Promises and Limits of Inferring Protected-Class Data for Disparate Impact Testing of AI Systems: Conference report

Jacob Appel, Bennett Borden, Cathy O'Neil, Dan Svirsky, Sam Tyner-Monroe

On September 8, 2023, ORCAA and DLA Piper hosted a conference in Boston about the use of Bayesian-Improved Surname Geocoding (BISG) and related race/ethnicity inference methods in fairness analyses, including disparate impact testing. The event brought together academics, practitioners, and regulators from a range of disciplines including insurance, law, voting rights, fair lending, and statistics. Our goals were to learn from each other about how these inference methods are being adopted and adapted in different areas, and to identify shared issues and open questions across these varied applications. In short, we aimed to convene a community of practice around this topic. This paper serves as a summary and report-out from the conference.

The paper is organized in three main sections. In the Technical Report section, we discuss the capabilities and requirements of BISG and some key variants and alternatives. In the Contextual section, we discuss how these methods are being applied. In the Policy section, we recap a discussion from the conference about fairness in insurance, as an example of how using inference methods can add clarity and substance to policy discussions. Finally, we conclude with a list of common issues and open questions that surfaced at the conference and could be taken up in the future by this nascent community of practice.

Technical Report

Basics of BISG

BISG is a method for inferring the self-reported race/ethnicity of individuals based on their surname and address, leveraging publicly available data from the U.S. Census Bureau. It was initially developed by the RAND Corporation with fairness testing in mind. RAND writes that it "can help U.S. organizations produce accurate, cost-effective estimates of racial and ethnic disparities within datasets — and illuminate areas for improvement."[1]

Two datasets are required to implement BISG: (1) a dataset of surnames, with the race/ethnicity distribution of all individuals who have each surname; and (2) a dataset of Census block groups[2], with the race and ethnicity distributions of each block group.

[1] https://www.rand.org/health-care/tools-methods/bisg.html

[2] A Census block group (BG) is the smallest geographic entity for which the decennial census tabulates and publishes sample data. A BG is a combination of census blocks that is a subdivision of a census tract or block numbering area. A census block is the smallest

The U.S. Census Bureau publishes these datasets[3] in connection with each decennial census.

Basically, BISG involves making one "guess" based on the person's surname, another based on his or her address, and combining them statistically. In slightly more detail,[4] it proceeds as follows:

1. The person's surname is used to query dataset (1), returning a race/ethnicity[5] distribution
2. The person's street address is matched to a Census block group using the federal geocoding server[6]
3. The resulting Census block group is used to query dataset (2), returning a race/ethnicity distribution
4. Bayes' Theorem[7] is applied to combine the information from the two race/ethnicity distributions
5. The result is a vector with entries corresponding to the race/ethnicity categories in datasets (1) and (2). Each entry represents the probability that the person would have declared that race/ethnicity category in their Census response. Being probabilities, the entries are decimals between 0 and 1. Because the categories are mutually exclusive and collectively exhaustive, the entries sum to one.

Decision rules: Using the BISG vector

There are different methods of using this vector of probabilities. There are basically two approaches, which we will call "definite labels" and "vectorized."

Definite labels involve declaring a rule that results in a single race/ethnicity label for each individual. One possible rule is "choose the race/ethnicity category with the

---

geographic area for which the Bureau of the Census collects and tabulates decennial census data. More detail on Census blocks and block groups is available at https://www2.census.gov/geo/pdfs/reference/GARM/Ch11GARM.pdf .

[3] The surname dataset published by the Census bureau is available at https://www.census.gov/topics/population/genealogy/data/2010_surnames.html. It excludes uncommon surnames, defined as those with less than 100 occurrences. The block group dataset is available on https://data.census.gov/.

[4] A thorough description is available in Technical Appendix A of the 2014 CFPB report "Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment"

[5] Race and ethnicity are distinct concepts related to ancestry; either or both may come into play when using BISG or related inference methods. Choosing a set of demographic categories to use is itself an important, evolving topic (see Conclusion and footnote 34). In this paper we use "race/ethnicity" as a shorthand for a given set of such categories, without claiming that any particular conceptual approach is right.

[6] API documentation is available at: https://geocoding.geo.census.gov/geocoder/Geocoding_Services_API.pdf

[7] https://en.wikipedia.org/wiki/Bayes%27_theorem

highest probability." Another possibility is "if some race/ethnicity category has probability at least 75%, choose it; otherwise label this person 'unknown'." Applying a probability threshold implies a tradeoff: the higher the threshold, the less likely a given label is to be wrong, but the more people are labeled 'unknown'.

The vectorized approach uses the whole vector of probabilities instead of applying a single label to each person. When aggregating information by race/ethnicity group, the vector entries are used as weights, to distribute a given person's information across race/ethnicity groups. An example to illustrate: imagine the task is to estimate the race/ethnicity breakdown of a group of people, and consider a person whose BISG vector entry is 0.85 for Black, 0.15 for White, and 0 for all other categories. Under "definite labels" this person would contribute 1 towards the count of Black people; under "vectorized" they would contribute 0.85 towards the count of Black people and 0.15 towards the count of White people.

Thinking about "accuracy" of BISG

It is natural to ask, "How accurate is BISG?" but it is not obvious how to answer. The literature on classifiers[8] offers many ways to measure the accuracy of classification labels. Among the most basic are metrics like precision, recall, and F1 score, which are calculated by comparing inferred labels to ground truth[9] across many individuals. These basically consider the proportion of cases where classifiers like BISG "guessed right" versus "guessed wrong" relative to ground-truth.

However, as discussed, BISG is not necessarily a classifier. Under the vectorized approach, one can think about accuracy in terms of calibration. BISG could be "well calibrated" in the sense that if you took many individuals with a 90% probability of being Hispanic (per BISG), 90% would indeed be self-reported Hispanic.

Depending on the context, the alignment between individual race/ethnicity inferences and ground truth may not be the relevant notion of accuracy. In many applications (see the Contextual section for further discussion), BISG is an intermediate step towards measuring differences across races in something else – such as political party membership, insurance premiums, or interest rate on a loan. Cory McCartan, Assistant Professor of Data Science, and co-author of the BISG-related inference method BIRDiE made the point that whether or not BISG is good at predicting race on an individual basis is neither necessary nor sufficient for doing a good job with measuring racial disparities. Specifically, there are scenarios where you could make more errors on race

[8] Classifiers are algorithms that identify which of a set of categories a given observation belongs to. A general overview is available at: https://en.wikipedia.org/wiki/Statistical_classification

[9] Ground-truth means the answer we are accepting as true. In the context of race/ethnicity, ground-truth usually refers to self-report: the race/ethnicity a person declares when asked, for instance on their Census survey.

inference (i.e., label more individuals wrongly), yet end up with a more accurate estimate of the racial disparity in an outcome. Indeed, this is a possibility with the BIRDiE, as discussed in the next section.

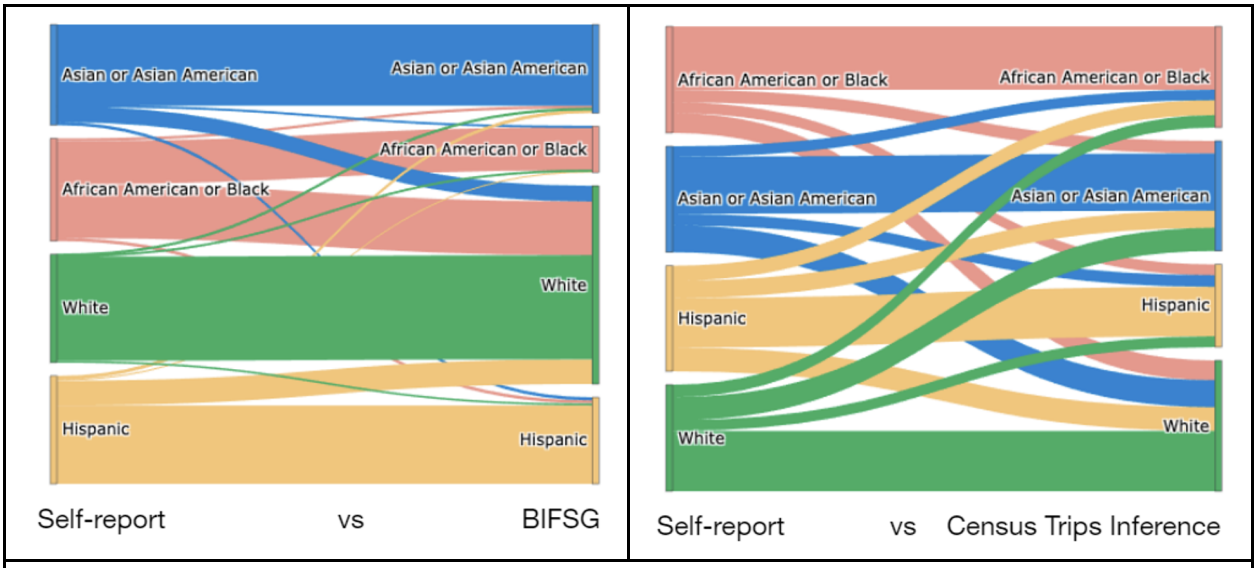
The broader point is that "accuracy" is itself a contextual notion. Accurate in what sense, and to what end? Considering these questions will focus discussions and help practitioners navigate trade-offs like the one between accurate labels for individuals, and an accurate estimate of overall population makeup and racial disparity.

BISG Variants, Alternatives, and an Ensemble Approach

BISG is one of many options analysts and practitioners have for filling in missing race/ethnicity data. In this section we mention other options that were discussed at the conference. Some of these, which we call BISG "variants", use similar data and/or statistical techniques to BISG, but incorporate additional information in the inferences. Others, which we call "alternatives", rely on different information or techniques.
All of these inference methods are imperfect, and the patterns of errors differ between methods. Using BISG as part of an ensemble of methods can exploit these differences, which achieves two goals: helping to understand the mechanism of a disparity and providing a more robust estimate of a disparity.

The way this approach works is illustrated in a paper by Rieke et al[10], which uses self-reports from Uber riders, Uber trip data, and different inference methods to assess demographic differences in outcomes, such as exposure to pollution, iPhone usage, and so forth. In that paper, the specific methods used were Ethnicolr[11], BIFSG, and a location-based method based on the demographics of a rider's Uber trip pickups and destinations. These inferences were compared to riders' self-reports, as depicted in the graphs below. Of note – different inference methods made different types of mistakes, and this is a useful discrepancy.

[10] Aaron Rieke, Vincent Southerland, Dan Svirsky, and Mingwei Hsu. 2022. Imperfect Inferences: A Practical Assessment. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 767–777. https://doi.org/10.1145/3531146.3533140
[11] https://github.com/appeler/ethnicolr



Each graph depicts the accuracy of a demographic inference method versus the self-reported race data among four groups: African American or Black, Asian or Asian American, Hispanic, and White.

Why use this approach? The first reason relates to the discrepancies between methods: because each inference method relies on different signals, each inference method will pick up different types of discrimination. For example, the location-based approaches were more effective at measuring discrepancies in exposure to pollution across space. A name-based approach will pick up on cases where a name is used to discriminate (e.g., Edelman et al[12] finding different host rejection rates on Airbnb when a name is modified).

In short, an ensemble approach can help discipline or guide an analyst's thinking about why disparities might be occurring. Approaches like BIRDiE or a raking adjustment (discussed below) work by using an understanding of real-world disparities to adjust inference methods. The ensemble approach uses differences in inference methods to develop an understanding of real-world disparities.

Ensembles can include bespoke methodologies that leverage data the analyst has at hand, as Rieke et al. do with Uber trip data. A hypothetical example along similar lines: imagine a car insurance company that has no self-reported race/ethnicity data about its customers, but that can see where policyholders drive (provided they consented to monitoring via telematics). They could make an inference based on the average demographics of every census tract a given driver is in for more than X minutes. Such an approach is different and creative and could be more informative -- knowing where and how someone moves through a city might reveal more than a self-report.

[12] Edelman, Benjamin, Michael Luca, and Dan Svirsky. 2017. "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment." American Economic Journal: Applied Economics, 9 (2): 1-22. DOI: 10.1257/app.20160213

The second advantage of an ensemble approach is a more robust understanding of whether disparities exist. Rieke et al. find that different inference methods give different answers on how much of a disparity there is. But in each outcome measured, either all 5 methods agreed on the direction of the disparity or 4 out of 5 agreed. This suggests that in some cases, using just one method could lead to a conclusion that no disparity exists when 4 other methods would lead to the opposite conclusion.

When starting a race disparity analysis, it is worth considering:

- What are the reasonable candidate methodologies to use?
- What different information do they each leverage, and how does that relate to our hypotheses about *how* the disparity or discrimination is (potentially) happening?
- What could we learn about the mechanism behind the disparity, by measuring it using an ensemble of methodologies and comparing the different answers?

In the rest of this section, we discuss BISG variants and alternatives that could be part of such ensembles.

Variant: BIFSG ("Bayesian-Improved Firstname Surname Geocoding")

This variant, proposed by Voicu,[13] incorporates information about an individual's first name, in addition to their surname and address, to further refine the inference. To do this, BIFSG requires a list of first names with their associated demographics. The Census does not publish such a table; instead the original implementation uses a dataset derived from mortgage application data.[14]

In the paper proposing BIFSG, Voicu does a validation study using mortgage application data that includes self-reported race/ethnicity as ground-truth. He finds BIFSG is slightly more accurate – in multiple of the senses discussed above – than BISG. Notably, "the largest improvements occur for non-Hispanic Blacks, a group for which the BISG performance is weakest."

Variant: BIRDiE (Bayesian Instrumental Regression for Disparity Estimation)

This variant[15] is designed for settings where the goal is to measure the disparity between race/ethnicity groups in some other outcome – not just to infer race/ethnicity. In some sense, the individual-level inferences are a by-product of this methodology; the disparity estimate is the focus here. By adopting a different identification strategy

---

[13] Ioan Voicu (2018) Using First Name Information to Improve Race and Ethnicity Classification, Statistics and Public Policy, 5:1, 1-13, DOI: 10.1080/2330443X.2018.1427012

[14] Tzioumis, K. Demographic aspects of first names. *Sci Data* 5, 180025 (2018). https://doi.org/10.1038/sdata.2018.25

[15] Paper available at: https://arxiv.org/pdf/2303.02580.pdf ; code (in R) available at: https://github.com/CoryMcCartan/birdie

than BISG – specifically, different assumptions about the statistical independence between a person's outcome and their surname, race, location, and other characteristics – it actually incorporates information from the distribution of the outcome, into the inference.

To build intuition with this tricky concept, consider a toy example: estimating differences across race/ethnicity groups in political party membership. Suppose this is in a state where White residents skew Republican and Black residents skew Democratic. And suppose you had a list of voters' surnames and addresses, with their party affiliations, but no race/ethnicity information. Now imagine a Black person with a predominately-White surname,[16] living in a predominately-White Census block, and they are registered Democrat. Standard BISG applied to this dataset would assign this person a high probability of being White, given the demographics of their surname and location. In comparison, BIRDiE would assign this person a slightly lower probability of being White, and a slightly higher probability of being Black, because it incorporates the information that they are a registered Democrat.

Precisely how BIRDiE's identification strategy differs from standard BISG's is beyond the scope of this paper, but it points to an important issue. Each variant of BISG relies on making certain assumptions about conditional independence between variables, which really means making certain assumptions about how the world works. For example, standard BISG assumes that "once we know an individual is White, knowing their surname is Smith tells us nothing about their residence location and other observed characteristics." However, "unlike the Smith example, knowing that an Asian individual's surname is Gupta makes it more likely that they have a higher income and live in the Eastern U.S (Budiman et al., 2019)."

Variant: Fully-Bayesian BISG with augmented name data (fBISG)

This variant[17] addresses two issues with standard BISG that hamper its performance for minority populations. First, published Census tables include inaccurate zero counts for minorities in many census blocks. Reasons for the inaccuracies include undercounting, the fact that people move, and the addition of statistical noise in Census publications for the sake of privacy. In standard BISG, because of the multiplication in applying Bayes' Theorem, zero counts force estimated probability to zero. For instance, a person living in a census block with zero Hispanics will have zero probability of being Hispanic, no matter their surname. Given that zero counts are often wrong – and are always undercounts when they are – this leads to undercounting minorities overall. fBISG addresses this issue by treating the counts in the Census tables as

[16] This is fairly common due to the history of slavery in the US, particularly the fact that many Black enslaved people adopted – or were forced to take – the surnames of their White owners.
[17] Paper available at: https://www.science.org/doi/10.1126/sciadv.adc9824; code (in R) available at: https://cran.r-project.org/web/packages/wru/index.html

measurements with error rather than as ground truth. In effect this "softens" the zero counts and thus improves performance for minority populations.

The second issue relates to uncommon names. The Census surname list used in BISG includes all common surnames, defined as those with at least 100 occurrences in the decennial Census. This list covers about 90% of Americans overall but, importantly, coverage varies by race. For example, 97% of Black Americans' surnames are on the list, compared with just 86% of Asian Americans'. If a person's surname is not on the list, standard BISG effectively skips that step and makes a prediction based solely on the address. These predictions are less accurate; and this happens more for some groups, particularly Asian Americans. fBISG addresses this issue by incorporating additional name information derived from state voter files that include self-reported race.[18] Their version includes a longer list of surnames, as well as first and middle names, all with associated demographic breakdowns.

Validating their method against voter file data, the authors of fBISG show that these two changes lead to significant improvements in the quality of inferences, especially for minority populations.

Variant: BISG with Raking Adjustment

Like with the BIRDiE method, the raking-adjusted BISG estimate proposed by Greengard and Gelman (2023)[19] attempts to correct for the violation of the implicit independence assumption in BISG. Raking is a reweighting method often used in survey statistics whose goal is to adjust observed count data to match a marginal population distribution. Thus, raking-adjusting BISG estimates can be used when marginal population distributions are known in order to improve BISG estimates.

To demonstrate the raking method, Greengard and Gelman use North Carolina and Florida voter files as sources of surname and geolocation information (the voter's county). Voter files for these two states contain self-reported race information, allowing for accuracy calculations. The authors obtain information on marginal distribution from the U.S. Census' Current Population Survey (CPS) Voter Supplement, which contains race information on a sample of registered voters in the US. The marginal race distribution from the CPS data for a state is used to adjust the BISG values so that the distribution of race according to BISG matches the distribution obtained from the CPS data.

For North Carolina and Florida voter files, the authors show that the raking adjusted BISG estimates are more accurate for identifying the racial makeup of registered voters

[18] The voter files they used are from AL, FL, GA, LA, NC, and SC.
[19] Paper available at https://arxiv.org/pdf/2304.09126.pdf

in each county. Notably, as population increases, BISG severely underestimates minority populations while raking-adjusted BISG does not.

Note that the raking method requires the racial makeup of the overall population from which the sample was drawn to be known. Thus, in situations where marginal population statistics are not available, the raking method cannot be used.

Alternative: Broker data

A data broker is a company that specializes in the acquisition of personal information, encompassing data elements such as income, ethnicity, political affiliations, and location. This data is predominantly gleaned from publicly available sources, albeit occasionally procured through private channels, with the ultimate purpose of subsequently offering it for sale or licensing to third-party entities for various applications, often marketing.

We have used broker data to assist with estimating racial disparities and have compared it to BISG results on multiple occasions. We have found that broker data tends to be less complete than BISG estimates, where 15-25% of a sample are unable to be assigned a race/ethnicity, compared to 5-10% of a sample for BISG or BIFSG. In addition, we have observed persistent differences in the estimated racial makeup of samples computed with broker data as compared to those of the same sample computed with BISG. Typically, in four-category estimates of non-Hispanic White/Other, non-Hispanic Black, non-Hispanic API, and Hispanic, the estimates for API and Hispanic are very similar from both methods (difference of less than one percentage point), while the broker population consistently measures as three to five percentage points more White with a corresponding decrease in percent Black than the BISG estimated population.

There are notable limitations to broker data as well. In addition to being less complete than BISG estimates, broker data varies in accuracy across the four main race categories (Hispanic, White, Black, and Asian). For example, Hispanic individuals' age is correctly listed in Experian data 73% of the time, compared to 82% for White individuals' age. In addition, there are concerns about broker data's potential to further disadvantage populations of color and those living in poverty, due to worse quality of data available for populations with higher poverty rates and for populations of color.[20]

[20] Giridhari Venkatadri, Piotr Sapiezynski, Elissa M. Redmiles, et al 2019. Auditing Offline Data Brokers via Facebook's Advertising Platform. In The World Wide Web Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, 1920–1930. https://doi.org/10.1145/3308558.3313666

Alternative: Perceived race

The implicit ground-truth of BISG is the U.S. decennial census. That is, a BISG output vector is an answer to, "How would this person have self-reported their race/ethnicity on the Census Bureau form?" Depending on the kind of discrimination we are concerned about, this might not be the best question to ask. For example, Airbnb was concerned about hosts discriminating against certain users in their decisions about whom to accept as guests. To test for discrimination in this context, a more relevant notion of race/ethnicity is how hosts would *perceive* a given user, not how that user would self-report.

Appropriately to this context, Airbnb devised a way to measure the *perceived race* of users for the purpose of fairness testing around this issue. They engaged an outside firm to review users' profile pictures and apply race/ethnicity labels based on their assessment of the picture. Airbnb also designed a secure, privacy-preserving system for the data that ensured the perceived-race labels were always aggregated and could not be associated with individual users, nor accessed by business units other than the anti-discrimination team.[21]

Contextual

One goal of the conference was to share information about how BISG and related methods are being adopted and adapted by practitioners in different settings and for different purposes.

Used to investigate potential unfair discrimination in life and auto insurance

The Colorado Division of Insurance (CO DoI) and Washington, D.C., Department of Insurance, Securities and Banking (DISB) have recently used BIFSG in connection with inquiries into potential unfair discrimination in life[22] and auto[23] insurance, respectively.[24] In these analyses inferred race/ethnicity labels are applied to policy-level data, which facilitates comparisons across groups. For instance, the regulators could measure differences across race/ethnicity groups in the average premiums customers pay.

These analyses are a significant step forward given the paucity of race/ethnicity information in insurance data generally. Broadly speaking, today insurers are not

---

[21] https://news.airbnb.com/wp-content/uploads/sites/4/2020/06/Project-Lighthouse-Airbnb-2020-06-12.pdf

[22] https://doi.colorado.gov/for-consumers/sb21-169-protecting-consumers-from-unfair-discrimination-in-insurance-practices

[23] https://disb.dc.gov/page/evaluating-unintentional-bias-private-passenger-automobile-insurance

[24] Full disclosure: ORCAA assisted the regulators with these inquiries.

allowed to make decisions (about rating, underwriting, or pricing, for instance) on the basis of a person's race or ethnicity. Although to our knowledge there is no prohibition on asking about – or inferring – race/ethnicity, insurers generally avoid collecting or holding any such information about individual customers. This gives them plausible deniability against claims of intentional discrimination while limiting their ability to measure outcomes by race/ethnicity. Analyses of insurance data that aim to do this without using inference often rely on geographic (e.g. zip code) modeling of race.[25] This is basically a coarser version of BISG: all policies within a given geography are associated with its aggregate demographics.

Given the industry's sensitivities, both CO DoI and DISB set the stage by socializing the idea of using inference methods for their analysis. Each held public stakeholder meetings, including presentations to stakeholders about BISG methodology; they also considered comments, and engaged industry directly for input. This gave stakeholders a chance to ask questions and voice concerns for the record.

Regulatory authority, not just voluntary stakeholder engagement, was crucial to making these analyses possible. DISB developed a data call collaboratively, including publishing a draft and considering stakeholders' comments; then it used its market conduct examination authority to require all carriers writing private passenger auto policies in D.C. to respond to the final data call.

It is also worth mentioning that both of these inquiries were structured so that insurers did not have to perform inference or hold individual-level race/ethnicity data. Instead, insurers submitted the relevant data to the regulator, who performed BIFSG and did the disparity analysis. Going forward, insurers may have to do more of the work themselves: CO DoI's current draft testing regulation for life insurers[26] requires them to estimate their insureds' race/ethnicity using BIFSG, do quantitative testing, and report the results annually to CO DoI.

Used in voting rights litigation

One panel at the conference, focused on BISG used in relation to voting rights, included Michael Rios of the UCLA Voting Rights Project, Loren Collingwood of the University of New Mexico, and Cory McCartan of New York University. In this context, inference methods are used to analyze turnout and gerrymandering. For instance, given a proposed redrawing of district lines, BISG may be used to assess the likely impact on race/ethnicity breakdown of voters in each district.

---

[25] For a discussion including many examples of zip-code-based race analyses, see Section 3 of "Matching Rate to Risk: Analysis of the Availability and Affordability of Private Passenger Automobile Insurance", by Robert Klein, published by the National Association of Mutual Insurance Companies (NAMIC)

[26] DRAFT Proposed Algorithm and Predictive Model Quantitative Testing Regulation.pdf

The panelists discussed ways of overcoming skepticism about BISG, both from judges and opposing counsel. At a high level, it comes down to convincing people the inferences are accurate enough. One panelist mentioned a validation analysis that used the California voter file to infer race/ethnicity using BISG, then compared the inferences to self-reports. Another said that even small "gut-check" demonstrations can be helpful – for instance, showing that in a set of BISG inferences, people with the surname "Hernandez" have a very high probability of being Hispanic.

The panelists agreed that in this area self-report data is preferable if it is available – for instance, in states that require voters to declare their race/ethnicity.[27] Where this data is not available, a little ground-truthing goes a long way. For example, a small study comparing inferences to self-reports within the relevant context/geography could greatly strengthen the case for using BISG. Even when self-report data is available, it has gaps since some people simply leave race/ethnicity questions blank or abandon the survey. One can measure and account for these gaps by estimating the demographics of the people who did not disclose their race.

Used in algorithmic auditing

Another panel focused on the use of inference methods in algorithmic auditing, where it is often used to investigate the performance of algorithms and automated decision systems across race/ethnicity groups. For instance, are there differences across groups in the errors the system makes? Are outcomes different across groups?

The panelists represented a range of what an "algorithmic auditor" can be. Judah Axelrod from the Urban Institute, a nonprofit think tank that does audits in the public interest using publicly-available information, spoke about work in algorithmic home appraisals with a focus on racial equity. Other panelists were Jacob Appel from ORCAA, a consultancy whose clients include both organizations commissioning audits of their own algorithms, as well as regulators and enforcers that investigate others' algorithms, and Kasey Matthews from zest.ai, a company that specializes in AI-driven lending and developed its own open-source race inference method.[28]

The panelists discussed ways to do these analyses while accommodating the sensitivities of audit targets around holding and using data with race/ethnicity labels. For instance, ORCAA created an analysis platform with a "double firewall" so that clients could upload raw data and receive fairness analysis based on inferred race/ethnicity, but (1) the client never sees any individual's race/ethnicity label, and (2) the analysis platform never sees any individual's name or address. Another example of

[27] According to Pew research, "In 16 states or portions of states, largely in the South, the Voting Rights Act of 1965 mandated that states list voters' race on the state voter rolls." https://www.pewresearch.org/methods/2018/02/15/demographic-data/
[28] https://github.com/zestai/zrp

designing around such sensitivities is Airbnb's system for perceived race data, described in the BISG Alternatives section.

Beyond client sensitivities, algorithmic auditors might run into legal challenges in some settings. In the EU, GDPR could hamstring auditors since individual-level race/ethnicity data is considered sensitive and is subject to additional requirements. Panelists mentioned a recent paper[29] that argues that a system along the lines of Airbnb's would be compliant with most EU member states' rules.

Used in fair lending analysis

US laws prohibit lenders from discriminating on the basis of race. Race inference is used in statistical tests that are part of fair lending analysis, showing whether a given lender is complying with – or violating – these laws. These analyses may be done in-house by lenders to monitor their own compliance, on behalf of lenders by third parties (like zest.ai), or by regulators or enforcers investigating potential violations. These analyses generally follow the legal doctrine of disparate impact. They use race/ethnicity labels to compare averages (e.g. average APR) between similarly-situated members of different groups.

This area is notable for having a broadly accepted approach to testing for fairness. This is indeed what codified equity rules look like. We can expect to see more along these lines in other regulated industries – like housing, hiring,[30] and insurance[31] – as laws and enforcement catch up with the ongoing adoption of algorithmic decision systems. For example, on October 30, 2023, just weeks after the conference, the White House published an Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.[32] The Order calls for the development of guidelines and tools to enable federal agencies – including those covering consumer finance (CFPB) and housing (FHA, HUD) – to implement minimum risk-management practices around the use of AI in their sectors.

Policy Discussion: BISG in Insurance

Another conference panel, which included Maryland's Commissioner Kathleen Birrane, D.C.'s Associate Commissioner Philip Barlow, Colorado Division of Insurance's Big Data and AI Policy Director Jason Lapham, discussed how policy in insurance can be informed by the existence of race inference methodologies like BISG. In particular, BISG can and has been used to measure disparities in outcomes between race categories.

---

[29] https://hstalks.com/article/7972/pilot-project-lighthouse-a-proposed-gdpr-compliant/
[30] New York City Local Law 144, requiring Bias Audits of Automated Employment Decision Tools, is an example.
[31] Colorado's draft testing regulations, discussed in the "Policy Discussion" section, is an example.
[32] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

BISG and its cousin BIFSG represent a big step forward in the discussions among insurance commissioners and insurers, because up to this point the conversation has been theoretical; most insurers have not really been investigating the question, in part because they wanted to keep at arm's length the notion that they are collecting or inferring the race of their customers.

Also, it's not even clear that there's a problem, so the panelists emphasized that the approach is thoughtful, careful, and does not assume there's a problem until one arises. For that matter, it's not at all obvious what exactly would comprise a "problem," since premiums might reflect losses pretty accurately but still land more heavily on minority populations.

With that said, what does testing look like? What is the definition of an unexplained difference in outcome, and what is the threshold for that difference? One reference, which was not available at the time of the conference but is now, is the Colorado draft regulation for life insurance, available here[33]. But having said that, a reasonable approach to measuring differences in outcome, after using BISG to infer race, has not yet been codified. It's also not clear how prescriptive the insurance regulators want to be in general, and they are actively having conversations with the industry to discuss appropriate setups and implementations.

With that, the panel moved to a series of open questions, including the critical question of whether a given factor, which is correlated to both losses and race, for example blood pressure in the case of life insurance, can be used either as a predictor by insurers or as a "legitimate explainer of differences" in outcomes by race. Could this question end up resulting in a balancing test on a per-factor basis? Could it rely on a list of "traditional underwriting factors"? The answer is probably both.

And, according to some panelists, the social justice element of this issue might best be left for new regulation, which arguably already exists in Colorado, or could be tackled by existing law under the aegis of the Insurance Commissioner's office.

In other words, thanks to BISG, the trickiest thing for a regulator to address the question of rates not being "inadequate, excessive, or unfairly discriminatory" – the relevant standard in insurance – is how exactly to interpret that last part, not the math.

Conclusion

Discussions at the conference ranged from technical details and causal diagrams, to questions of political economy and winning over skeptical judges. This reflected the diversity of participants, both across functional roles – regulators, practitioners,

[33] https://drive.google.com/file/d/1BMFuRKbh39Q7YckPqrhrCRuWp29vJ44O/view

litigators, advisers/consultants, researchers – and industries. Many attendees remarked on how valuable, and unusual, it was to gather such a diverse set of perspectives.

Two takeaways from the day were (1) BISG and its variants/alternatives are an important part of fairness analyses today, are used and accepted in many contexts, and are not going away anytime soon. And (2) we identified a number of open questions and common issues that practitioners are grappling with across settings:

- When should we use which variant (BIFSG, fBISG, BIRDiE, etc)? One could imagine creating a flowchart or structured questionnaire to guide practitioners to the appropriate variant based on characteristics of their intended analysis.
- Under what conditions could BISG *overestimate* race disparities? This is a key question in practice, since those conducting fairness analyses (e.g., algorithmic auditors) often want to provide assurance to the targets of those analyses that they will not wrongly conclude there is a problem.
- When should we use which decision rule (definite label vs. vectorized), and what is the impact of different decision rule choices on disparity measurements? In some settings there may not be much choice; for instance, if individuals must get discrete race/ethnicity labels for reasons of due process, then you must use definite labels. Even within definite labels, to choose a threshold you must navigate a tradeoff between the accuracy of each label and the number of individuals labeled "unknown".
- What adjustments should we make when applying BISG to a population that may not mirror the US Census? The "BISG with Raking Adjustment" method described in "BISG variants" is an option, but it requires demographic information of the population in question.

In terms of making race inference useful to practitioners in the field, some issues that emerged from the panel discussions were:

- Keeping these methods current amid a changing demographic landscape: updating from 2010 to 2020 Census data, and incorporating new race/ethnicity categories.[34]
- What options exist for inferring race outside the US?
- When requiring the use of inference methodology, e.g., in a regulation or a corporate risk management policy, how specific should you be? For instance, should you prescribe which variant of BISG to use, which race categories, which decision rule?
- Agreeing on language: Should the output of BISG be presented as a calculation, an estimate, a prediction – or something else?

[34] For a general discussion, see "An American Puzzle: Fitting Race in a Box" by K. K. Rebecca Lai and Jennifer Medina in the New York Times

These questions call for further research and collaboration. The nascent community of practice represented by the conference participants would be well suited to take them on.