will demonstrate the predictive performance of my method when validated against the combined North Carolina and Florida voter file ($n = 21{,}000{,}000$).

Data

Outputs

Before discussing the methodology further, it is important to first define what "predicting ethnicity or race" means. These are categories which, although relatively immutable compared to other

---

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

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 categories.

There are a few benefits to using the Census ethnoracial categorization. These groups capture a common understanding of race and ethnicity in the US, and correspond to the groups studied most frequently in social science research. The data sources of these groups' distributions that serve as inputs to the prediction formula 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 an Other Race category ensures that important sources of diversity are washed over.[4]

Inputs

First Names

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

---

[2]Non-Hispanic Other Race includes individuals who identify as belonging to two or more race/ethnicities, or mixed race, 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 abysmal predictive performance for the Other Race category.

benefits from using first name data, as I will demonstrate, likely overwhelms 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). 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 descending 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. It is worth pointing out that this phenomenon is, in part, the consequence of generations of segregationist housing policies in the US. The fact that knowing an individual's ZIP code or Census block gives us so much knowledge about their race is an indictment of the US system more generally.

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").

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 formula helps slightly.

---

[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

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.

Methodology

In general terms, the method described here 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.

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

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:

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

Where x is the vector of individual attributes indexed by j. The particular attribute x' comes from the use of the chain rule when decomposing 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. Using last names for x' appears to help predictions of Whites—but to the detriment of other categories. 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 ethnorace to generate final predicted probabilities that an individual belongs to a particular ethnorace. The end result is more balanced predictions across each ethnorace category.

The conditional independence assumption necessary for transforming (1) to (2) means that knowing both a particular attribute, and an 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

Validation

Conclusion

cite list:

methods:

- (Voicu 2018)

- (Tzioumis 2018)

- (Shah and Davis 2017)

- (Imai and Khanna 2016)

- (Elliott et al. 2009)

- (Crabtree and Chykina 2018) (independence assumption)

- (Lauderdale and Kestenbaum 2000) (asian surnames)

naive bayes:

- (Rish 2001)

- (Lewis 1998)

- (Domingos and Pazzani 1997)

applications:

- (Velez and Newman 2019) (use NC)

- (Studdert et al. 2020) (medical study)

- (Kuk, Hajnal, and Lajevardi 2020) (aggregate)

- (Cascio and Washington 2012) (aggregate)

- (Cantoni 2020) (RDD)

- (Hersh and Nall 2016) (catalist)

- (Ghitza and Gelman, n.d.) (catalist)

- (Grumbach and Sahn 2020)

- (Grumbach, Sahn, and Staszak 2020)

- (Fraga 2018) (turnout)

- (Enos, Kaufman, and Sands 2019) (protests and voting)

- (Enos 2016) (racial threat)

race:

- (Omi and Winant 2014)

References

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.

Cascio, Elizabeth U, and Ebonya L Washington. 2012. "Valuing the Vote: The Redistribution of Voting Rights and State Funds Following the Voting Rights Act of 1965," 53.

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. 2016. "What the Demolition of Public Housing Teaches Us About the Impact of Racial Threat on Political Behavior." *American Journal of Political Science* 60 (1): 123–42. https://doi.org/10.1111/ajps.12156.

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.

Fraga, Bernard L. 2018. *The Turnout Gap. Race, Ethnicity, and Political Inequality in a Diversifying America.* Cambridge University Press.

Ghitza, Yair, and Andrew Gelman. n.d. "Voter Registration Databases and MRP: Toward the Use of Large-Scale Databases in Public Opinion Research," 25.

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.

Hersh, Eitan D., and Clayton Nall. 2016. "The Primacy of Race in the Geography of Income-Based Voting: New Evidence from Public Voting Records: RACE AND INCOME-BASED VOTING." *American Journal of Political Science* 60 (2): 289–303. https://doi.org/10.1111/ajps.12179.

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.

Lauderdale, Diane S, and Bert Kestenbaum. 2000. "Asian American Ethnic Identification by Surname." *Population and Development Review* 19: 18.

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.

Shah, Paru R., and Nicholas R. Davis. 2017. "Comparing Three Methods of Measuring Race/Ethnicity." *The Journal of Race, Ethnicity, and Politics* 2 (1): 124–39. https://doi.org/10.1017/rep.2016.27.

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.

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.