RESEARCH METHODS

Addressing census data problems in race imputation via fully Bayesian Improved Surname Geocoding and name supplements

Kosuke Imai[1]*, Santiago Olivella[2], Evan T. R. Rosenman[3]

Prediction of individuals' race and ethnicity plays an important role in studies of racial disparity. Bayesian Improved Surname Geocoding (BISG), which relies on detailed census information, has emerged as a leading methodology for this prediction task. Unfortunately, BISG suffers from two data problems. First, the census often contains zero counts for minority groups in the locations where members of those groups reside. Second, many surnames—especially those of minorities—are missing from the census data. We introduce a fully Bayesian BISG (fBISG) methodology that accounts for census measurement error by extending the naïve Bayesian inference of the BISG methodology. We also use additional data on last, first, and middle names taken from the voter files of six Southern states where self-reported race is available. Our empirical validation shows that the fBISG methodology and name supplements substantially improve the accuracy of race imputation, especially for racial minorities.

INTRODUCTION

Social scientists and public health researchers often must predict individual race and ethnicity when assessing disparities in policy and health outcomes. Bayesian Improved Surname Geocoding (BISG), which uses Bayes' rule to combine information from the census surname list with the geocoding of individual residence, has emerged as a leading methodology for this prediction task (*1*–*4*). Recent applications of the BISG methodology include studies on racial disparity in police violence (*5*), eviction (*6*), suicide (*7*), and turnout (*8*).

Here, we address two census data problems that hinder accurate prediction of individual race and ethnicity when using the BISG. First, the decennial census often contains zero counts for minority groups in the census blocks where some members of those groups reside. This may happen for several reasons. Some individuals may have moved after the decennial census. There may also be undercounts. Another possibility is that the census may inject measurement error for privacy protection [see (*9*) and references therein].

Second, the decennial census surname files only include the racial composition of surnames that occur 100 or more times in the population. According to the Census Bureau, these names account for about 90% of people with surnames recorded in the 2010 Census (*10*). This means that no racial breakdown statistic is available for the remaining 10%. This lack of information may disproportionately affect minority groups if their surnames are less frequently occurring than those of the majority group.

Using the data from six Southern states in which individual self-reported race of registered voters is available for validation, we show how these problems can result in a deterioration of predictive quality for the standard BISG approach. To address the census zero count problem, we introduce a fully Bayesian generalization of the BISG approach (fBISG). The proposed fBISG models the observed census counts using a measurement error model so that zero counts for minority groups do not necessarily imply nonexistence of their members. The model is fitted via a collapsed Gibbs sampler and can incorporate additional names and covariates such as the standard BISG.

To address the problem of missing surnames, we supplement the census surname list with the lists of additional surnames, middle names, and first names from these voter files. This allows us to markedly reduce the proportion of missing surnames and further improve the race prediction. The proposed methodology, including the additional name lists, is publicly available as part of the open-source software package wru: Who Are You? Bayesian Prediction of Racial Category Using Surname and Geolocation (*11*). A complete description of the proposed methodology is given in Materials and Methods below.

Our empirical validation study demonstrates that these proposed solutions yield substantial improvements in classification accuracy without reducing a high degree of calibration—particularly among racial minorities. Specifically, a model that incorporates all our proposed improvements increases classification accuracy by an average of about 14% among all five major racial groups (vis-à-vis the standard BISG), with improvements as high as 26% among Asian voters. Moreover, these gains in classification accuracy do not come at the expense of the calibration of predicted probabilities across racial groups (i.e., the extent to which predicted probabilities match observed sample proportions of cases), which is already high for predictions made by the standard BISG methodology for white and Black voters.

Last, we conclude with a brief discussion about the applicability of our proposed modeling approach to various domains.

[1]Department of Government and Department of Statistics, Institute for Quantitative of Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA. [2]Department of Political Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA. [3]Harvard Data Science Initiative, Harvard University, Cambridge, MA, 02138, USA.
*Corresponding author. Email: imai@harvard.edu

RESULTS

The census data problems in race imputation

In this subsection, we first briefly review the standard BISG methodology. We then describe the census data problems and quantify the degree to which they negatively affect the predictive performance of BISG.

BISG: A review

The goal of BISG is to predict the race of individual i, defined as $R_i \in \mathscr{R}$ where $|\mathscr{R}| = J$ is the total number of (mutually exclusive) racial categories. Here, we will have $J = 5$, with the categories $\mathscr{R}$ = {"White," "Black," "Hispanic," "Asian," "Other"}.

Suppose that we observe the individual's surname $S_i \in \mathscr{S} = \{1,2, ..., K\}$ and geolocation $G_i \in \mathscr{G} = \{1,2, ..., L\}$ where the latter is typically recorded as a census geographical unit (e.g., census block), in which his or her residence is located. The BISG methodology is an application of naïve Bayes prediction, where the key assumption is given by the following conditional independence relation between geolocation and surname, given race.

Assumption 1. (Independence between Surname and Geolocation within Racial Group).

$$G_i \perp\!\!\!\perp S_i \mid R_i$$

Under Assumption 1, the BISG prediction of an individual's race is given by

$$P(R_i \mid S_i, G_i) \propto P(S_i \mid R_i, G_i)P(R_i \mid G_i) = P(S_i \mid R_i)P(R_i \mid G_i) \quad (1)$$

One may also use the following equivalent formula obtained via another application of Bayes' rule

$$P(R_i \mid S_i, G_i) \propto P(R_i \mid S_i)P(G_i \mid R_i)$$

In practice, the decennial census surname files are used to compute $P(S_i \mid R_i)$, whereas for $P(R_i \mid G_i)$, it is common to use the Census Bureau's cross-tabulations of racial category by geographic location (e.g., census blocks). Although we do not address the potential violation of Assumption 1 in this paper, it is important to acknowledge its limitations. The assumption is violated if, for example, among Asian Americans, various ethnic groups (Chinese, Indians, Japanese, Korean, Vietnamese, etc.) have distinct surnames and tend to live in different areas. A similar problem might also arise among Hispanic Americans. The surname "Santos," for instance, may be common among Hispanics in some areas, but it may also be a common last name among Brazilian Americans (who are classified as non-Hispanic whites according to the census) in other areas. We now turn to two census data problems that negatively affect the predictive performance of BISG.

Table 1. Count of individuals of each race (in millions) by 2010 Census racial counts in the individual's home census block. The top row gives the count of individuals for whom the 2010 Census states that there are zero members of that racial group living within the individual's home block; the bottom row gives the count for whom the census states that there is at least one member of that racial group in the block. Data are sourced from voter files from AL, FL, GA, LA, NC, and SC, and racial data are self-reported on the file.

Census tally	White	Black	Hispanic	Asian	Other
Zero counts	0.12 (1%)	0.32 (4%)	0.16 (5%)	0.13 (20%)	0.22 (30%)
Nonzero counts	22.13 (99%)	7.55 (96%)	2.84 (95%)	0.51 (80%)	0.51 (70%)

Consequences of zero census counts

The decennial census is intended to provide a full accounting of where each resident of the United States lives as of April 1 on the census year. Reported census distributions are considered reasonably reliable as of this date, although still imperfect [see, e.g., (*9*)]. Over time, however, this accuracy degrades even further, as individuals move within the nation's borders at significant rates (*12*). At fine levels of resolution, such as census blocks, this means that the racial distributions may not fully capture the diversity of residents within a short time after the census is conducted. Among rapidly growing minority groups, such as Asian Americans and Hispanic Americans, errors may be particularly large.

Prior studies have shown that the use of the census block level data, rather than the data at a higher level of geographical aggregation, tends to yield more accurate BISG predictions of individual race and ethnicity [e.g., (*4*)]. This, however, can result in a greater chance of measurement error. In particular, when census counts are used to obtain the prior distribution $P(R_i \mid G_i)$ at the census block level, some blocks may record zero individuals of certain ethnic and racial categories. In these cases, $P(R_i \mid G_i)$ would be set to zero, making the posterior probability of belonging to these groups automatically zero for all individuals who reside in these blocks. For example, someone with the last name "Gutiérrez"—a distinctively Hispanic last name—living in a neighborhood where the census failed to count anyone of Hispanic descent would have a zero posterior probability of being classified as such according to the standard BISG methodology.

Thus, the predictive accuracy of the BISG methodology can suffer markedly when probabilities are zeroed-out a priori. To quantify this error, we consider the voter files of six Southern states—Alabama, Florida, Georgia, Louisiana, North Carolina, and South Carolina—sourced between October 2020 and February 2021 (before the release of 2020 census data). The voter files were provided by L2 Inc., a leading national nonpartisan firm and the oldest organization in the United States that supplies voter data and related technology to candidates, political parties, pollsters, and consultants for use in campaigns. These files tally all registered voters (approximately 37.8 million voters) in the states as of the production date, geocoding the census blocks of their home addresses. About 91% of these voters provided self-reported race data.

Table 1 shows the counts of voters (in millions) by self-reported race, further divided by whether the 2010 Census indicates that exactly zero members of the individual's racial group live within their home census block. Because of internal mobility and other forms of measurement error, just shy of 1 million voters (2.8%) live in a census block where the 2010 Census tallies indicate that no members of the individual's racial group reside. Notably, these errors are not shared evenly across races. While fewer than 1% of white voters live in a census block in which the census data say no white individuals reside, a full fifth of Asian voters live in census blocks in which the 2010 Census says there are no Asian residents. Even these aggregates mask substantial heterogeneity by state. In South Carolina, for example, 19% of Hispanic voters and 31% of Asian voters reside in zero-Hispanic and zero-Asian blocks, according to the 2010 Census.

This mismatch presents a significant challenge for the BISG methodology. A naïve application of BISG would yield a prediction of 0% for the true racial group of all individuals in the first row of Table 1—comprising a relatively large proportion of all minority voters in the South—simply as a mechanical result the census reporting no members of these racial groups living in the corresponding geographies.

The impact on BISG prediction is summarized in Table 2, where we compute misclassification rates for each racial group, by assigning each individual to the group whose predicted probability is the greatest (i.e., maximum a posteriori) and comparing against their true, self-reported race. The table reports overall error and false-positive and false-negative rates by racial group, which can be quite substantial among minorities. Among Latinos, for instance, the hard zeros induced by the census counts raise false-negative rates from 16% to about 20%. In addition, among Asian voters, the increase is even starker, raising the false-negative rates from 33 to 46%. Overall, and across racial groups, a zero prior probability caused by zero census counts increases classification errors from 14.5 to 16.9%.

Table 2. Overall classification error rate as well as false-positive (type I error) and false-negative (type II error) rates for white, Black, Latino, Asian, and other voters using the standard BISG prediction as implemented in the wru package. Each voter is classified to the racial category with the highest predicted probability. We compare rates for individuals in blocks for whom the census sets a nonzero prior for their true racial group ("nonzero census blocks") against individuals in blocks from who the census sets a zero prior ("zero census blocks"). All individuals are classified to the wrong racial group in zero census blocks, so false-negative rates are 100%, while false-positive rates are undefined. NA, not available.

Ethnicity	Data	Nonzero census blocks	Zero census blocks	Total
Overall error rate		14.5%	100%	16.9%
White	False negative	5.6%	100%	6.1%
	False positive	31.4%	NA	31.4%
Black	False negative	33.7%	100%	36.4%
	False positive	3.5%	NA	3.5%
Hispanic	False negative	15.7%	100%	20.3%
	False positive	2.2%	NA	2.2%
Asian	False negative	33.2%	100%	46.6%
	False positive	0.7%	NA	0.7%
Other	False negative	92.7%	100%	94.9%
	False positive	0.3%	NA	0.3%

These results suggest that individual race prediction can be improved by addressing the possibility that block-level racial priors may be inaccurate or out of date, especially if they are equal to zero. Further evaluation statistics, underscoring the same point, can be found in table S1.

Consequences of missing race-name data

A second source of error arises from the use of surname data. In most applications of the BISG methodology, surname racial distributions are drawn from the Census Bureau's surname list. The 2010 Census surname list, for example, provides the racial distribution of surnames appearing at least 100 times, which amounts to a total of about 160,000 names. In the original version of the wru software package, these data are supplemented with the census's Spanish surname list—a list of about 12,000 common Hispanic surnames, approximately half of which are not in the census surname list.

While these data are quite broad, they do not account for the possibility of rare surnames. In our sample of Southern states as of late 2020 and early 2021, we find that about 2 million voters (5.9%) have surnames that cannot be matched to the census name dictionary, even after the data are cleaned and stripped of punctuation to improve the chance of a match. The distribution of this mismatch across racial groups is given in Table 3. Although Asian voters are particularly unlikely to have their surnames matched (14%), the same is true for a significant portion of white voters (7%).

In the absence of a surname match, the default behavior of a common implementation of BISG [viz. the software package wru (*11*)] is to use the approximate 2010 national race proportions as an estimate for $P(R_i \mid S_i)$. This approximation yields a degradation in predictive performance among these records. As before, we compute misclassification rates for each racial group. These results can be found in Table 4. The results are somewhat less marked in this case, because the default behavior in the absence of a name match does not automatically yield a misclassification. Nonetheless, we can see that misclassifications occur for less than one-sixth of individuals whose names are matched but nearly a quarter of individuals whose names are unmatched, increasing the overall error rate. In particular, among white voters, the false-positive rate is much higher with unmatched surnames than with matched surnames, whereas the false-negative rate increases for individuals of the other racial groups. Further evaluation statistics can be found in table S2. Once again, a data limitation yields a reduction in the predictive performance of the BISG methodology. Accordingly, better name coverage would improve the quality of our predictions.

Table 3. Count of individuals (in millions) of each race for whom the individual's surname cannot be matched to a name in the census surname dictionary or Hispanic surname file. Data are sourced from voter files from AL, FL, GA, LA, NC, and SC, and racial data are self-reported on the file.

Name match?	White	Black	Hispanic	Asian	Other
No	1.47 (7%)	0.27 (3%)	0.13 (4%)	0.09 (14%)	0.08 (10%)
Yes	20.79 (93%)	7.61 (97%)	2.88 (96%)	0.55 (86%)	0.66 (90%)

Table 4. Overall classification error rate as well as false-positive (i.e., type I) and false-negative (i.e., type II) error rates for white, Black, Latino, Asian, and other voters using prediction using standard BISG as implemented in the wru package. Each voter is classified to the racial category with the highest predicted probability. We compare rates for individuals whose names are matched to our name dictionary against those whose names are not matched (in which case a national racial prior is used). Error rates are significantly higher among those whose names are unmatched.

Ethnicity	Error	Name matched to dictionary	Name unmatched to dictionary	Total
Overall error rate		16.4%	24.7%	16.9%
White	False negative	6.1%	6.5%	6.1%
	False positive	30.1%	58.8%	31.4%
Black	False negative	35.5%	63.7%	36.4%
	False positive	3.6%	2.1%	3.5%
Hispanic	False negative	18.4%	65.4%	20.3%
	False positive	2.1%	4.1%	2.2%
Asian	False negative	40.3%	84.6%	46.6%
	False positive	0.6%	2.7%	0.7%
Other	False negative	94.4%	99.3%	94.9%
	False positive	0.3%	0.2%	0.3%

To address these issues, we develop a fully Bayesian version of BISG (termed "fBISG") to account for the census zero counts problem while introducing additional name data to address the missing surname issue. Before describing these proposed solutions in detail in Materials and Methods below, we first show how our corrections to the above data quality issues can substantially improve the prediction accuracy of BISG.

Empirical validation of the proposed solutions

To empirically validate our proposed improvements, we fit both the standard BISG and our fBISG to the combined voter files from AL, FL, GA, LA, NC, and SC, from L2 Inc. As discussed in the previous section, this combined dataset contains information for roughly 38 million voters.

The setup

In our validation, we treat the self-reported race of each record as unobserved and use other available information for that record to obtain posterior probability distributions over their race. Specifically, we use the last, first, and middle names of each voter, as well as the census block in which their reported home address is located. We then compare predictions based on these posterior distributions to the known racial categories of each record to evaluate the overall quality of our fBISG predictions vis-à-vis those of the standard BISG methodology.

As evaluation metrics, we use false-positive and false-negative error rates based on each model's predictions. Furthermore, we also rely on two validation metrics throughout our empirical exercise: the area under the receiver operating characteristic curve (AUROC) and the calibration of predicted probabilities. Both offer distinct windows into the quality of a model's predictions, and we are interested in improving upon classical BISG on both fronts simultaneously.

The AUROC measures the probability that a randomly chosen member of each racial group will have a higher predicted probability of belonging to that racial group than a randomly chosen nonmember. Accordingly, the AUROC gives us a sense of the extent to which predicted probabilities can help us sort cases correctly into instances and noninstances of the category under consideration, with higher values thus indicating better accuracy.

While the AUROC is a useful tool for gauging a model's classification accuracy, it is less useful for understanding the extent to which predicted probabilities accurately reflect true relative frequencies. To measure the accuracy of the predicted probabilities, we compute the observed relative frequency with which records fall under a given racial category. If this relative frequency matches the predicted probability assigned to the observations (so that, for example, 10% of observations with predicted probabilities of being Hispanic equal to 0.1 are, in fact, Hispanic), then we would conclude that these probabilities are well calibrated and, thus, that the model as a whole has good calibration. In practice and because probabilities are rarely exactly the same for a group of observations, we assess calibration by considering a small range of values around a discrete set of target relative frequencies (e.g., values between 0 and 1 by 0.1 increments).

To compute these validation metrics while reducing the risk of overfitting, we estimate models separately by state. When computing predictions on a given state, we do not include that state's voter file in the compilation of the name-race dictionaries used to form $P(S_i \mid R_i)$. For instance, in sampling the race probabilities of voters in North Carolina, we only use the name-given-race distributions derived from augmenting the original census dictionary with records from the other five states. This ensures that the name-given-race distributions are not obtained from the validation voter file. After we obtain predictions using this leave-one-out strategy, we compute validation metrics using the combined predictions from all states.

Last, to obtain samples from the fBISG posterior distribution over races for each voter in our combined voter file, we rely on the latest version of the wru package in R (*11*). We initialize the global counts in the Gibbs updates of Eq. 6 using the predictions based on the standard BISG methodology and run a single Markov chain for 1500 iterations, discarding the first 500 samples as burn-in. We completed all analyses on a laptop computer with an M1 Max central processing unit and 64 gigabytes of random-access memory in less than 3 hours of wall time.

Correcting the zero census counts problem

Figure 1 shows, for each racial category, the AUROC based on posterior predictions generated by the standard BISG (blue bars) and fBISG (red bars) methods. For all but the Other racial category, the classification performance of the fBISG methodology represents a substantial improvement over that of the standard BISG (as we

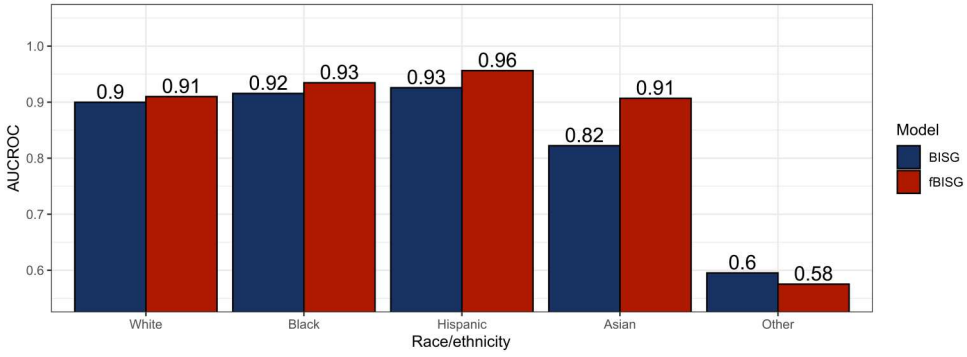

Fig. 1. AUROC for race predictions obtained using the standard BISG methodology (blue) and our fBISG methodology (red). All results are based on the 2010 Census surname dictionary. A greater value of AUROC indicates more accurate race classification. For all but the Other category, the fBISG methodology has better classification performance than the standard BISG methodology, generating the most marked improvements among Asian minorities.

will discuss below, adding more name information rectifies this small dip in AUROC performance relative to the standard BISG). These performance gains are most marked for Hispanic and Asian racial groups—with the latter yielding an 11% increase (from 0.82 using the BISG to 0.91 using the fBISG). In general, the use of fBISG effectively eliminates the performance gap observed between major racial categories when using BISG, which disproportionately affected members of the Asian category.

The source of these improvements in classification accuracy varies by racial category, as indicated by changes in false-positive and false-negative error rates (see columns under "Census last name" in Table 5). Table 5 also contains two additional columns under "All names," which present results based on our second proposed solution to the census data quality issues (we will discuss the results under those last two columns in the "Correcting the missing race-name data problem" section). Among white voters, fBISG—with census last names only—substantially reduces the false-positive rate from 25 to 18% while keeping the false-negative rate below 10%. Among Black voters, the improvement comes primarily from reducing false negatives, bringing type II error down to about 21% from the 27% achieved by BISG.

In turn, while error reduction among Hispanics is small, accounting for the zeroes in the census counts substantially affects the accuracy of classification among Asian voters. For the latter, fBISG substantially reduces the false-negative rate (from almost 50% to about 40%) while keeping the false-positive rate effectively constant. Given the large percentage (20%) of Asian voters living in a block for which the 2010 Census tallies register zero Asians, the improvement induced by fBISG is expected. Overall, classification errors are reduced by about four percentage points—an almost 22% decrease in classification errors across all racial categories.

Although fBISG substantially improves classification accuracy, this gain does not come at the cost of the calibration of predicted probabilities. As we discussed earlier, a model with well-calibrated predictions generates probabilities that match observed proportions of cases, so that about x% of observations are predicted to have the same x% chance of being in a given racial category. On a plot of the latter against the former, well-calibrated probabilities are thus evidenced by a curve that falls close to the 45° line.

Figure 2 shows the benchmark of perfect calibration (black 45° line), as well as curves that track predicted probabilities versus observed sample proportions of voters in each racial category, for both BISG (solid blue curve) and fBISG (dashed red curve) methods. The figure shows that fBISG produces better calibrated predictions (i.e., curves that lie closer to the 45° line) than the standard BISG methodology for voters in minority racial categories (especially Hispanics) while maintaining the high calibration levels of BISG among white and Black voters. In sum, correcting the zero-count measurement error issues can yield substantial improvements in race classification without reducing a high degree of calibration across all major racial categories.

Correcting the missing race-name data problem

Addressing name undercoverage issues and adding additional name information also result in substantial improvements in classification. Figure 3 shows the overall improvement in classification accuracy that results from using additional name-race data from the L2 voter files, for both BISG (blue bars) and fBISG (red bars). While using augmented name dictionaries when fitting the BISG model leaves the zero count issue unaddressed, fBISG models fit with the new name data address both identified problems simultaneously. We fit both models to distinguish gains attained by implementing solutions both separately and in combination. In addition, while these results aggregate across voter files, recall that we mitigate overfitting by sampling each state separately, leaving names from that state out of the augmented dictionaries. Figure S1 presents the results separately for each state that are similar to the overall results.

Consistent with prior findings [e.g., (*13*)], we find that using first and middle name information typically improves the predictive performance of models. Moving from top to bottom, panels in Fig. 3 show the steady improvement in model performance as the predictions rely on an increasing amount of name information (i.e., surnames only; surnames and first names; and surnames, first names, and middle names together). While both BISG and fBISG benefit from the progressively larger name sets being used in the prediction, fBISG (red bars) is able to make the most of the additional information. This is true even among white voters, for whom classification accuracy can be improved by as much as 4.4% (from and AUROC of 0.91 to 0.95 using fBISG for generating predictions).

Once all of our proposed solutions are implemented, improvements in predictive accuracy over the standard BISG methodology are substantial. Across major racial categories, the average increase in AUROC is about 7%, with improvements among Asian voters being as large as 15%—from 0.82 using the standard, census

Table 5. Overall classification error rate as well as false-positive (i.e., type I) and false-negative (i.e., type II) error rates for white, Black, Latino, Asian, and other voters using predictions from standard BISG and from our proposed fBISG model. Each voter is classified to the racial category with the highest posterior probability. For all but the Other category, both types of errors are reduced as we move from standard, census-dictionary BISG to fBISG using the augmented name dictionary.

		Census last name		All names	
Ethnicity	**Error**	**BISG**	**fBISG**	**BISG**	**fBISG**
Overall error rate		16.70%	13.15%	13.20%	11.98%
White	False negative	8.96%	6.59%	8.71%	6.79%
	False positive	25.53%	18.55%	23.00%	15.65%
Black	False negative	27.48%	20.88%	22.70%	17.77%
	False positive	6.41%	4.28%	8.06%	4.28%
Hispanic	False negative	19.43%	16.55%	24.06%	11.76%
	False positive	2.22%	2.03%	2.15%	2.11%
Asian	False negative	49.87%	39.87%	41.44%	30.25%
	False positive	0.44%	0.45%	0.46%	0.49%
Other	False negative	95.18%	91.55%	95.33%	91.78%
	False positive	0.22%	0.97%	0.18%	0.72%

surname-only BISG to 0.94 using all our proposed solutions. Using our fully specified model renders classification quality across major racial categories effectively identical in terms of AUROC, bringing it over 0.94 for all major racial groups.

These gains in classification accuracy, after implementing all of our proposed solutions, are mainly due to substantial reductions in the number of false negatives among all voters—and particularly among non-white voters—as can be seen by comparing the third and sixth columns of Table 5. For non-white voters, type II error is reduced, on average, by about 30% once all our solutions are implemented. These improvements primarily come from correcting false positives attributed to the white category, where we see a corresponding type I error rate reduction of almost 39%. Overall, classification errors are reduced from 16.7% incorrect classifications to about 12% incorrect classifications. That is, among registered voters in the six states for which we have self-reported race data, about 1.7 million more people are correctly classified into their self-reported racial groups.

Moreover, gains in accuracy from adding information on first and middle names are not achieved at the expense of calibration, with calibration curves that are effectively the same across most rows of Fig. 4 for white and Black voters. While the inclusion of all names among Hispanic and Asian voters somewhat worsens model calibration, predictions based on both BISG and fBISG appear to suffer from this problem.

DISCUSSION

Here, we consider the problem of predicting an individual's race. This task is especially relevant to modern research on racial equity in areas including public health, elections, and finance. The current state-of-the-art approach is BISG, which uses surname and geolocation data to generate a probabilistic prediction for each individual over racial classes. However, as we have shown, BISG predictions can underperform for minority groups because of two consistent challenges: inaccurate census counts and name undercoverage.

To address these challenges, we introduce a fully Bayesian analog called fBISG that addresses the problem of census zero counts. Moreover, we augment our name dictionaries, including additional surnames, as well as first and middle names, sourced from voter files in six Southern states provided by L2 Inc. Together, these methodological improvements yield substantial performance gains in predictive accuracy—as measured by AUROC, as well as false-positive and false-negative error rates—while simultaneously improving the calibration of predictions. Moreover, the gains are most pronounced among Hispanics and Asian Americans, drawing their predictions almost to parity with those for white and Black voters in terms of accuracy. We believe that these improvements, which we discuss in detail in Materials and Methods below, will be useful for practitioners, allowing them to obtain improved individual-level racial predictions and better characterize disparate racial impacts.

The results discussed herein are implemented in the latest version of the wru package (*11*), available for download on the Comprehensive R Archive Network (CRAN). The name dictionaries are also separately available for download from the Harvard Dataverse (*14*). While the latest software provides the option to use exclusively the census surname distribution, we believe that using the augmented dictionaries, and including first and middle names, will yield superior performance in most use cases. The merits of using the voter file data—including increased name coverage and the ability to leverage informative names other than surnames—are discussed in this manuscript.

The sole potential drawback stems from the possibility of regional biases in the estimates of race-name probabilities $P(F_i \mid R_i)$, $P(M_i \mid R_i)$, and $P(S_i \mid R_i)$. Recall that under Assumption 1, we have $G_i \perp\!\!\!\perp S_i \mid R_i$. To include first and middle names, we are also implicitly assuming

$$G_i \perp\!\!\!\perp F_i \mid R_i \text{ and } G_i \perp\!\!\!\perp M_i \mid R_i$$

That is, race-name probabilities are unchanged depending on the geography of analysis. These assumptions will not hold exactly in practice; certain names will be more popular among members of a given ethnic group in some locales and less popular among members of the same ethnic group in other locales.

Nonetheless, we have observed meaningful improvements in aggregate performance across the Southern states when using a leave-one-out approach to building the race-name dictionaries. This is significant, as there is considerable heterogeneity across the Southern states themselves, with Florida having a much larger Hispanic population than the other states; Georgia and Louisiana having

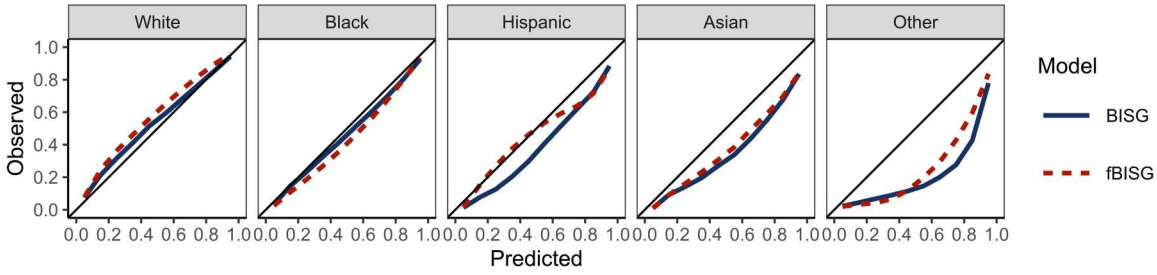

Fig. 2. Calibration curves for race predictions obtained using the standard BISG (blue) and fBISG (red) methods. The curves plot predicted probabilities of being in each racial category against the observed proportion of cases that actually fall in that category. Thus, curves closer to the 45° line indicate better calibrated predictions. The results are based on the 2010 Census surname dictionary. The red curve (i.e., the fBISG calibration curve) is either identical to or closer to the 45° than the blue curve (i.e., the BISG calibration curve), indicating that fBISG's predictive accuracy is at least on par to that of BISG.

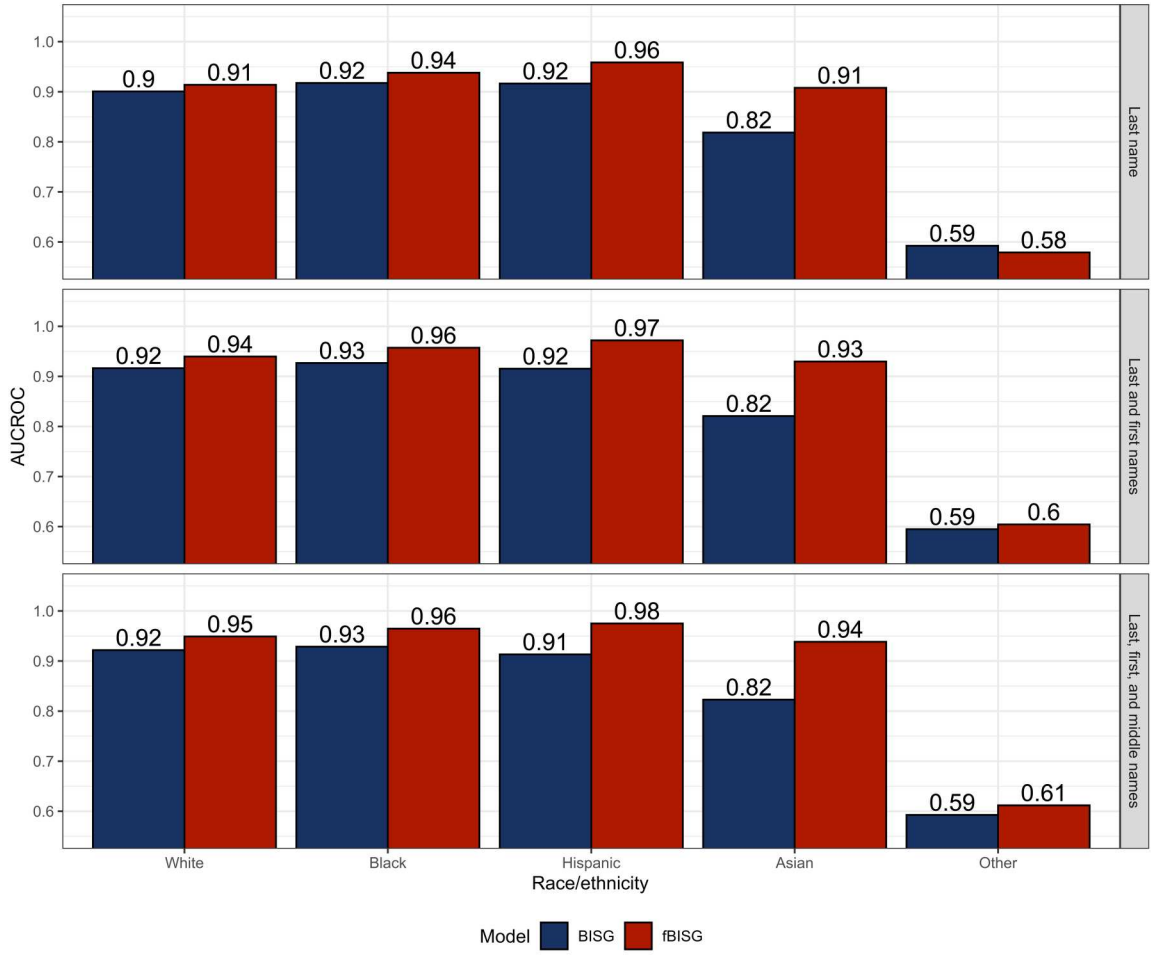

Fig. 3. AUROC for race predictions obtained using the standard BISG (blue) and fBISG (red) methods. The results are based on progressively more name information, starting with the L2-augmented surname dictionary (top). As before, higher values indicate better predictive accuracy. Overall, using more name information uniformly improves the accuracy of models, and using fBISG combined with more name information produces the most accurate models.

disproportionately large Black populations; and Asian residents concentrated in Florida and Georgia. We anticipate similar performance improvements when applying to other regions of the United States.

MATERIALS AND METHODS

In this section, we describe the proposed solutions to the measurement error problems described above. We begin by introducing a measurement error model designed to address potential for error

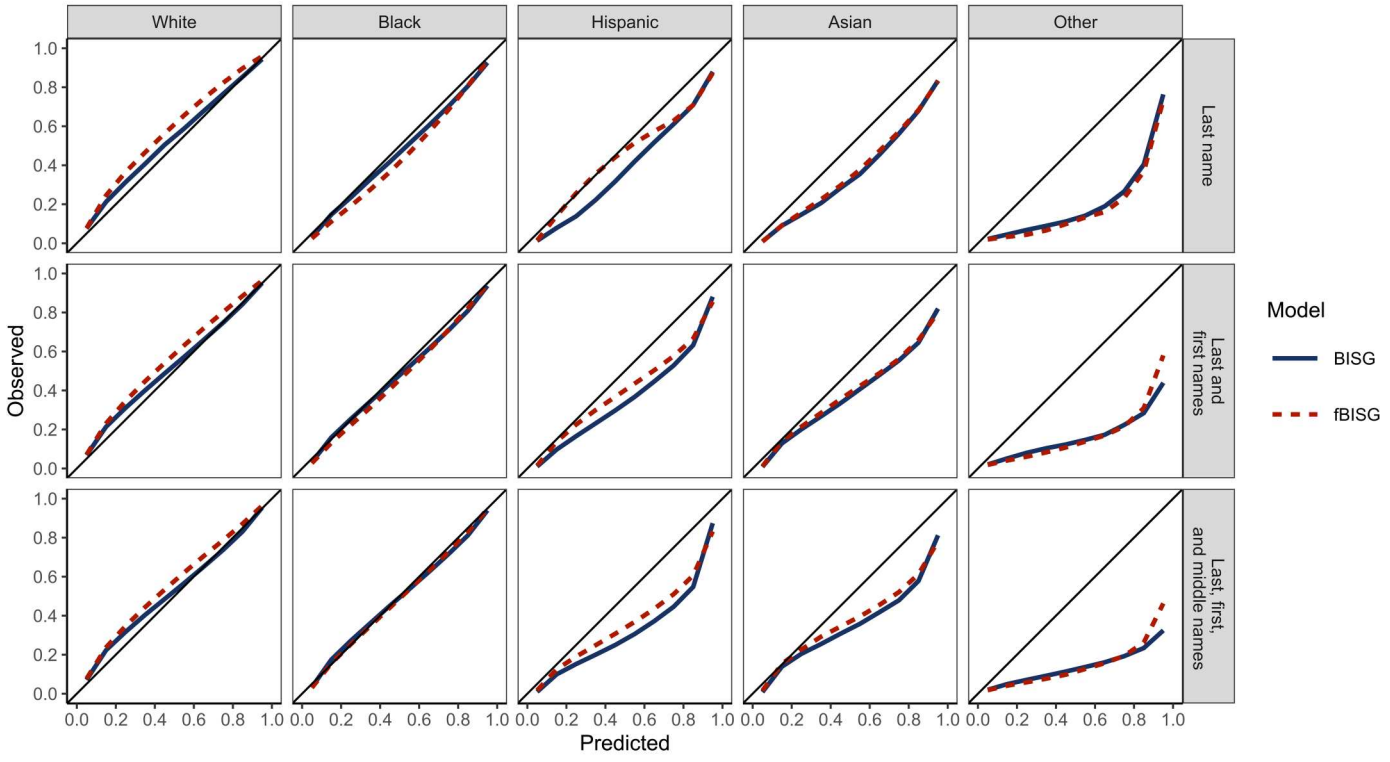

Fig. 4. Calibration curves for race predictions obtained using BISG (blue) and fBISG (red), using progressively more name information from dictionaries augmented with L2 data (from top to bottom rows). The curves plot predicted probabilities of being in each racial category against the observed proportion of cases that actually fall in that category, for each model. As before, curves closer to the 45° line indicate better calibrated predictions.

in census tallies. Our model generalizes the naïve Bayes BISG methodology to a fully Bayesian model. We complete our discussion by describing our name augmentation strategy, designed to correct for lack of coverage in commonly used name-by-race dictionaries.

Accounting for the measurement error in census counts

We use a fully Bayesian modeling strategy to account for potential measurement error that arises when quantifying the racial distribution within each geography. We begin by modeling the observed census counts as a draw from a multinomial distribution with the true, but unknown, race proportions in geolocation g, denoted by $\boldsymbol{\zeta}_g = (\zeta_{1g}, \zeta_{2g}, \ldots, \zeta_{Jg})$

$$\boldsymbol{N}_g \overset{\text{indep.}}{\sim} \text{Multinom}(N_g, \boldsymbol{\zeta}_g) \tag{2}$$

where $\boldsymbol{N}_g = (N_{1g}, N_{2g}, \ldots, N_{Jg})$ is the J-dimensional vector of census counts for individuals who belong to different racial groups and live in geolocation g and $N_g = \sum_{r \in \mathcal{R}} N_{rg}$ is the observed total census population count in geolocation g.

Next, we place the following conjugate prior distribution over the unknown race distribution for the geolocation g

$$\boldsymbol{\zeta}_g \overset{\text{indep.}}{\sim} \text{Dirichlet}(\boldsymbol{\alpha}) \tag{3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_J)$ is the J-dimensional vector of prior hyperparameters. In our implementation, we define a uniform prior distribution with $\boldsymbol{\alpha} = \mathbf{1}$. As it will become clear in Eq. 6 below, this value of α offers enough smoothing over observed zero counts, while injecting a negligible amount of information into the posterior relative to information that comes from the census and the voter file. Furthermore, this prior information does not give preeminence to any particular racial or ethnic group over any another.

We call this measurement error model the fBISG. Letting $P(S_i \mid R_i = r) = \boldsymbol{\pi}_r$, the full posterior distribution of the fBISG is given by

$$\begin{aligned} &P(\{R_i\}_{i=1}^n, \{\zeta_g\}_{g \in \mathcal{G}} \mid \boldsymbol{S}, \boldsymbol{G}, \{\pi_r\}_{r \in \mathcal{R}}, \alpha) \\ &\propto \prod_{i=1}^{n} \prod_{r \in \mathcal{R}} \prod_{g \in \mathcal{G}} \left\{ \left(\prod_{s \in \mathcal{S}} \pi_{sr}^{1\{S_i = s\}} \right) \zeta_{rg}^{1\{G_i = g\}} \right\}^{1\{R_i = r\}} \times \prod_{r \in \mathcal{R}} \prod_{g \in \mathcal{G}} \zeta_{rg}^{N_{rg} + \alpha_r - 1} \\ &= \prod_{s \in \mathcal{S}} \prod_{r \in \mathcal{R}} \pi_{sr}^{m_{sr}} \times \prod_{r \in \mathcal{R}} \prod_{g \in \mathcal{G}} \zeta_{rg}^{n_{rg} + N_{rg} + \alpha_r - 1} \end{aligned} \tag{4}$$

where $n_{rg} = \sum_{i=1}^{n} 1\{R_i = r, G_i = g\}$ is the number of individuals on the voter file who belong to race r and live in geographical unit g, and $n_{rg} = \sum_{i=1}^{n} 1\{R_i = r, G_i = g\}$ is the number of individuals in the voter file who belong to race r and have surname s. The last line of Eq. 4 implies that the posterior distribution over races in fBISG is, once again, a function of two terms: the probability of observing a surname s for a given race r (i.e., π_{sr}) and the (unknown) true racial composition of geolocation g (i.e., ζ_{rg}). As we will detail below, treating this racial composition as unknown and giving it a prior distribution are what allows us to overcome the zero-count

issues that affect BISG, while correctly incorporating information from both the census (viz., the counts N_{rg}) and the voter file under study (viz., the counts n_{rg}) into the prediction problem.

To simplify computation, we integrate out $\boldsymbol{\zeta}_g$, obtaining the following marginalized posterior distribution

$$P(\{R_i\}_{i=1}^n \mid \{\pi_r\}_{r\in\mathcal{R}}, \boldsymbol{S}, \boldsymbol{G}, \alpha) \propto \prod_{r\in\mathcal{R}} \left\{ \prod_{s\in\mathcal{S}} \pi_{sr}^{m_{sr}} \prod_{g\in\mathcal{G}} \Gamma\left(n_{rg} + N_{rg} + \alpha_r\right) \right\} \quad (5)$$

To sample from this joint posterior distribution, we construct a Gibbs sampler. Using the fact that $\Gamma(x + y) = x^y\Gamma(x)$ for $y \in \{0,1\}$, we can derive the following conditional posterior distribution for R_i given the race of the other individuals

$$\Pr\left(R_i = r \mid R_{-i}, S_i = s, G_i = g, G_{-i}, \alpha\right) \propto \pi_{sr}(n_{rg}^{-i} + N_{rg} + \alpha_r) \quad (6)$$

This posterior probability summarizes all the estimation uncertainty about predicting individual race. Note that $n_{rg}^{-i} = \sum_{i'\neq i} 1\{R_{i'} = r, G_{i'} = g\}$ is the only parameter that needs to be updated throughout the sampling process. In particular, we do not need to store the posterior draws of individual race. After the corresponding Markov chain has converged to its stationary distribution, if one wishes to impute each individual's race, then this posterior prediction for R_i can be obtained by iteratively sampling from the full set of conditional distributions in Eq. 6.

Note that the conditional posterior in Eq. 6 factorizes over locations g, which allows us to fit the model separately across any level of geographic aggregation defined on $\mathcal{G}$. While the size of each voter file in our sample did not require parallelization to make computation feasible, factorization over g allows researchers to parallelize their analyses to fit our model efficiently on much larger data files.

The comparison of Eq. 6 with Eq. 1 shows how the fBISG addresses the problems caused by zero census counts. The only difference between these two formulae is that the race-geolocation probability $\Pr(R_i \mid G_i)$ in the BISG prediction formula, which is given by $N_{rg}/\sum_{r'g} N_{r'g}$, is replaced with the ratio $(n_{rg}^{-i} + N_{rg} + \alpha_r)/\sum_{r'}(n_{r'g}^{-i} + N_{r'g} + \alpha_{r'})$ in the fBISG formula. In the BISG methodology, if $N_{rg} = 0$, then the posterior prediction for this racial group r in the geolocation g is zero. In contrast, the fBISG methodology gives nonzero probability of belonging to the racial group with zero census counts by adding a prior α_r and partially pooling other individuals of the same racial group who live in the same geolocation n_{rg}^{-i}. Because both of these additional parameters are nonzero, the fBISG methodology will not give zero probability even in the presence of census zero counts. Note that a larger sample size will typically improve the performance of fBISG because there will be a greater number of individuals who live in the same geography, thereby increasing n_{rg}^{-1} and making partial pooling more effective.

Last, it is straightforward to incorporate additional covariates X_i such as age and sex into the proposed fBISG methodology. Typically, researchers use these covariates in the BISG by assuming the following conditional independence relation (*4*)

$$\{G_i, X_i\} \perp\!\!\!\perp S_i \mid R_i$$

instead of Assumption 1. Thus, the fBISG methodology can incorporate these additional covariates by simply replacing G_i with the random variable jointly defined by G_i and X_i.

Increasing surname coverage and incorporating first and middle names

In addition to the surname, we may also have first and middle names of each individual whose racial group we wish to predict. Let $F_i \in \mathcal{F} = \{1,2, ..., K_F\}$ and $M_i \in \mathcal{M} = \{1,2, ..., K_M\}$ denote the first and middle names of individual i, respectively. Using the same voter file data from L2 Inc., we construct the racial composition of each first name and that of each middle name. This allows us to further approximate the joint distributions $P(R_i, F_i)$ and $P(R_i, M_i)$.

In (*13*), Voicu shows that incorporating the first name can improve the performance of the BISG. The author makes the assumption, similar to Assumption 1, that the first name is independent of geolocation conditional on race. In addition, it is assumed that the first name is independent of the surname given race. If we make the same assumption about middle names, then the prediction formula becomes

$$P(R_i \mid F_i, M_i, S_i, G_i) \propto P(F_i \mid R_i)P(M_i \mid R_i)P(S_i \mid R_i)P(R_i \mid G_i)$$

Combining this information with our fully Bayesian model for smoothing over zero census counts results in the following updated full conditional distribution over individual i's race

$$\Pr\left(R_i = r \mid R_{-i}, F_i = f, M_i = m, S_i = s, G_i = g, G_{-i}, \alpha\right) \propto \pi_{fr}^{\mathcal{F}}\pi_{mr}^{\mathcal{M}}\pi_{sr}^{\mathcal{S}}(n_{rg}^{-i} + N_{rg} + \alpha_r) \quad (7)$$

where $\pi_{fr}^{\mathcal{F}} = P(F_i = f \mid R_i = r)$, and similarly with $\pi_{mr}^{\mathcal{M}}$ and $\pi_{sr}^{\mathcal{S}}$.

We demonstrate the empirical benefit of incorporating voter file surname racial distributions, as well as those of first and middle names. Note that in the companion paper (*15*), we also discuss the relative advantages of using voter file data over the state-of-the-art data source for the racial distribution of first names, given by (*16*). Using the same set of voter files from L2 Inc., we consider matching individuals to name dictionaries under several schemes. First, we consider surnames exclusively and compute the proportion of individuals from each racial group who do not have a surname matched to the census dictionaries (as in Table 3). Next, we compute the proportion of individuals of each race who do not have a surname matched to the census dictionaries, augmented with data from the L2 voter files themselves. Third, we compute the proportion of individuals of each race who do not have a surname matched to the augmented surname dictionary or a first name matched to the separate first name dictionary compiled from the L2 data. Last, we compute the proportion of individuals of each race who do not have a name matched to any of the augmented surname dictionary or to first and middle name dictionaries compiled from the L2 data.

Because the voter file data are used both to compile the dictionaries and to assess coverage, as before, we iteratively hold out each of the six states (Alabama, Florida, Georgia, Louisiana, North Carolina, and South Carolina) and consider coverage using a dictionary compiled from the other five states. The results in Fig. 5 show that dictionary augmentation—and inclusion of additional names—substantially decreases the proportion of individuals who cannot be matched to any dictionary. For non-Asian voters, all

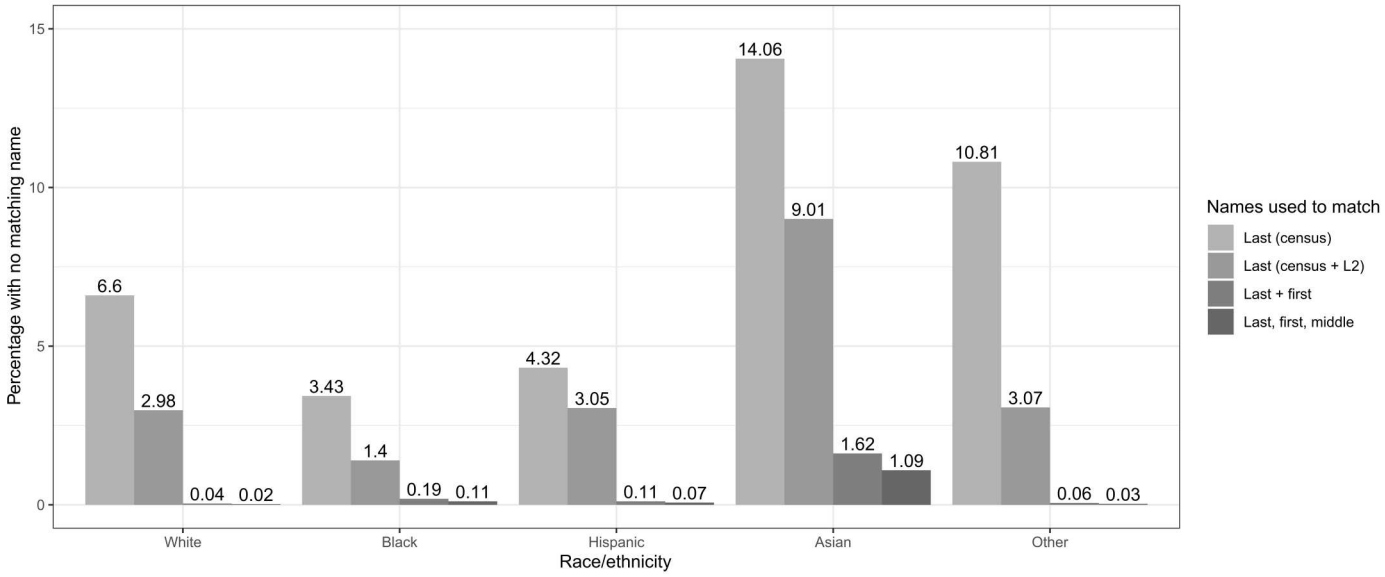

Fig. 5. Percentage of individuals in each racial group who cannot be matched to any name dictionary, under four different matching schemes. The schemes include matching to census last names only; matching to census and L2 last names; matching last names and first names; and matching last, first, and middle names. Data are drawn from the voter files of Alabama, Florida, Georgia, Louisiana, North Carolina, and South Carolina.

but a negligible fraction of voters can be matched to at least one dictionary once all their names are included. Among Asians, approximately 1% of voters still cannot be matched when using first, middle, and last names. This, however, represents a marked improvement relative to the case of exclusively using surnames and sourcing data only from the census.

Supplementary Materials

This PDF file includes:

Fig. S1
Tables S1 and S2

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. M. N. Elliott, A. Fremont, P. A. Morrison, P. Pantoja, N. Lurie, A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health Serv. Res.* **43**, 1772–1736 (2008).
2. M. N. Elliott, P. A. Morrison, A. Fremont, D. F. McCaffrey, P. Pantoja, N. Lurie, Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Serv. Outcomes Res. Methodol.* **9**, 69–83 (2009).
3. K. Fiscella, A. M. Fremont, Use of geocoding and surname analysis to estimate race and ethnicity. *Health Serv. Res.* **41**, 1482–1500 (2006).
4. K. Imai, K. Khanna, Improving ecological inference by predicting individual ethnicity from voter registration records. *Polit. Anal.* **24**, 263–272 (2016).
5. F. Edwards, H. Lee, M. Esposito, Risk of being killed by police use of force in the United States by age, race-ethnicity, and sex. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 16793–16798 (2019).
6. P. Hepburn, R. Louis, M. Desmond, Racial and gender disparities among evicted Americans. *Sociol. Sci.* **7**, 649–662 (2020).
7. D. M. Studdert, Y. Zhang, S. A. Swanson, L. Prince, J. A. Rodden, E. E. Holsinger, M. J. Spittal, G. J. Wintemute, M. Miller, Handgun ownership and suicide in California. *N. Engl. J. Med.* **382**, 2220–2229 (2020).
8. B. L. Fraga, *The Turnout Gap: Race, Ethnicity, and Political Inequality in a Diversifying America* (Cambridge Univ. Press, 2018).
9. C. T. Kenny, S. Kuriwaki, C. McCartan, E. T. R. Rosenman, T. Simko, K. Imai, The use of differential privacy for census data and its impact on redistricting: The case of the 2020 U.S. census. *Sci. Adv.* **7**, eabk3283 (2021).
10. J. Comenetz, Frequently occuring surnames in the 2010 census, in *Technical Report* (United States Census Bureau, 2016); www2.census.gov/topics/genealogy/2010surnames/surnames.pdf.
11. K. Khanna, B. Bertelsen, E. Rosenman, S. Olivella, K. Imai, *wru: Who Are You? Bayesian Prediction of Racial Category Using Surname and Geolocation* (2022). R package version 1.0.0. available at https://CRAN.R-project.org/package=wru.
12. G. Basso, G. Peri, Internal mobility: The greater responsiveness of foreign-born to economic conditions. *J. Econ. Perspect.* **34**, 77–98 (2020).
13. I. Voicu, Using first name information to improve race and ethnicity classification. *Stat. Public Policy* **5**, 1–13 (2018).
14. E. Rosenman, S. Olivella, K. Imai, *Name Dictionaries for "wru" R Package* (Harvard Dataverse, V1, 2022); https://doi.org/10.7910/DVN/7TRYAC.
15. E. T. R. Rosenman, S. Olivella, K. Imai, Race and ethnicity data for first, middle, and last names. arXiv:2208.12443 [stat.OT] (26 August 2022).
16. K. Tzioumis, Demographic aspects of first names. *Sci. Data* **5**, 180025 (2018).

Acknowledgments: We thank B. Willsie, chief executive officer of L2 Inc., for providing us with the voter files we use in this paper. We also thank D. Ho for insightful comments on this manuscript. Last, we thank B. Bertelsen for help with updating the wru package. **Funding:** We have no funding to declare. **Author contributions:** K.I. conceived the project and supervised the research. K.I. and S.O. derived the fBISG computation algorithm, and S.O. implemented the algorithm. E.T.R.R. compiled the data dictionaries. E.T.R.R. and S.O. implemented the data analyses. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Code is available in the wru package, version 1.0.0, available at https://CRAN.R-project.org/package=wru. Name data are available at https://doi.org/10.7910/DVN/7TRYAC. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 14 May 2022
Accepted 31 October 2022
Published 9 December 2022
10.1126/sciadv.adc9824

ScienceAdvances

Addressing census data problems in race imputation via fully Bayesian Improved Surname Geocoding and name supplements

Kosuke ImaiSantiago OlivellaEvan T. R. Rosenman

Sci. Adv., 8 (49), eadc9824. • DOI: 10.1126/sciadv.adc9824

Science Advances (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.