CENTER FOR
INCLUSIVE
DEMOCRACY

Inclusive Data Advisory Committee

An Overview: Race and Ethnicity Identification in Current Voting and Election Research Methods

Conducting research aimed at understanding the historic and current patterns of civic and electoral underrepresentation in the U.S has its challenges. Currently, there is no method employed to identify a voter's race and ethnicity that is ideal for election and voting data collection. While self-reported race and ethnicities are highly reliable, low self-response rates create a small sample of voters in most states, which is not representative of the full picture of voting behaviors. To further examine this issue, in 2021 the Center for Inclusive Democracy (CID) formed the Inclusive Data Advisory Committee, a group of leading U.S. researchers and practitioners in the voting field. In this memo, the Committee summarizes the most common research methods used to identify race and ethnicities of voters and the limitations of each method.

Background

Residents in most U.S. states are not required to identify their race or ethnicity on voter registration forms, which makes the analysis of voter turnout and other voting behaviors for racial and ethnic groups challenging. The percent of voters who self-identify their race or ethnicity when registering to vote varies across states, with some states seeing self-identification by less than a quarter of their registered voters. For this reason, when conducting elections-related research, alternative methods must be deployed in order to identify the race and ethnicity of individual registered voters. However, no method used to identify these characteristics in voter files is ideal, especially with growing concerns involving the quality and limitations of census population data.

With many methods, some race and ethnicities, such as Black and Indigenous voters, are more difficult to identify than others. Misidentifying or failing to identify historically underrepresented voters in the voter file can lead to an incomplete or incorrect picture of those groups' voting behaviors. It should also be noted that common research methods do not capture the complex nature of a person's racial or ethnic self-conception, which is not determined by their last name or where they choose to live.

Methods for Identifying Race and Ethnicity of Voters

Surname Matching

Surname matching entails comparing surnames in voter registration records to available ethnic surname lists. Surname analysis infers race and ethnicity from surnames that research has found are distinctive to racial and ethnic groups. Latinos are identified using the Passel-Word Spanish surnames list generated by the U.S. Census, which identifies common Spanish surnames. Asian-Americans are identified using surname lists derived and separately evaluated by researchers to identify persons belonging to six principal Asian American ethnic groups: Chinese, Japanese, Asian Indian, Korean, Filipino, and Vietnamese.

Names that are more exclusive to a particular racial or ethnic group are helpful in identifying members of that group. Surname matching is commonly used and trusted to identify Latino and Asian-American voters by many researchers in the elections field. However, the method is far from perfect. For instance, the Passel-Word list published by the U.S. Census assigns any person with a Latino surname to the Latino category, which can misassign some people such as Filipinos. There are also separate lists for each of the six Asian American groups mentioned above, but they are less reliable and not commonly used for research purposes. Black and white populations have some common surnames, but they are not exclusive to those demographic groups, making surname matching alone a less reliable identifier for these groups. Further, a growing multi-racial U.S. population can make analysis by surname a greater challenge going forward.

Geocoding can address some of the limitations of surname analysis. Geocoding links an individual's given address to a census measure of their census tract's racial and ethnic population makeup and uses that measure as a basis for inferring the individual's race or ethnicity. This method can be used in combination with surname matching for Latinos and Asian Americans. For white, non-Latino and Black voters, geocoding can produce some level of accuracy at an aggregate level, especially for Black voters who are more likely than other groups to live in segregated neighborhoods. In neighborhoods where most voters are Black, it is reasonable to assume that any given registrant from that neighborhood is also Black—more so if other information like surname analysis also suggests the same conclusion. But in most parts of the U.S. (including California), the Black population is too small, and geocoding can erroneously assign Black voters to other racial or ethnic groups. This problem makes geocoding unreliable for Black voters in most areas within the U.S. While reasonable estimates for counties that have substantial Black populations can be generated through geocoding at the census tract level, the concern is that reliable estimates for Black voters cannot be produced in counties where every tract has a small Black population.

Bayesian Improved Surname Geocoding (BISG)

Geocoding and surname analysis methods alone have limitations (geocoding is limited in identifying Asian Americans and Latinos, and surname analysis is limited in differentiating between Blacks and white, non-Latinos. In response to the challenges of identifying the race and ethnicities of voters using surname matching and geocoding alone, researchers have looked to expand the geocoding with accompanying data from the voter file, such as home address, gender, age, and party affiliation. Bayesian Improved Surname Geocoding (BISG) relies on a combination of census surname analysis and geocoding with census block-level racial demographics to provide an overall probability assessment of the individual's race or ethnicity. In recent years, many researchers in the elections field have adapted BISG and produced findings validating it as an appropriate method.

In 2007, the U.S. Census Bureau released a comprehensive surname list detailing surnames classified by self-reported race and ethnicity from almost 270 million U.S. residents based on the 2000 Census. The BISG method uses the updated surname list and calculates probabilities for six race and ethnicity categories (Latino, white, Black, Asian American and Pacific Islander, American Indian/Alaska Native, and multiracial). The BISG approach has since been updated by many researchers to use the 2010 Census surname list.

The first step in BISG is the surname analysis. Surname analysis in BISG compares last names on voter registration cards to the published surname directories created by the Census Bureau, which assigns probabilities of the racial or ethnic group associated with each last name. The probabilities were developed from the U.S. Census when individuals record their last name and their self-identified race or ethnicity. The second step in BISG is using geocoding to cross-reference the voter's home address with the census data on the self-reported race of residents. Based on census statistics for the racial and ethnic composition of the block in which a voter resides, the geographical race and ethnicity probabilities can be used to refine the initial estimate of voter race or ethnicity by surname alone. This method is particularly helpful in identifying Black voters, who are more difficult to identify using surname lists and are more likely to live in segregated neighborhoods.

Expansion of BISG Methods

Some researchers in the U.S. have looked to expand BISG to include other demographic information, such as age, gender, and party affiliation to further improve the method's accuracy. Imai and Khanna, political methodologists at Princeton University, developed the Who Are You (WRU) package in R to incorporate party affiliation into the model to predict race and ethnicity probabilities. The researchers used publicly available Gallup polling data to obtain the distribution of partisanship by race in order to extend the model.

In recent years, some researchers have looked to incorporate first names into the BISG model to further increase its accuracy. Voicu (2018) used a first name list derived from mortgage application data and included 4,250 first names associated with the six race and ethnicity categories to create the Bayesian Improved First Name Surname Geocoding (BIFSG) method. BIFSG expands BISG by including first-name based probabilities. The probability is approximated based on the proportion of the population of the given race or ethnicity who bears the respective first name. During his validation process, Voicu found that BIFSG outperforms BISG in terms of accuracy in all major racial and ethnic categories. While overall improvements were small, the identification of Black, non-Latinos saw the greatest increase in identification, a group for which the BISG has one of the lowest accuracy rates.

BISG Limitations

The level of geography used to geocode the voter file is important to consider. Since neighborhood characteristics are a crucial component to BISG, researchers generally agree the Census block level is the most accurate geography level to use. Some researchers have explored other geography levels, such as voting precinct and ZIP code. Imai and Khanna tested both voting precinct and Census block and found that the Census block returns fewer false negatives in all race and ethnicity categories. The Census block level, however, has been identified as a potential issue for some communities that do not perfectly fit inside a Census block. Recently, Clark et al (2021) found that forgoing geocoding and using ZIP codes in the BISG model can work better for Black and white, non-Latino voters than for Latino and Asian-American voters. Since ZIP codes are the smallest unit of publicly known geography, they do not need to be geocoded for use. The authors argue that ZIP codes are a good alternative for data that cannot be geocoded due to lack of full addresses.

The use of BISG is becoming more common in the elections and voting rights field. Barreto et al (2021) examined the potential use of BISG in the 2021-22 redistricting round and voting rights cases in federal courts. The researchers argued that BISG is an improved method for identifying Black, Latino, and Asian-American voters to prevent the dilution of minority votes that can occur when solely relying on ACS and Census data. Additionally, Barreto et al reasoned that BISG can overcome concerns related to surname matching alone, such as the inability to identify Latinos without a Latino surname or incorrectly identifying non-Latinos as Latino who have a surname gained through marriage. For example, the researchers state that a voter with an 83% Hispanic occurring surname living in an 80% Hispanic populated census block has an extremely low probability of being white, Black, or Asian American. BISG also improves issues with surname matching alone when the surname is not exclusive to a particular race or ethnicity. The surname Williams, for example, is 45.7% white and 47.7% Black. The geocoding step in BISG can help narrow down the possibilities. If a voter with the surname Williams lives in a census block that is overwhelmingly Black, it is statistically very likely that the voter is Black too. Barreto et al. conclude that the combination of both surname analysis and census block level data provides a more precise estimate of each voter's race or ethnicity than just using either method alone.

In recent years, BISG has been applied to a variety of research areas where race and ethnicity data is difficult to obtain, including campaign finance (Alvarez, Katz, and Kim 2020), police related deaths (Edwards, Esposito, and Lee 2018), and the intersection of race and gender (Signorella 2020). However, another potential limitation of BISG is the quality and limitations of source Census data. For example, 2020 American Community Data was released as experimental with limited geographic coverage. Further, Decennial Census differential undercounts could possibly make this approach a bit less reliable, as well as new approaches to data disclosure avoidance implemented by the Census Bureau.

Summary

Currently, there is no method employed to identify a voter's race and ethnicity that is ideal. While self-reported race and ethnicities are highly reliable, low response rates create a small sample of voters in most U.S. states that can change the picture of voting behaviors. Surname matching has a high level of accuracy for some demographics, such as Latinos and Asian Americans (although to a lesser degree), but it is, generally, not reliable for identifying others, such as Black and Indigenous voters. BISG reduces the unreliability found in surname matching by utilizing neighborhood characteristics to better detect hard to identify demographics. BISG's accuracy in identifying these demographics, however, is dependent on the neighborhood voters live in and could under- or over-count voters based on the community's demographic makeup. BISG is generally not reliable for calculating voter turnout rates for many groups in many locations, including Black and Indigenous voters, and is unable to identify Asian-American subgroups at county or higher geographic levels.

Notes

See: https://www.census.gov/topics/public-sector/voting/about.html
See: https://cran.r-project.org/web/packages/wru/wru.pdf
See: https://www.rand.org/pubs/research_reports/RR1162.html

References

Alvarez, R. M., Katz, J. N., & Kim, S. Y. S. (2020). Hidden Donors: The Censoring Problem in US Federal Campaign Finance Data. Election Law Journal: Rules, Politics, and Policy, 19(1), 1-18. Retrieved from https://preprints.apsanet.org/engage/apsa/article-details/5e12e22ccd361a001afed251

Barreto, M., Cohen, M., Collingwood, L., Dunn, C., & Waknin, S. (2021). A Novel Method for Showing Racially Polarized Voting: Bayesian Improved Surname Geocoding. New York University Review of Law & Social Change, Forthcoming. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3834818

Clark, J. T., Curiel, J. A., & Steelman, T. S. (2021). Minmaxing of Bayesian Improved Surname Geocoding and Geography Level Ups in Predicting Race. Political Analysis, 1-7. Retrieved from https://www.cambridge.org/core/journals/political-analysis/article/abs/minmaxing-of-bayesian-improved-surname-geocoding-and-geography-level-ups-in-predicting-race/2B259C0A8B66EFB00C4AD05B19CCFF4A

Deluca, K., & Curiel, J. A. Validating the Applicability of BISG to Congressional Redistricting. Retrieved from https://electionlab.mit.edu/sites/default/files/2021-07/deluca-curiel_validating_bisg.pdf

Edwards, F., M. H. Esposito, and H. Lee. 2018. “Risk of Police-Involved Death by Race/Ethnicity and Place, United States, 2012–2018.” American Journal of Public Health 108(9):1241–1248. Retrieved from https://ajph.aphapublications.org/doi/pdfplus/10.2105/AJPH.2018.304559

Elliott, M. N., Fremont, A., Morrison, P. A., Pantoja, P., & Lurie, N. (2008). A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. Health services research, 43(5p1), 1722-1736. Retrieved from https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1475-6773.2008.00854.x?casa_token=jUIHtM8E6HkAAAAA:y3_UDnNdHdxP3o-QT1QZqsWEMv5AD3E0SlNgoS_19p3KJE-UHh4tN0FTXrpgKj2CBCEf4wWc19boyI0

Elliott, M., Morrison, P., et al. (2009). Using the Census Bureau’s Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities. Health Serv Outcomes Res Method 9:69–83. Retrieved from https://www.thecre.com/insurance/wp-content/uploads/2010/06/CFPB-Paper.pdf

Lauderdale, D., Kestenbaum, B. (2000). Asian American Ethnic Identification by Surname. Population Research and Policy Review 19: 283–300, 2000. Retrieved from https://statewidedatabase.org/info/metadata/asian_american_ethnic_id_by_surname.pdf

Voicu, I. (2018). Using first name information to improve race and ethnicity classification. Statistics and Public Policy, 5(1), 1-13. Retrieved from https://doi.org/10.1080/2330443X.2018.1427012