On imputing Race/Ethnicity with Bayesian Improved Surname Geocoding - Opportunity and implications to community research

Kailas Venkitasubramanian, Charlotte Urban Institute

Eric Moore, Charlotte Urban Institute

2023-02-14

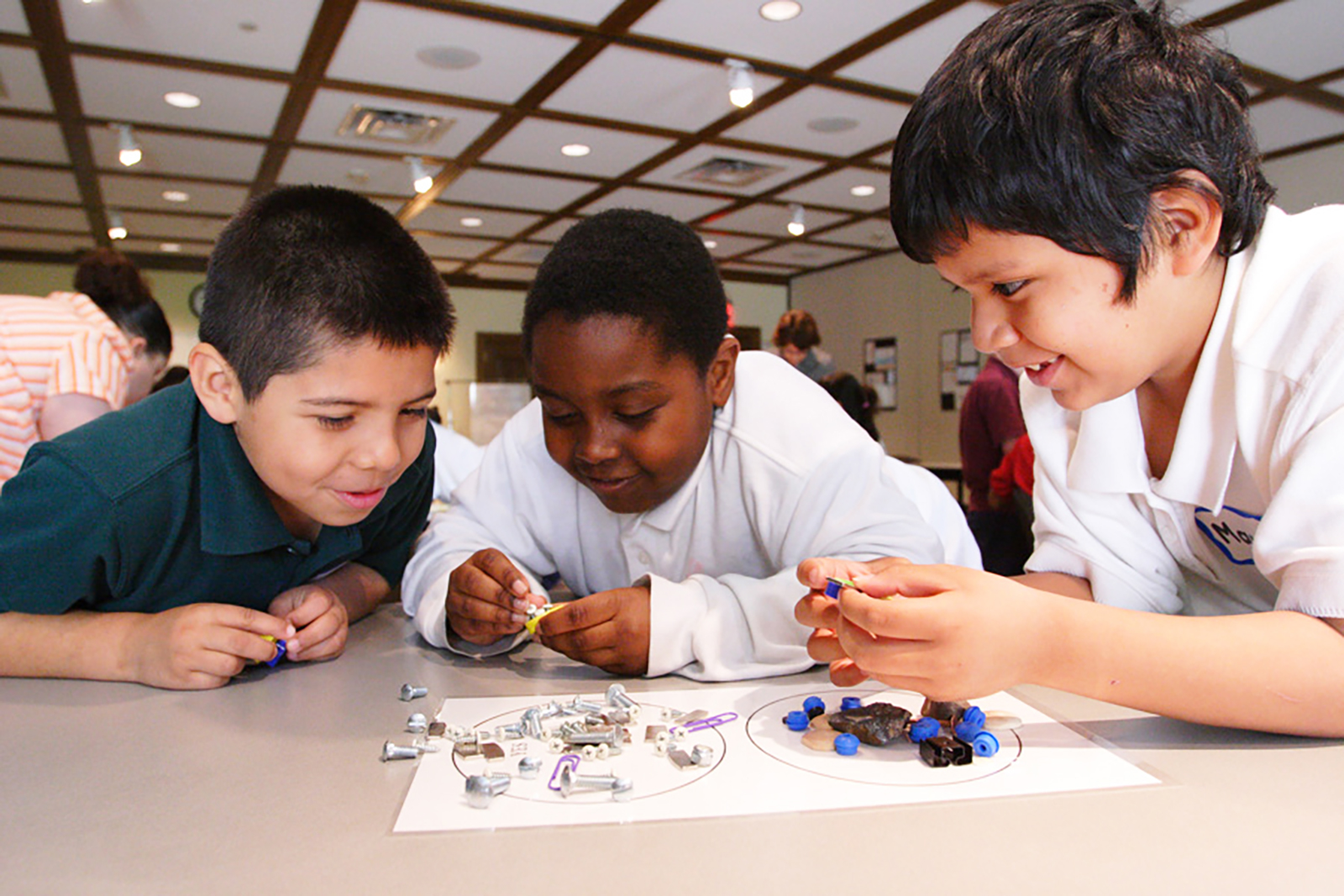


Photo: Bakken Museum

## Why impute race/ethnicity data?

Real-world data being messy and incomplete is more of a rule than exception. Non-existent, unreliable or incomplete race/ethnicity data can be detrimental to insights drawn from community research and policies that are informed by research.

For example, if outcomes of a program to address health disparities among Hispanic/Latino people in Mecklenburg county was evaluated using a mix of survey or administrative data, and the self-reported race/ethnicity responses in this dataset were largely missing or of poor quality, conclusions drawn from the analyses about the success of this program can be circumspect or unusable. High quality race/ethnicity data is important to accurately represent the demographic characteristics of communities, identify and address disparities, center equity and make research inclusive and sensitive to a diverse population.

Imputation is an important technique in social science research to address missing data issues. However, traditional imputation methods require at least part of the data to be non-missing in order to guess missing values. When race/ethnicity data is unavailable either due to restrictions in the source data system, or inability to record data at the time of collection, it is impossible to use traditional imputation methods to recover race/ethnicity data. Thus the value and utility of Bayesian improved surname geocoding.

What is Bayesian improved surname geocoding?

Bayesian Improved Surname Geocoding (BISG), originally developed by the [RAND Corporation](https://www.rand.org/pubs/periodicals/health-quarterly/issues/v6/n1/16.html), is a probabilistic algorithm to estimate the race/ethnicity of an individual based on their surname and the racial/ethnic composition of the geographic area where they reside.

BISG calculates the probability that an individual belongs to a particular racial group based on their name and geolocation using the US Census Bureau’s Surname list and demographic data of their area (county, tracts, block groups, blocks etc.). BISG gives us a set of probabilities that quantify whether a given person with missing race data is White, Black, Hispanic, Asian, or Other.

For example, if your last name is ‘Walker’ and you live in a neighborhood that is around 80% White, the BISG algorithm will place a high probability that you are White. Similarly, if your surname is Martinez and you live in a Hispanic dominated census tract, the imputation will guess that you are more likely to be Hispanic. Please refer to [our technical article](WRU_Paper_314.html) for a primer on implementing BISG during data analysis.

Several studies that have applied BISG or [validated BISG-derived estimates](https://pubmed.ncbi.nlm.nih.gov/23855558/) indicate a high level of accuracy of the algorithm in imputing racial/ethnic composition. The algorithm thus presents both opportunities and challenges for the institute’s work, and in general community research.

## How the Urban Institute uses BISG

The Urban Institute often works with survey data and other microdata that include participants’ names. Also, program/policy evaluations usually leverage secondary data collected by the program’s administration also include participant surnames and location. BISG allows recovery of race/ethnicity data when none is available or when the collected race/ethnicity data is known to be unreliable.

Secondly, the [Charlotte Regional Data Trust](https://ui.charlotte.edu/our-work/charlotte-regional-data-trust) integrates several administrative datasets for research. Our data science team’s work on improving record linkage algorithms suggest that race/ethnicity data helps identify people across different agency databases with greater accuracy. BISG therefore can augment our effort to leverage linked administrative data to address important community research questions.

BISG remains a relatively new technique in social science research, with [Imai and Khanna](https://www.cambridge.org/core/journals/political-analysis/article/abs/improving-ecological-inference-by-predicting-individual-ethnicity-from-voter-registration-records/9DC8EBA269C25B1C606040196A3CB779) (2016) publishing their approach to using Bayes rule with the Census surname list less than a decade ago. Yet, this method has been used to answer diverse research questions that include comparing political behavior and participation, understanding disparities in housing evictions, estimating the risk of police use of force, mobility patterns, addressing racial discrimination in lending practices, healthcare disparities, increased partisan sorting by racial groups, and transportation outcomes (Derose et al. 2012; Luo et al. 2016; Edwards, Lee, and Esposito 2019; Hepburn, Louis, and Desmond 2020; Einstein, Glick, and Palmer 2020; Brown and Enos 2021; Sartin et al. 2021).

As our work continues to touch similarly diverse domains in the Charlotte region, the ability of BISG to remediate or improve data quality issues on race and ethnicity allows the institute’s practice to bring an equity lens to research projects that otherwise may be stifled due to lack of critical data. But our explorations so far also raise concerns and pointers on how BISG method may be used in community research.

## Concerns and Practice Considerations

### Accuracy

BISG implicitly assumes that surnames are a reliable indicator of a person’s race/ethnicity and geographic distribution of surnames remains stable over time. Surnames can be influenced by cultural and historical factors such as immigration, adoption, and intermarriage. Rapidly growing or declining areas may witness significant changes in their race/ethnicity composition thus may affect BISG’s accuracy. Studies have shown that BISG can achieve high accuracy rates for some racial/ethnic groups, such as non-Hispanic whites and non-Hispanic blacks. However, accuracy rates may be lower for other groups, such as Hispanic/Latino individuals or those with multiracial backgrounds. While some of these discrepancies can be remedied by using the most recent Surname dictionaries (Imai, Olivella, and Rosenman 2022) and/or utilizing the lowest geographic level possible (Clark, Curiel, and Steelman 2021), it continues to be an issue that scholars need to be aware of.

### Data quality

BISG relies on accurate and reliable data sources, such as census data or administrative records, to estimate the race/ethnicity of individuals. The quality of these data sources can vary, and errors or inconsistencies can affect the accuracy of BISG estimates. BISG is also often used to estimate race/ethnicity at the neighborhood or census block level, which can result in small sample sizes and increased uncertainty in the estimates.

### Data Privacy

BISG relies on publicly available data sources, which can raise privacy concerns if the data are linked to identifiable individuals. Individuals could be re-identified based on BISG estimates if the data are not properly anonymized. This risk is higher when BISG estimates are combined with other sources of data that could be used to identify individuals, such as medical records or social media profiles.

### Fairness and Non-discrimination

BISG estimates can be used to stigmatize or discriminate against certain groups if they are presented in a way that reinforces negative stereotypes or assumptions about those groups. BISG estimates are based on probabilities and can be used to group individuals into racial or ethnic categories. It is important to ensure that BISG is used in a way that promotes fairness and non-discrimination.

### Informed consent

Informed consent is an ethical principle that requires researchers to obtain the voluntary and informed consent of study participants before collecting data or conducting research. BISG relies on publicly available data sources, such as census data or voter registration records, which may not have been collected specifically for research purposes. It is important to consider whether informed consent was obtained from individuals whose data are being used for BISG and to ensure that the data are being used in a way that respects their privacy and autonomy.

### Rare populations and groups

BISG relies on surname information to estimate race and ethnicity, which may not be accurate for rare populations or groups. Some examples of rare populations or groups include immigrants, refugees, or indigenous populations. BISG estimates may be less accurate for rare populations or groups, which may lead to inaccurate conclusions or unfair comparisons.

## Conclusion

Based on the scientific evidence and applications of BISG in the last decade, our work sees value and benefits of using this tool in community research especially in racially diverse regions like Charlotte. In addition to addressing the concerns above, work remains to be done to assess neighborhood level performance of BISG in estimating racial/ethnic composition to ensure high quality reporting using imputed data. But as our exploration progresses, we see engagement with partners and the community as a critical prerequisite to applying BISG to community research projects. Conducting research in a respectful and culturally sensitive manner remains a cornerstone of the institute’s research practice even as rapid technology advances in data science continue to empower our work in the region.

## More Resources

Brown, Jacob R., and Ryan D. Enos. 2021. “The Measurement of Partisan Sorting for 180 Million Voters.” *Nature Human Behaviour* 5 (8): 998–1008. <https://doi.org/10.1038/s41562-021-01066-z>.

Clark, Jesse T., John A. Curiel, and Tyler S. Steelman. 2021. “Minmaxing of Bayesian Improved Surname Geocoding and Geography Level Ups in Predicting Race.” *Political Analysis* 30 (3): 456–62. <https://doi.org/10.1017/pan.2021.31>.

Derose, Stephen F., Richard Contreras, Karen J. Coleman, Corinna Koebnick, and Steven J. Jacobsen. 2012. “Race and Ethnicity Data Quality and Imputation Using U.S. Census Data in an Integrated Health System.” *Medical Care Research and Review* 70 (3): 330–45. <https://doi.org/10.1177/1077558712466293>.

Edwards, Frank, Hedwig Lee, and Michael Esposito. 2019. “Risk of Being Killed by Police Use of Force in the United States by Age, Race–Ethnicity, and Sex.” *Proceedings of the National Academy of Sciences* 116 (34): 16793–98. <https://doi.org/10.1073/pnas.1821204116>.

Einstein, Katherine Levine, David M. Glick, and Maxwell Palmer. 2020. “Neighborhood Defenders: Participatory Politics and America’s Housing Crisis.” *Political Science Quarterly* 135 (2): 281–312. <https://doi.org/10.1002/polq.13035>.

Hepburn, Peter, Renee Louis, and Matthew Desmond. 2020. “Racial and Gender Disparities Among Evicted Americans.” *Sociological Science* 7: 649–62. <https://doi.org/10.15195/v7.a27>.

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>.

Imai, Kosuke, Santiago Olivella, and Evan T. R. Rosenman. 2022. “Addressing Census Data Problems in Race Imputation via Fully Bayesian Improved Surname Geocoding and Name Supplements.” *Science Advances* 8 (49). <https://doi.org/10.1126/sciadv.adc9824>.

Luo, Feixiong, Guofeng Cao, Kevin Mulligan, and Xiang Li. 2016. “Explore Spatiotemporal and Demographic Characteristics of Human Mobility via Twitter: A Case Study of Chicago.” *Applied Geography* 70 (May): 11–25. <https://doi.org/10.1016/j.apgeog.2016.03.001>.

Sartin, Emma B., Kristina B. Metzger, Melissa R. Pfeiffer, Rachel K. Myers, and Allison E. Curry. 2021. “Facilitating Research on Racial and Ethnic Disparities and Inequities in Transportation: Application and Evaluation of the Bayesian Improved Surname Geocoding (BISG) Algorithm.” *Traffic Injury Prevention* 22 (sup1): S32–37. <https://doi.org/10.1080/15389588.2021.1955109>.