



## Using First Name Information to Improve Race and Ethnicity Classification

Ioan Voicu

**To cite this article:** 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](https://doi.org/10.1080/2330443X.2018.1427012)

**To link to this article:** <https://doi.org/10.1080/2330443X.2018.1427012>



© 2018 The Author. Published with License by American Statistical Association© Ioan Voicu



[View supplementary material](#)



Published online: 21 Mar 2018.



[Submit your article to this journal](#)



Article views: 12055



[View related articles](#)



[View Crossmark data](#)



Citing articles: 17 [View citing articles](#)

# Using First Name Information to Improve Race and Ethnicity Classification

Ioan Voicu

U.S. Department of the Treasury, Office of the Comptroller of the Currency (OCC), Washington, DC

## ABSTRACT

This article uses a recent first name list to develop an improvement to an existing Bayesian classifier, namely the Bayesian Improved Surname Geocoding (BISG) method, which combines surname and geography information to impute missing race/ethnicity. The new Bayesian Improved First Name Surname Geocoding (BIFSG) method is validated using a large sample of mortgage applicants who self-report their race/ethnicity. BIFSG outperforms BISG, in terms of accuracy and coverage, for all major racial/ethnic categories. Although the overall magnitude of improvement is somewhat small, the largest improvements occur for non-Hispanic Blacks, a group for which the BISG performance is weakest. When estimating the race/ethnicity effects on mortgage pricing and underwriting decisions with regression models, estimation biases from both BIFSG and BISG are very small, with BIFSG generally having smaller biases, and the maximum a posteriori classifier resulting in smaller biases than through use of estimated probabilities. Robustness checks using voter registration data confirm BIFSG's improved performance vis-a-vis BISG and illustrate BIFSG's applicability to areas other than mortgage lending. Finally, I demonstrate an application of the BIFSG to the imputation of missing race/ethnicity in the Home Mortgage Disclosure Act data, and in the process, offer novel evidence that the incidence of missing race/ethnicity information is correlated with race/ethnicity.

## ARTICLE HISTORY

Received June 2016  
Accepted January 2018

## KEYWORDS

Bayesian methods;  
Classification; Ethnicity;  
Measurement errors; Race

## 1. Introduction



The ability to accurately classify individuals into racial or ethnic groups plays a crucial role in studying racial and ethnic disparities in a wide range of areas, including but not limited to: health care, access to financial services and labor markets, educational outcomes, socio-economic status, and political science. Yet this ability is hampered by the existence of significant gaps in the collection of accurate racial and ethnic data at the population level, largely due to the absence of a mandate for collecting such information, and personal identification information (PII) concerns. For example, until recently many viewed the collection of race/ethnicity data from the users of the U.S. health care system as illegal (Fremont and Lurie 2004). Additionally, in the financial services area, lenders are not required to collect information on the race/ethnicity of applicants for nonmortgage products. Even when lenders are required to collect such information for mortgage applications, under the Home Mortgage Disclosure Act (HMDA) of 1975, a nontrivial portion of these applications are missing this information, primarily because the applicants declined to provide it in mail, Internet, or telephone applications.

Absent direct information on race/ethnicity, practitioners and researchers have turned to methods of estimating these demographics indirectly, based on other responses such as name and address, which are readily available from various

sources (e.g., loan applications, medical records). Such indirect methods also require publicly available data that help determine how the relevant responses are associated with a specific race/ethnicity. Several indirect methods for estimating race/ethnicity have been proposed, some based on surname information, some on geographic location, and others on a combination of surname and geographic location or surname and first name.

Surname-based methods typically infer race/ethnicity by matching the relevant surnames with well-established dictionaries of Hispanic or Asian surnames (e.g., Perkins 1993; Lauderdale and Kestenbaum 2000) or with the comprehensive list of surnames and their associated race/ethnicity prevalences<sup>1</sup> compiled by the U.S. Census Bureau in 2007 based on the Decennial Census 2000. On the other hand, methods based on geographic location, also known as geocoding methods, use an individual's address to link individuals to census demographics of the geographic areas where they live. The race/ethnicity prevalences for the geographic area of residence are then used by themselves, as the probabilities of an individual belonging to each of the identified racial/ethnic groups, or in conjunction with a threshold to classify the individual in a specific group.

Both surname-based and geocoding methods have well known limitations: the former has limited ability to distinguish non-Hispanic (NH) Blacks from NH Whites because their

**CONTACT** Ioan Voicu  [ioan.voicu@occ.treas.gov](mailto:ioan.voicu@occ.treas.gov)  U.S. Department of the Treasury, Office of the Comptroller of the Currency (OCC), 400 7th Street SW, Washington, DC 20219.

<sup>1</sup>The term "prevalence" refers to the percentage of people with a given surname that belong to a specific racial/ethnic group.

© 2018 Ioan Voicu

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The moral rights of the named author(s) have been asserted.

Published with License by American Statistical Association

surnames are relatively nondistinctive, and the latter has little ability to identify Hispanics or NH Asians because these groups are not spatially segregated.<sup>2</sup> For this reason, hybrid approaches have been suggested that attempt to improve the accuracy of race/ethnicity estimates by combining the different strands of information. To combine said information, researchers have typically employed either arbitrary mathematical functions, such as the multiplicative, linear, and maximum functions in Coldman, Braun, and Gallagher (1988), or naive Bayesian models, such as the Bayesian Improved Surname Geocoding (BISG) algorithm proposed by Elliott et al. (2009). The latter is the most recent and advanced hybrid approach. It first creates a “prior” probability of belonging to a given racial/ethnic category using the Census 2000 surname list, and then applies the naive form of Bayes’ theorem to update this probability using the demographic characteristics of the geographic area (census block group in this case) where the individual resides. Several studies (Elliott et al. 2009; Consumer Financial Protection Bureau [CFPB] 2014; Baines and Courchane 2014; Adjaye-Gbewonyo et al. 2014) performed evaluations of BISG using health plan enrollment data and mortgage data, and found that it predicts race/ethnicity, especially NH Black and NH White, more accurately than the surname-only and geography-only approaches.<sup>3</sup>

First name information has been under used in prior research on race/ethnicity proxies, largely due to the lack of comprehensive tabulations covering a wide range of first names and their associated race/ethnicity prevalences.<sup>4</sup> Tzioumis (2017) has recently filled this gap by compiling a comprehensive list of 4,250 first names drawn from mortgage applications and classified by self-reported race/ethnicity, henceforth the “first name list.” In this article, I take advantage of this first name list, in order to improve on the existing hybrid approaches. Specifically, I use this information to enhance the naive Bayes classifier proposed by Elliott et al. (2009) by creating a new algorithm, which I refer to as the Bayesian Improved First Name Surname Geocoding algorithm (BIFSG) that incorporates first name information into the BISG framework. In essence, BIFSG is an extension of the BISG formula from two to three conditional attributes, that is, a naive Bayesian updating formula that updates the surname-based probabilities of membership in one of six racial/ethnic categories with the first name and geographic location proportions for these same six groups. Given that first name demographic information could be helpful in predicting race/ethnicity as documented in Tzioumis (2017), its addition to the Bayesian updating framework has the potential to improve the classification accuracy both directly, by adding to the information content associated with surnames and geography, and indirectly, by improving the imputation of missing surname and geography information. In developing the new algorithm, I also

take advantage of an updated surname list released by the U.S. Census Bureau in 2017 based on the Decennial Census 2010, referred to hereafter as the “Census 2010 surname list.”

To evaluate the performance of BIFSG relative to BISG, I use a wide array of metrics that compare the accuracy of the two algorithms in terms of how closely the race/ethnicity estimates that they produce match the self-reported race/ethnicity for the same individuals. The evaluation is based on a large mortgage lending dataset composed of information from several lenders. As a robustness check, I also evaluate BIFSG using voter registration data for North Carolina.

Following evaluation, as an empirical application, I apply BIFSG to the imputation of missing race/ethnicity information in the HMDA data, and in the process, offer novel evidence that race/ethnicity are somewhat correlated with the incidence of missing race/ethnicity information in mail/Internet/telephone applications.

To preview the main results based on the mortgage data, I find that the new approach offers improvements, in terms of accuracy and coverage, over BISG for all major racial/ethnic categories. The improvements are most substantial for the group for which BISG is least accurate—NH Blacks. Moreover, the improvements provided by the BIFSG could become even stronger if a new and improved first name list became publicly available. Additionally, I find that when estimating the effects of race/ethnicity on mortgage pricing and underwriting decisions with reasonably well-specified regression models, the estimates based on either BIFSG or BISG proxies are remarkably close to those based on actual race/ethnicity. This finding should significantly alleviate bias-related concerns raised in Baines and Courchane (2014). The evaluation also shows that discrete classification schemes based on either the BIFSG or BISG probabilities may result in smaller biases in the estimation of the race/ethnicity effects on mortgage lending outcomes than probabilities themselves. Furthermore, evaluation results based on voter registration data for North Carolina show significant improvements in accuracy and coverage from BIFSG, suggesting that the new algorithm is also applicable to fields other than mortgage lending. Finally, the results from the BIFSG empirical application offer evidence that the incidence of missing race/ethnicity information in the mortgage data is generally higher among NH Blacks than among other groups.

The remainder of the article is organized as follows. Section 2 offers a detailed description of BIFSG, including the underlying data, relevant computations, and approaches to evaluate its accuracy in comparison with BISG. Section 3 discusses the evaluation results, and Section 4 presents the empirical application of the BIFSG to the imputation of missing race/ethnicity information in the HMDA data. Section 5 concludes the article.

## 2. Method

The new method to estimate race/ethnicity uses three publicly available data sources combined with naive Bayesian methods. A fourth source including proprietary mortgage data is used to illustrate computations and validate the approach. In this section I describe the data, then the Bayesian algorithm, and finally the

<sup>2</sup> For example, see Fiscella and Freemont (2006) and Elliott et al. (2009) for a detailed description of the advantages and drawbacks of each approach.

<sup>3</sup> Another study, Martino et al. (2013), applied BISG to evaluate health plan performance by race/ethnicity.

<sup>4</sup> A number of studies, primarily for non-U.S. populations (e.g., U.K., Canada, Germany, Netherlands), have used first name information to create race/ethnicity proxies. However, the focus of these studies has been limited to a very narrow range of ethnic minorities, usually of Asian origin (Mateos 2007). Additionally, Mateos (2007) developed a method to classify the U.K. population into groups of common national origin based on surnames and first names; however, to my knowledge, this method has not yet been validated on external, individual-level data.

methods to evaluate the accuracy of the new approach compared with BISG.

## 2.1. Input Data

### 2.1.1. Surname Data

For surname demographic information, I use the Census 2010 surname list.<sup>5</sup> This list includes all surnames occurring 100 or more times in the Decennial Census 2010 (more than 160,000 surnames, covering about 90% of the U.S. population), along with demographic information showing the percentage of people with a given surname that belong to one of six groups: Hispanic, NH Black, NH White, NH Asian/Pacific Islander (API), NH American Indian/Alaskan Native (AIAN), and NH Multiracial.<sup>6</sup>

### 2.1.2. First Name Data

I use the list of first names in Tzioumis (2017). This list draws information from a large pool of recent mortgage application data, and includes 4250 first names with information on their respective counts and proportions across six mutually exclusive and collectively exhaustive racial/ethnic categories that are consistent with the categories used in the Census 2010 surname list. For the great majority of cases, the first names' demographic information is calculated using more than 30 observations.<sup>7</sup> The coverage of this first name list is very similar to that of the Census 2010 surname list. As noted by Tzioumis (2017), since this list is not based on census data for the entire U.S. population, but on mortgage applications, its demographic information becomes more representative when the sample for which one wishes to create the proxies has characteristics similar to those of mortgage applicants (e.g., adult population, employed population).

### 2.1.3. Geo-Demographic Data

I use census-block-group level data on counts and proportions across the six aforementioned racial/ethnic groups, derived from the Decennial Census 2010 SF1 dataset. I choose geocoding at the census-block-group level partly to enhance comparison with other closely related studies using the same geography level, and partly because the census block group is smaller than other commonly used geographies (census tract, zip code) and the degree of correspondence between area and individual characteristics generally increases when smaller, more homogenous units of analysis are used (Krieger et al. 2002).<sup>8</sup>

**Table 1.** Match rates for names and geography.

Match rate for first names	0.884
Match rate for surnames	0.886
Match rate for geography <sup>2</sup>	0.957
Match rate for matches on both first name and surname	0.795
Match rate for matches on both surname and geography	0.848
Match rate for matches on all three features	0.761
<i>N obs</i> <sup>1</sup>	279,404

NOTE. <sup>1</sup>Includes all applications with valid race/ethnicity from single applicants and dual applicants with same race/ethnicity.

<sup>2</sup>Applications that were matched with a geographical unit without any people are considered failed matches and thus are excluded from the numerator of the match rate.

## 2.2. Validation Data

In order to demonstrate the calculation of BIFSG proxies and assess their accuracy, I extract information from several distinct proprietary databases of mortgage transactions, each belonging to a different lender and a given year in the 2012–2014 interval. The extracted information includes applicant first and last names, geographic location,<sup>9</sup> self-reported race/ethnicity, action taken on the application (e.g., denied, originated, approved but not accepted), and the cost of borrowing as measured by the annual percentage rate (APR).<sup>10</sup> I combine these datasets, partly for data confidentiality purposes and partly to improve the representativeness of the test population. The combined dataset contains 279,404 observations after excluding applications with invalid race/ethnicity information.<sup>11</sup> It is worth noting that the requirements for collection and reporting of information on race/ethnicity are consistent across the original datasets since the respective lenders comply with HMDA requirements. Additionally, HMDA race/ethnicity classifications allow exact replication of the six groups from the Census 2010 surname list and the first name list.

## 2.3. Implementation of the BIFSG Algorithm

Constructing the BIFSG proxies for race/ethnicity requires several steps. These steps involve cleaning the validation mortgage data to merge with the two name lists (surnames and first names) and the census geo-demographic data; then constructing the probabilities that are inputs in the BIFSG formula; and, finally, applying the BIFSG formula to compute the proxies.

As shown in Table 1, the respective match rates for the three relevant attributes are quite high—about 88% for first names and surnames, and 96% for geography (i.e., census block group). As Tzioumis (2017) also mentions, the overlap between the missing/unmatched first names and the missing/unmatched surnames is very small (no overlap corresponds to a match rate for matches on both first name and surname of 0.770, which is very close to the actual match rate of 0.795).

<sup>9</sup> The data maintain the confidentiality of the identity of any individual person, family, households, or properties, either directly or indirectly.

<sup>10</sup> For dual applications of most lenders, while the mortgage data includes race/ethnicity for both applicant and co-applicant, only the applicant's name and geographic location are available.

<sup>11</sup> Mortgage applications with missing values for race/ethnicity and dual applications for which race/ethnicity of the applicant and co-applicant are not identical are excluded. The reason for keeping only dual applications with the same race/ethnicity is because, for most lenders, only one name is available per application and, although that name can reasonably be assumed to belong to the applicant, it is possible that this is not always the case.

<sup>5</sup> This list is publicly available at [https://www.census.gov/topics/population/genealogy/data/2010\\_surnames.html](https://www.census.gov/topics/population/genealogy/data/2010_surnames.html).

<sup>6</sup> For confidentiality purposes, the Census Bureau has suppressed exact counts for racial/ethnic categories with fewer than five occurrences for a given surname; and when only a single category had fewer than five occurrences for a given surname, both its count and the count of the category with the second fewest occurrences were suppressed. Similar to CFPB (2014) and Elliott et al. (2009), in these cases, I distribute the sum of the suppressed counts for each surname equally across all groups with missing nonzero counts.

<sup>7</sup> The only exception is when the proportion is unity for a single category and zero for all other five categories, and it is based on 15–29 observations. This exception is intended to capture strictly ethnic first names that appear infrequently in the U.S. population (e.g., Eleftherios, Slobodan, Tomislav, Xiaoping).

<sup>8</sup> Fiscella and Fremont (2006) provided a good comparison of the various geographic area definitions typically used in geocoding approaches.



Following the initial data cleaning and matching steps, the probability inputs to the BIFSG formula are computed. There are three sets of probabilities—surname-based, first-name-based, and geography-based. I describe each of them below.

- (a) *Surname-based probabilities.* For each surname in the validation data that matches the Census 2010 surname list, the probability that a person belongs to a given racial/ethnic group given the person's surname is approximated by the proportion of all people with the given surname who report being of the given race/ethnicity, which is readily available in the Census 2010 surname list. For applications with missing surnames or surnames that cannot be matched with the census list, the surname-based probabilities are considered missing.
- (b) *First-name-based probabilities.* For each first name in the validation data that matches the first name list, the probability of a person's having that first name, given the person's race/ethnicity, is approximated by the proportion of the population of the given race/ethnicity who bear the respective first name. This proportion is derived from the full sample used by Tzioumis (2017) to develop the first name list.<sup>12</sup> For applications with missing first names or first names that cannot be matched with the first name list, the first-name-based probabilities are considered missing.
- (c) *Geography-based probabilities.* For each application with nonmissing geo-demographic information, I compute the proportion of the U.S. adult population for each race/ethnicity residing in the census block group that pertains to that application (e.g., the proportion of all Hispanics in the United States that reside in the census block group associated with the given application). For applications with missing geo-demographic information, the geography-based probabilities are considered missing.

The BIFSG algorithm is then built based on a naive Bayesian updating formula that updates the prior probability of membership in each racial/ethnic category as defined by the surname-based probabilities with the first-name-based and geography-based probabilities, respectively. This formula, which is an extension of the BISG formula from two conditional attributes to three, calculates the probability that a person with surname ( $s$ ), first name ( $f$ ), and geographic area of residence ( $g$ ) belongs to racial or ethnic group ( $r$ ) as follows:

$$p(r|s, f, g) = \frac{p(r|s) \cdot p(f|r) \cdot p(g|r)}{\sum_{r=1}^6 p(r|s) \cdot p(f|r) \cdot p(g|r)}, \quad (1)$$

where:  $p(r|s, f, g)$  is the updated (posterior) probability of being of race/ethnicity  $r$ , given surname  $s$ , first name  $f$ , and geographic

area  $g$ ;  $p(r|s)$  is the probability that a person is of race/ethnicity  $r$ , given that the person has surname  $s$ , (i.e., the surname-based probability described above);  $p(f|r)$  is the probability that a person has first name  $f$ , given that the person is of race/ethnicity  $r$  (i.e., the aforementioned first-name-based probability);  $p(g|r)$  is the probability that a person resides in geographic area  $g$ , given that the person is of race/ethnicity  $r$  (i.e., the aforementioned geography-based probability); and the summation in the denominator occurs over the six race/ethnicity categories defined previously.<sup>13</sup> Details on the derivation of this formula are provided in Appendix A.

The statistical validity of the BIFSG formula relies on the assumption of conditional independence among surnames, first names, and geographic areas, that is,  $p(g|r, s) = p(g|r)$  and  $p(f|r, s, g) = p(f|r)$ . In other words, this assumption implies that the probability of residing in a given geographic area, given a person's race/ethnicity, does not vary by surname, and that the probability of having a given first name, given a person's race/ethnicity, does not vary by surname or geographic area. These assumptions are likely to be strong and cannot be tested or relaxed with the available data, due to very small sample sizes for the various combinations of name, geography, and race/ethnicity that are required for testing. However, Domingos and Pazzani (1997) showed that the naive Bayesian classifier performs quite well in practice in terms of classification accuracy—though not necessarily in terms of the accuracy of estimated probabilities—even when strong attribute dependences are present. This feature is at least in part due to the fact that the classifier does not depend on attribute independence to be optimal for classification purposes.

The BIFSG formula requires that all input probabilities are nonmissing. However, to the extent that this requirement results in significant data attrition, it may be desirable to enhance the algorithm so that it also creates proxies when one or more of the input probabilities are missing. This can be easily accomplished by computing proxies using a BISG-like formula if two attributes have nonmissing values, and using surname-only (SO), first-name-only (FO), or geography-only (GO) probabilities if a single attribute has nonmissing values.<sup>14,15</sup> For brevity, I do not cover the performance of the various algorithm extensions or related weighting schemes in this article. However, it is worth noticing that if BIFSG is more accurate than BISG when all

<sup>13</sup> The Bayesian updating formula can be expressed using different permutations of the three attributes, for example,  $p(r|f) \cdot p(s|r) \cdot p(g|r)$ , depending on the choice of how to define an individual's "prior" probability of membership in a given racial/ethnic category. In preliminary work, I experimented with two alternative permutations, using the racial/ethnic distribution of the individual's first name,  $p(r|f)$ , and geographic location,  $p(r|g)$ , respectively, to define the "prior," and the performance was very similar across alternatives, with the one presented in the paper resulting in marginally better overall performance.

<sup>14</sup> The SO probability is computed like the surname-based probability described above; the FO probability is given by the proportion of all people with a specific first name who report being of a given race/ethnicity, which is readily available in the first name list; and the GO probability is given by the proportion of all people in a given geographic area who report being of a given race/ethnicity.

<sup>15</sup> To the extent that methods that use fewer attributes produce less accurate proxies, it may be worthwhile for future research to explore the use, in statistical analyses, of some penalty function for observations that are assigned proxies based on one or two attributes.

<sup>12</sup> Specifically, the denominator of the proportion is derived from the size of the full sample used by Tzioumis (2017), including all mortgage applicants with valid first names regardless of how common the name is—not just the subset of applicants with the most common first names (which have at least 30 observations) that make up the first name list. The denominators used to compute the relevant proportions are available upon request from the author.

attributes are nonmissing, the BIFSG will also maintain its lead when “extended” with BISG if only the first name is missing.<sup>16,17</sup>

## 2.4. Outcomes of the BIFSG Algorithm: Probabilities vs. Classifications

In the statistical analysis of variation in mortgage outcomes by race/ethnicity, BIFSG probabilities can be used to proxy for race/ethnicity either *directly* (i.e., using the predicted probabilities *per se*) or *indirectly* (i.e., creating a binary measure based on a threshold rule). Each of these approaches has its own strengths and weaknesses. For example, McCaffrey and Elliott (2008) showed that the loss of efficiency from modeling with discrete classifications is larger than that from modeling with continuous probabilities, although this result is demonstrated in the case of dichotomous variables and the authors acknowledge that the relative efficiency of the methods in the case of polytomous variables (like race/ethnicity) is an important area of further research. For another example, the use of probabilities for classification purposes may result in biased estimates because of classification errors, and so does modeling with probabilities if the probabilities are biased. The latter is a distinct possibility in the context of the naive Bayes updating, where, as discussed above, the accuracy of estimated probabilities may be negatively impacted if the attribute independence assumption is violated (Domingos and Pazzani 1997).

Statistical considerations aside, discrete classifications have the advantage of allowing identification of individuals of a specific race/ethnicity, which can be very valuable in certain areas. For example, restitution schemes used to settle discrimination lawsuits would be difficult to accomplish without discrete classification.

## 2.5. Evaluation of the BIFSG Algorithm Using Mortgage Data

Elliott et al. (2008, 2009), CFPB (2014), and Baines and Courchane (2014) have already shown—using health plan enrollment data or mortgage data for which race/ethnicity are self-reported—that the BISG proxy method is more accurate than either the surname-only or geography-only methods. Therefore, my evaluation of the BIFSG algorithm focuses on its performance relative to BISG. Similar to the approach of CFPB (2014) and Baines and Courchane (2014), I test the accuracy of these two methods using the aforementioned mortgage data reported under HMDA.

The evaluation employs several metrics that compare the accuracy of the two proxy algorithms in terms of how closely the race/ethnicity estimates that they produce match the self-reported race/ethnicity for the same individuals. Given that

proxies can be defined either using the BIFSG and BISG probabilities directly or using discrete classifications based on these probabilities, evaluation covers both these approaches.

Following previous research on the BISG performance (Elliott et al. 2008 and 2009; CFPB 2014; and Baines and Courchane 2014), I assess the accuracy of the proxies in three ways: (1) by comparing the distribution of race/ethnicity across applicants based on the proxies to the distribution based on the self-reported attributes; (2) by assessing how well the proxies are able to sort applicants into the self-reported race/ethnicity classes; and (3) by evaluating the biases that proxies may cause in estimating the effects of race/ethnicity on mortgage lending outcomes, using mortgage pricing and underwriting decisions as examples of outcomes. As noted by Elliott et al. (2008), methods (1) and (2) are complementary in that the first detects systematic classification errors and the second finds unsystematic errors. Method (3) evaluates the impact that these errors may have on the specific outcomes that are analyzed on the basis of proxies.

Method (2) employs two sets of metrics: the Pearson correlation coefficient between the proxy probability and the self-reported race/ethnicity, which measures the extent to which applicants of a given race/ethnicity are assigned higher proxy probabilities of belonging to that race/ethnicity; and a set of accuracy metrics—false negative rate and false positive rate—which together measure the diagnostic power of discrete classification schemes based on the proxy probabilities.<sup>18</sup> Related to the accuracy metrics, false negatives in a particular group are individuals who self-report belonging to that group, but whom the proxy method categorizes into another group; and false positives in a given group are individuals whom the proxy method assigns to that group, when in fact they belong to another group. Then, the corresponding rates for a particular group are calculated as follows: the false negative rate is the ratio of the number of false negatives in that group to the total population that self-reports belonging to that group; and the false positive rate is the ratio of the number of false positives categorized in that group to the total population that the proxy classifies in that group.<sup>19</sup> Individuals whom the proxy method cannot assign to any particular group are not included in the above calculations; however, the proportion of unassigned cases within a given group is separately calculated as a measure of the coverage of the proxy method. In addition to calculating the aforementioned metrics, I provide an explanation in Appendix B of how the addition of first name information helps improve the sorting quality of the proxy probabilities.

<sup>16</sup> Of course, the difference in accuracy between BISG and the BIFSG “extended” with BISG narrows as the missing rate for first names increases, since the extended BIFSG becomes more similar to BISG as more observations have race/ethnicity imputed based on BISG.

<sup>17</sup> In preliminary research, using a smaller dataset, I also found that a “fully extended” BIFSG (i.e., with proxies computed for all observations with at least one nonmissing attribute) is more accurate than a similarly extended BISG. Selected results from this preliminary research are available upon request from the author.

<sup>18</sup> In preliminary research, I also calculated the area under the receiver operating characteristic curve (AUC), and found that: (1) the AUC statistics for all four major racial/ethnic groups are high for both BIFSG and BISG, suggesting that the two proxy methods have very good discriminatory ability; and (2) the AUC statistics for BIFSG are statistically significantly larger than those for BISG, for each major racial/ethnic group. The AUC statistics are available upon request. I chose not to include them in the article because my data are highly skewed towards NH Whites (see ), and AUC can present an overly optimistic view of an algorithm’s performance and consequently may mask differences between different algorithms, if there is a large skew in class distribution (e.g., see Davis and Goadrich 2006).

<sup>19</sup> Traditionally, the accuracy rate, given by the sum of true positives and true negatives (i.e., cases where the proxy method correctly determines that an individual does not belong to a specific group) for a given group as a proportion of the whole sample, has been the most commonly used empirical measure. However, in the framework of unbalanced datasets, it may lead to erroneous conclusions and thus it is not appropriate to use (e.g., see Lin and Chen 2013).

**Table 2.** Self-reported race/ethnicity composition.

Data Source	Composition						N obs
	Hispanic	NH Black	NH White	NH API	NH AIAN	NH Multiracial	
National average (Census 2010)	11.1%	11.3%	70.5%	7.0%	0.9%	0.8%	
Validation Sample <sup>1</sup>	11.7%	6.6%	75.2%	6.0%	0.2%	0.4%	212,628

NOTE. <sup>1</sup>Includes all applications with valid race/ethnicity from single applicants and dual applicants with same race/ethnicity, and with nonmissing name and geographic information.

**Table 3.** Distribution of mortgage applications by race/ethnicity: proxies vs. self-report.

Method	Hispanic	NH black	NH white	NH API	Weighted average overall deviation from self-report	% Diff. in average deviation BIFSG - BISG	Relative efficiency of BIFSG vs. BISG
Self-Report	11.7%	6.6%	75.2%	6.0%	(0)		
BIFSG	12.4%	7.5%	72.6%	6.0%	2.1%	— 38.6%	164.8%
BISG	12.6%	8.1%	70.9%	6.5%	3.5%		

NOTE. Differences in prevalences between BIFSG and BISG are statistically significant at the 5% level for all racial/ethnic groups, as indicated by both pairwise t-tests and nonparametric sign tests.

To enhance the comparability of my evaluation results to the broader literature, I follow Elliott et al. (2008, 2009) in: (1) summarizing certain metrics—distributional differences between the proxy-based race/ethnicity and the self-reported attributes, and correlation coefficients—across all racial/ethnic categories by computing the weighted average of the specific metric across the six categories, with weights given by the true proportion of applicants in each category; and (2) measuring the relative efficiency of the two methods in matching the actual distribution of race/ethnicity across applicants and in predicting individual race/ethnicity.<sup>20</sup>

To ensure that the comparative accuracy of BIFSG and BISG is not driven by the underlying sample and that we obtain an unbiased account of the improvement that first name information can bring in the prediction of race/ethnicity, the evaluation samples exclude any observations that have missing values for either set of proxies.<sup>21</sup> As can be inferred from Table 1, the rate of missing values is somewhat higher (by 9 percentage points) for the BIFSG probabilities than for the BISG probabilities due to observations with missing first name information and non-missing surname and geo-demographic information. However, the somewhat lower coverage of BIFSG is not concerning, since as previously mentioned, one can always impute missing BIFSG probabilities with BISG probabilities if only the first name information is missing. Table 2 shows the size and the self-reported race/ethnicity composition of the evaluation sample compared with that for the whole United States from the Decennial Census 2010. The sample includes 212,628 applications (representing

76% of all applications with valid race/ethnicity), a sample size that is comparable to those in the studies by CFPB (2014) and Baines and Courchane (2014). Although the validation dataset has a somewhat higher proportion of NH Whites and a lower proportion of NH Blacks relative to the nation as a whole, it generally reflects well the diversity of the U.S. population—in fact, much better than any of the recent studies that use mortgage data to evaluate the BISG methodology.

### 3. Evaluation Results

This section presents the evaluation results based on the aforementioned metrics, illustrating the accuracy and coverage improvements that BIFSG achieves compared with BISG. Due to poor performance of both methods in identifying NH AIAN and NH Multiracial individuals, the results for these two categories are not shown in the paper.<sup>22</sup> However, this is not of significant concern because these two groups account for less than 1% of the application population (see Table 2).

#### 3.1. Distribution of Lending by Race/Ethnicity

Table 3 shows the distribution of mortgage applications by self-reported race/ethnicity, along with the distributions based on the predicted BIFSG and BISG probabilities. For the proxy methods, the percentages are calculated as the sum of probabilities for each category across all applicants divided by the total number of applicants, times 100. Both methods tend to underestimate the prevalence of NH Whites and overestimate the prevalence of minorities, especially NH Blacks. However, it is worth noting that BIFSG is more accurate overall, with smaller deviations from the true proportions than BISG for all racial/ethnic groups, and with a weighted average prevalence error (deviation from self-reported) that is about 39% lower than that of BISG.<sup>23</sup> Using the efficiency measure described in section 2.5, BIFSG is

<sup>20</sup>The relative efficiency in matching the actual racial/ethnic distribution is computed as the ratio of average squared deviations from the actual distribution of race/ethnicity for BIFSG and BISG; and the relative efficiency in predicting individual race/ethnicity is the ratio of squared correlations between actual race/ethnicity and the proxy probabilities for the two proxy methods. Then, to say, for example, that BIFSG has a relative efficiency of 2 compared with BISG—or, equivalently, that method BIFSG is 100% more efficient than BISG—means that the accuracy of an analysis testing for differences in a given outcome among racial/ethnic groups using BIFSG with a given sample size is the same as what would be obtained with twice the sample size using BISG. This approach was proposed by McCaffrey and Elliott (2008) and applied by Elliott et al. (2008, 2009) to evaluate BISG and its earlier version, the Bayesian Surname and Geocoding (BSG) method.

<sup>21</sup>In other words, the evaluation samples include only applicants with demographic information for all three attributes (surname, first name, and geography).

<sup>22</sup>The results for these two groups are available upon request from the author. Although these results are not shown in this article, they are included in the computation of the weighted average statistics.

<sup>23</sup>The absolute difference in deviations between the two methods is 0.60 percentage points for NH Blacks, 1.65 percentage points for NH Whites, 0.52 percentage points for NH APIs, and 0.27 percentage points for Hispanics. The relative difference in deviations between the two methods is very high for NH Blacks (40%),

**Table 4.** Correlation between proxy probability and self-reported race/ethnicity.

Method	Pearson correlation coefficients					% Diff. in average correlation BIFSG - BISG	Relative efficiency of BIFSG vs. BISG
	Hispanic	NH Black	NH White	NH API	Weighted average		
BIFSG	0.884	0.747	0.840	0.873	0.837	2.6%	5.4%
BISG	0.869	0.712	0.818	0.864	0.815		

NOTE. Differences between the BIFSG and BISG correlations are statistically significant at the 5% level for all racial/ethnic groups, as indicated by a comparison of the 95% confidence intervals for Fisher's Z transformation of the correlation coefficients.

almost 165% more efficient than BISG in estimating prevalences. To put these findings in perspective, the BIFSG improvement over BISG is considerably larger than the improvement of BSG (the earlier version of BISG proposed by Elliot et al. 2008) relative to the GO method.<sup>24</sup>

### 3.2. Predicting Individual Race/Ethnicity

#### 3.2.1. Correlations between the Proxy Probability and Self-Reported Race/Ethnicity

Table 4 displays the Pearson correlation coefficients between the self-reported race/ethnicity and the proxy probabilities generated by BIFSG and BISG. The correlation coefficients between BIFSG probabilities and self-reported race/ethnicity ranges from 0.747 to 0.884, with a weighted average across all categories of 0.837. The correlation coefficients for the BISG method are all statistically significantly smaller than their BIFSG counterparts, ranging from 0.712 to 0.869, with a weighted average across all categories of 0.815. Thus, BIFSG results in 2.6% higher average correlation and 5.4% higher overall efficiency than BISG. The difference in average correlation between the two methods also means that BIFSG removes 11.6% of the remaining predictive error from BISG, with the reduction being calculated as  $[1 - (1 - 0.837)/(1 - 0.815)] \times 100$ .

The improvements associated with BIFSG are largest for NH Blacks (5% higher correlation, 10.3% higher efficiency, and 12.4% reduction in the remaining predictive error from BISG) and NH Whites (2.8% higher correlation, 5.6% higher efficiency, and 12.5% reduction in the remaining predictive error from BISG). Moreover, the magnitude of improvement for NH Blacks depends on the level of minority concentration in a specific geography. In other words, the improvement is much higher when geography has low ability to distinguish NH Blacks, that is, in areas with low concentration levels for NH Blacks. Table 5 illustrates this point by showing the correlation between the self-reported race/ethnicity and the proxy probabilities across different levels of geographic concentration for NH Blacks.<sup>25</sup> The improvement associated with BIFSG in areas with low concentration—an increase of 11.6% in the correlation coefficient over BISG—is much larger than the improvements

**Table 5.** Correlation between proxy probability and self-reported race/ethnicity for NH blacks, by concentration level.

Concentration level	Correlation coefficient		% Diff. in correlation BIFSG - BISG
	BIFSG	BISG	
High	0.846	0.826	2.5%
Medium	0.730	0.692	5.5%
Low	0.614	0.551	11.6%
Overall	0.747	0.712	5.0%

NOTE. Concentration levels are determined based on county-specific dissimilarity indices for NH Blacks: High if index  $\geq 0.6$ , Medium if  $0.4 < \text{index} < 0.6$ , and Low if index  $< 0.4$ . The areas with low concentration level account for 38% of the validation sample, those with medium concentration level for 45%, and those with high concentration level for 17%.

obtained in areas with higher concentration, and over twice as large as the improvement estimated for the United States as a whole.

Again putting findings in perspective, the BIFSG improvement over BISG is similar to the improvement of BISG over BSG and that of BSG over SO for Hispanics, and to the improvement of BSG over GO for NH Blacks, and larger than the improvement of BSG over SO for NH API.

#### 3.2.2. Classification Accuracy

The improvements of BIFSG also extend to the accuracy and coverage of various classification schemes. As previously mentioned, accuracy is measured by the rates of false negatives and false positives (not including unclassified applicants), and the coverage is computed as the proportion of applicants of a given race/ethnicity that cannot be classified (the higher the proportion of unclassified applicants, the lower the coverage).

The evaluation results shown in the paper pertain to the maximum a posteriori (MAP) classification scheme, which is commonly used for naive Bayesian classifiers (e.g., see Dai and Su 2014). This scheme classifies each applicant in the racial/ethnic category corresponding to the largest proxy probability, and, since it classifies all applicants, it has maximum coverage. As shown in Table 6, the BIFSG proxies generally have higher accuracy compared with the respective BISG proxies, with the largest improvements occurring for NH Blacks: a 3.3 percentage point decrease in the rate of false negatives, and a 4 percentage point decrease in the rate of false positives. Notably, BIFSG's improved accuracy compared with BISG for NH Blacks is much larger in geographic areas with low concentration of NH Blacks.<sup>26</sup> An alternative scheme that has been suggested in

NH Whites (38%), and NH APIs (99%). Differences in prevalences between the two methods are statistically significant for all groups.

<sup>24</sup>Specifically, Elliott et al. (2008) found that BSG has 20% lower average deviation from self-reported and it is 56% more efficient than GO.

<sup>25</sup>The level of geographic concentration is identified with the dissimilarity index, which measures the percentage of the NH Black population that would have to change residence for each census block group in a given county to have the same percentage of NH Blacks as the county overall. The dissimilarity index ranges from 0.0 (complete integration) to 1.0 (complete segregation). The cut-offs for the High, Medium, and Low levels of the index are based on recommendations by the Diversity and Disparities data project of Brown University, available at: <https://s4.ad.brown.edu/projects/diversity/segregation2010/Default.aspx?msa=45780>.

<sup>26</sup>In geographic areas with low concentration of NH Blacks, the BIFSG proxy for NH Blacks has 7.8% lower false negative rate and 6.1% lower false positive rate than the BISG proxy. Detailed results on the variation of error rates with the level of concentration for NH Blacks are available upon request from the author.



**Table 6.** Accuracy metrics for the MAP classification method.

Method	Race/ ethnicity	Total applications in group		False negative rate (%)	False positive rate (%)
		Based on self-reporting	Estimated by proxy method		
BIFSG	Hispanic	24,794	25,671	10.3	13.3
BISG	Hispanic	24,794	25,561	11.7	14.4
BIFSG	NH Black	13,941	12,937	31.4	26.1
BISG	NH Black	13,941	13,023	34.7	30.1
BIFSG	NH White	159,960	161,374	4.2	5.0
BISG	NH White	159,960	161,209	4.8	5.5
BIFSG	NH API	12,700	12,342	14.7	12.2
BISG	NH API	12,700	12,722	14.3	14.5

NOTE. MAP classifies all applicants and thus has maximum coverage.

False negatives in a particular group are applicants who self-report belonging to that group, but whom the proxy method categorizes into another group. The false negative rate is the ratio of the number of false negatives in that group to the total population that self-reports belonging to that group.

False positives in a particular group are applicants whom the proxy method assigns to that group, when in fact they belong to another group. The false positive rate is computed as the ratio of the number of false positives categorized in that group to the total population that the proxy classifies in that group.

related research classifies the application in a given racial/ethnic category if the proxy probability for that category is larger than a specified threshold, where the threshold value is typically set at 80%. Notably, while MAP classifies all applicants, the threshold-based scheme does not classify applicants for which all six proxy probabilities are below the specified threshold. I do not show detailed results from the 80% threshold scheme in the paper, and instead include them in the online supplementary materials (see Appendix SA1), because my research shows that this threshold-based scheme is somewhat inferior to MAP<sup>27</sup> and results in similar findings on the relative accuracy of the BIFSG and the BISG methods, albeit with somewhat smaller differences as compared with MAP. However, it is worth noting that for the threshold-based scheme, BIFSG has significantly better coverage than BISG both overall and for each specific racial/ethnic category. Specifically, the overall proportion of unclassified applicants is 11.4% for BIFSG and 15.9% for BISG; and, across groups, the improvement in coverage associated with BIFSG (as measured by the difference in the proportion of unclassified applicants between the two proxy methods) ranges from 1.5 to 7.1 percentage points, with the largest improvement occurring for NH Blacks.<sup>28</sup>

### 3.3. Estimating the Race/Ethnicity Effects on Mortgage Lending Outcomes

The BISG method has been criticized by Baines and Courchane (2014) in that, despite its superiority relative to other alternatives, it is still subject to significant bias and estimation error, and overstates the effects of race/ethnicity on mortgage lending outcomes. To evaluate the biases that proxies may cause, I use mortgage pricing and underwriting decisions as examples of outcomes. Specifically, I regress APR and a denied/approved indicator on race/ethnicity, first using actual race/ethnicity,

<sup>27</sup> Specifically, while the 80% threshold scheme has slightly higher accuracy than MAP, it has significantly lower coverage, with the proportion of unclassified individuals based on BIFSG ranging from 9% for Hispanics and NH Whites to 36% for NH Blacks. Additionally, as shown in the online appendix (Appendix SA1), the threshold-based scheme results in larger biases when estimating the effects of race/ethnicity on mortgage lending outcomes, compared with MAP.

<sup>28</sup> The coverage improvements associated with BIFSG are even larger when measured by the relative (percentage) difference in the proportion of unclassified applicants between the two proxy methods, ranging from 10% to 36% across groups, and amounting to nearly 29% overall.

and then using the BIFSG and BISG proxies—both as continuous probabilities and as discrete classifications. Unlike Baines and Courchane (2014) who estimate raw (unconditional) race/ethnicity effects, I estimate adjusted (conditional) effects using regression models that control for important mortgage pricing and underwriting factors. Adjusted effects are much more informative regarding biases from the proxies' errors than raw effects for two related reasons: (1) analyses of the race/ethnicity effects on mortgage lending outcomes are typically carried out in a multivariate framework to account for various determinants of the lending outcome that may be correlated with race/ethnicity; and (2) biases in the race/ethnicity effects resulting from the errors of proxy methods depend on the correlations between these errors and the other explanatory variables included in the regression, as well as on the correlations between the true (self-reported) race/ethnicity and the other regressors (e.g., see Bound, Brown, and Mathiowetz 2000; Angrist and Krueger 1999).

The APR regressions are estimated using OLS on a sample that includes only originated loans, whereas the denied/approved decision is modeled using logistic regression on a sample that includes originated loans as well as approved but not accepted and denied applications. Both estimation samples include only observations for which both BIFSG and BISG proxies can be computed. The omitted race/ethnicity category is NH Whites. To control for mortgage pricing factors in the APR regressions, I include the following explanatory variables: FICO score, combined loan-to-value-ratio, loan amount (in logarithmic form), and indicators for: rate type (fixed vs. adjustable), loan type (conventional/FHA/VA), property type (1–4 family vs. other), owner occupancy, subordinate lien status, loan purpose (home purchase/refinance),<sup>29</sup> the year-quarter when the rate was locked, the state of property location, and origination channels (retail, broker, correspondent lender) or business units with different underwriting and pricing policies, which are specific to each lender. To control for mortgage underwriting factors in the denied/approved regressions, I include

<sup>29</sup> I exclude applications for home improvement loans because they typically have very distinct underwriting and pricing guidelines. Nonetheless, in alternative specifications, I also experimented with including these applications in the estimation samples while controlling for them with a dummy variable, and obtained similar results. These alternative results are available upon request from the author.

**Table 7.** Comparison of race/ethnicity effects in pricing and underwriting: self-report vs proxies.

A. Adjusted race/ethnicity effects on APR <sup>1</sup>											
Method	N obs	Hispanic			NH Black			NH API			
		Coef. (bps)	Bias (bps)	Bias difference (bps)	Coef. (bps)	Bias (bps)	Bias difference (bps)	Coef. (bps)	Bias (bps)	Bias difference (bps)	
self-report	122,836	5.4			5.5			−12.0			
BIFSG - prob	122,836	6.3	0.9	−1.4	6.9	1.4	−0.1	−14.1	−2.1	0.4	
BISG - prob	122,836	7.8	2.4		7.0	1.5		−13.7	−1.7		
BIFSG - MAP	122,836	4.9	−0.5	0.3	5.4	−0.1	−0.6	−12.8	−0.7	0.6	
BISG - MAP	122,836	5.3	−0.1		4.8	−0.7		−12.2	−0.1		

B. Adjusted race/ethnicity effects on denial odds <sup>2</sup>											
Method	N obs	Hispanic			NH Black			NH API			
		Odds ratio	Bias	Bias difference	Odds ratio	Bias	Bias difference	Odds ratio	Bias	Bias difference	
self-report	173,899	1.501			1.495			1.355			
BIFSG - prob	173,899	1.588	0.087	−0.026	1.667	0.172	0.006	1.423	0.068	0.022	
BISG - prob	173,899	1.613	0.113		1.661	0.166		1.401	0.046		
BIFSG - MAP	173,899	1.478	−0.022	−0.005	1.479	−0.016	−0.005	1.363	0.008	−0.002	
BISG - MAP	173,899	1.474	−0.027		1.474	−0.021		1.344	−0.010		

NOTE. The omitted race category is NH Whites. All effects are statistically significant at the 1% level. The bias is computed as the difference in the relevant coefficient or odds ratio between the proxy method and the self-report method. The bias difference is computed as the difference between the absolute value of the BIFSG bias and the absolute value of the BISG bias. Thus, a negative difference means that the magnitude of the BIFSG bias is smaller than that of the BISG bias. The instances where the bias or the bias difference appears to be slightly off are due to rounding.

<sup>1</sup>The adjusted race/ethnicity effects on APR are obtained from OLS regressions of APR on race/ethnicity indicators or probabilities, and controls for mortgage pricing factors. "Bps" stands for "basis points," where 1 basis point = 0.01%.

<sup>2</sup>The adjusted race/ethnicity effects on denial odds are obtained from logistic regressions of the denied/approved indicator on race/ethnicity indicators or probabilities and controls for mortgage underwriting factors.

the following explanatory variables: FICO score, combined loan-to-value-ratio, debt-to-income ratio, loan amount (in logarithmic form), and the aforementioned indicators for rate type, loan type, property type, owner occupancy, subordinate lien status, loan purpose, and lender-specific origination channels or business units. Since some of the control variables used in the underwriting and pricing specifications have missing values, both specifications also include missing value indicators for these variables, to alleviate potential sample selection problems that may arise from dropping observations with missing data.<sup>30</sup>

The relevant regression results are shown in Table 7. There are two noteworthy patterns that can be observed in this table. First, discrete classification (MAP) results in smaller biases in the race/ethnicity effects on both APR and denial odds than continuous probabilities. As an example, the BIFSG-induced bias in the denial odds ratio for NH Blacks is 0.172 when using continuous probabilities, but only 0.016 when using MAP. These findings are consistent with the analysis in Domingos and Pazzani (1997), which shows that the naive Bayesian classifier performs well in terms of classification accuracy but not necessarily in terms of the accuracy of estimated probabilities.

Second, and most importantly, the biases from both BIFSG and BISG proxies are generally very small—maximum 2.4 basis points (bps) in the APR regressions and, with few exceptions,<sup>31</sup> less than 0.11 in the denied/approved regressions—and, consequently, differences in these biases between the two methods are trivial—up to 1.4 bps in the APR regressions and up to 0.026 in

the denied/approved regressions. It is also worth noting that in the cases with the largest bias differences between the two methods (1.4 bps for APR for Hispanics and 0.026 for denial odds ratios for NH Blacks, using continuous probabilities), BIFSG produces the smaller bias.

### 3.4. Robustness Check—Evaluation Using Voter Registration Data

A formal evaluation of the applicability of BIFSG to other national contexts or to U.S.-based applications other than mortgage lending is outside the scope of the paper, largely due to considerable data challenges. Nonetheless, I provide a glimpse of the BIFSG applicability to other fields by evaluating its predictive performance on voter registration data for North Carolina, in conditions in which there is a growing literature on the use of names to identify race/ethnicity in political science that could benefit from leveraging improved proxy methods (e.g., see Grofman and Garcia 2014; Enos 2016). While voter registration data on a national scale would provide for a more representative test of the BIFSG capabilities to the extent that the algorithm performance varies across geographic areas, it would be very difficult to assemble such a dataset given that the great majority of states do not collect race/ethnicity information in their voter registration files. The evaluation sample, which has more than 3.6 million observations, includes all active voters in North Carolina as of September 2017, who were born on or after 1925, and for whom race/ethnicity, as well as the name and geographic information, is available. For brevity, the detailed evaluation results based on the North Carolina voter registration data are available in the online supplementary information (see Appendix SA2). These results illustrate that the BIFSG performs significantly better than the

<sup>30</sup>However, the extent of missing data is small—5% of the observations in the underwriting sample and 11% of those in the pricing sample have missing values for at least one variable—and regression analyses that exclude observations with missing values produce similar results with those presented in the article. These alternative results are available upon request from the author.

<sup>31</sup>The exceptions are biases in denial odds ratios of about 0.170 for NH Blacks based on continuous probabilities.

**Table 8.** BIFSG application: Distribution of loans by race/ethnicity.

Race/ethnicity	Self-Report	BIFSG-prob		BIFSG-MAP	
	Sample with race/ethnicity	Sample with race/ethnicity	Sample without race/ethnicity	Sample with race/ethnicity	Sample without race/ethnicity
Hispanic	13.6%	14.6%	13.0%	14.3%	12.8%
NH Black	5.9%	6.7%	9.2%	5.3%	7.7%
NH White	73.4%	70.4%	68.9%	73.9%	72.5%
NH API	6.5%	6.5%	7.0%	6.4%	6.8%

NOTE. Both the sample with race/ethnicity and the sample with missing race/ethnicity have 10,000 applications. Differences in prevalences between the two samples are statistically significant at the 5% level for all groups except NH API. The statistical significance of the difference in the probability-based prevalence is determined using *t*-tests with Satterthwaite approximation that account for unequal variances of the probability distributions in the two samples. Additionally, since the distribution of BIFSG probabilities for a given group is skewed, I also test whether the difference in the probability distribution between the two samples is statistically significant, using the Wilcoxon–Mann–Whitney nonparametric test. The statistical significance of the difference in the MAP-based prevalence is determined using chi-square tests.

BISG. In fact, the BIFSG improvements over BISG, in terms of prediction accuracy metrics (correlation, classification errors) and coverage, are significantly larger than those obtained with the nationwide mortgage data. These findings demonstrate that BIFSG is also applicable to fields other than mortgage lending.

#### 4. Application: Imputation of Race/Ethnicity in HMDA Data

This section demonstrates the application of BIFSG to the imputation of missing race/ethnicity information in the HMDA data. As mentioned in the Introduction, a portion of the mortgage applications in the HMDA data are missing race and/or ethnicity. For example, in HMDA 2014, about 22% of all applications and nearly 15% of the applications processed by the reporting institution (excluding purchased loans) are missing this information. The use of race/ethnicity proxies offers an easy way to check and, if necessary, alleviate selection bias and efficiency problems that may arise in studies of racial/ethnic trends in the HMDA data if race/ethnicity is missing for systematic reasons, and therefore represents a worthwhile application of BIFSG.

In this application, I randomly select 10,000 applications from the sample with available race/ethnicity used in the BIFSG evaluation, and 10,000 applications for which actual race/ethnicity are missing but name and geographic information is available (so that the BIFSG probabilities can be computed).<sup>32</sup> For each of the selected applications, I compute the BIFSG proxy probabilities and the associated MAP classifications, and then compare the racial/ethnic prevalences based on these proxies in the sample for which actual race/ethnicity information is available with those in the sample for which this information is missing. This comparative analysis is intended to provide novel evidence on the correlation between race/ethnicity and whether an application is missing this information.<sup>33</sup>

The application results are summarized in Table 8. First, notice that in the sample with nonmissing race/ethnicity, prevalences based on either probabilities or MAP classifications are very similar to those based on actual race/ethnicity—a result that is consistent with the evaluation exercises. Turning to the focus of the comparative analysis, results provide evidence of some correlation between race/ethnicity and whether an application is missing this information, with the incidence of missing race/ethnicity being higher among NH Blacks than among other groups. Specifically, NH Black is the only group with a higher prevalence in the sample with missing race/ethnicity information than in the sample with nonmissing race/ethnicity information, with the difference being around 2.5% and statistically significant at any conventional level. In contrast, NH Whites and Hispanics have lower prevalences in the sample with missing race/ethnicity than in the sample with nonmissing race/ethnicity.

#### 5. Conclusions

Previous indirect methods to estimate race/ethnicity have underutilized first name information due to the lack of comprehensive lists of first names and their associated race/ethnicity distributions. In this article, I develop an enhanced Bayesian method—the Bayesian First Name Surname Geocoding method (BIFSG)—for predicting race and ethnicity, using a new first name list offered in Tzioumis (2017). The new method improves on the existing BISG naive-Bayesian algorithm by considering first name information, along with surname and geographic information. Using data from mortgage applications and voter registration, and applying a wide array of performance metrics, I demonstrate that BIFSG results in nontrivial improvements over BISG, in terms of accuracy and coverage. These improvements hold for both the continuous probabilities and the discrete classification schemes across racial/ethnic groups.

While the overall magnitude of the improvement associated with BIFSG is somewhat modest, the largest improvements occur for NH Blacks, which is the group for which BISG is least accurate. Moreover, the improvement for NH Blacks is much higher where geography has low ability to distinguish NH Blacks. This aspect is particularly important as much of the research on the topic of racial/ethnic differences focuses on specific geographic areas rather than the entire United States. It is also worthwhile to note that the improvements of BIFSG over BISG are generally comparable to the improvements of BISG over simpler methods. Last but not least, when assessing the

<sup>32</sup>In additional analyses, I select the random samples from the set of applications for which at least one of the three input probabilities (surname-based, first-name-based or geography-based) is available, and then use the “extended” BIFSG algorithm that can create proxies even when just one of the input probabilities is available. Given the trivial number of applications for which all input probabilities are missing, this approach ensures that the results are representative of the whole population of applications with missing race/ethnicity (not just those for which all three input probabilities can be computed). The results from these alternative analyses are very similar with the ones presented in the paper. They are available upon request from the author.

<sup>33</sup>While Huck (2001) and Dietrich (2002) also analyzed this correlation, they impute missing race/ethnicity largely based on applicants’ geographic location—an approach that has been shown to produce much less accurate predictions than the Bayesian methods discussed in this article.

degree of improvement from BIFSG, one should consider that even the most advanced methods are likely to result in incremental improvements for Hispanics and NH Asians, given that surnames alone are highly predictive for these particular groups.

As an additional test, I evaluate the bias from using proxies rather than the actual race/ethnicity classification in underwriting and pricing regressions that control for a number of relevant creditworthiness parameters. The results demonstrate that the biases in the race/ethnicity effects on APR and denial odds from both BIFSG and BISG proxies are economically insignificant, with the BIFSG generally having a smaller bias. Additionally, I find that a discrete classification scheme such as MAP results in smaller biases than continuous probabilities.

Following evaluation, I demonstrate the application of BIFSG to the imputation of missing race/ethnicity information in the HMDA data, and in the process, provide novel evidence that race/ethnicity are somewhat correlated with the incidence of missing race/ethnicity information. Specifically, I find that the incidence of missing race/ethnicity information is higher among NH Blacks than among other groups.

Given its improvements over the previously most advanced method to estimate race/ethnicity, BIFSG represents a worthwhile alternative when direct information on these demographic characteristics is not available. Moreover, the BIFSG's discriminatory power has the potential to increase further once more comprehensive first name lists (akin to the existing U.S. Census surname lists) become available. And, to the extent that a first name list based on mortgage applications may limit the BIFSG's applicability to general populations, the advent of more comprehensive first name lists would also likely enhance the generalizability of the algorithm.

## Appendix A: Derivation of the BIFSG formula

Using the notation described in Section 2.2, the conditional probability  $p(r|s, f, g)$  can be written according to Bayes' Rule as follows:

$$p(r|s, f, g) = \frac{p(s, f, g|r) p(r)}{p(s, g, f)} \quad (A1)$$

The joint probability  $p(s, g, f)$  can be re-written by conditioning on  $r$  and summing over the discrete values of  $r$ , i.e.,

$$p(s, g, f) = \sum_{r=1}^6 p(s, f, g|r) p(r) \quad (A2)$$

Then, substituting (A2) into (A1), we obtain:

$$p(r|s, f, g) = \frac{p(s, f, g|r) p(r)}{\sum_{r=1}^6 p(s, f, g|r) p(r)} \quad (A3)$$

Since the conditional probability  $p(s, f, g|r)$  can also be written as  $p(r, s, f, g)/p(r)$ , (A3) becomes:

$$p(r|s, f, g) = \frac{p(r, s, f, g)}{\sum_{r=1}^6 p(r, s, f, g)} \quad (A4)$$

Then, we can use the chain rule to decompose the joint probability  $p(r, s, f, g)$  into:

$$p(r, s, f, g) = p(s) \cdot p(r|s) \cdot p(g|r, s) \cdot p(f|r, s, g) \quad (A5)$$

**Table B1.** BIFSG and BISG calculation example.

Attributes	Hispanic	NH Black	NH White	NH API	NH AIAN	NH Multiracial
Surname: DAVIS	2.4%	31.6%	62.2%	0.5%	0.8%	2.5%
First name: LATOYA	0.0022%	0.0759%	0.0002%	0.0000%	0.0000%	0.0000%
Geography: CBG 340297175012	0.0006%	0.0004%	0.0011%	0.0002%	0.0001%	0.0009%
<b>BIFSG probability</b>	<b>0.3%</b>	<b>98.5%</b>	<b>1.2%</b>	<b>0.0%</b>	<b>0.0%</b>	<b>0.0%</b>
<b>BISG probability</b>	<b>1.7%</b>	<b>15.6%</b>	<b>79.9%</b>	<b>0.1%</b>	<b>0.1%</b>	<b>2.6%</b>

**Table B2.** Distribution of the difference between the BIFSG and BISG probabilities.

Race/Ethnicity	% population with large differences between BIFSG and BISG probabilities		
	Total	BIFSG prob > BISG prob	BIFSG prob < BISG prob
Hispanic	21.4%	16.2%	5.2%
NH Black	34.8%	23.4%	11.5%
NH White	14.6%	11.2%	3.4%
NH API	19.5%	10.2%	9.4%

NOTE: Large probability differences are those larger than 10%age points.

Substituting (A5) into (A4), we obtain:

$$p(r|s, f, g) = \frac{p(s) \cdot p(r|s) \cdot p(g|r, s) \cdot p(f|r, s, g)}{\sum_{r=1}^6 p(s) \cdot p(r|s) \cdot p(g|r, s) \cdot p(f|r, s, g)} \quad (A6)$$

If one assumes conditional independence among  $s$ ,  $g$ , and  $f$ , i.e.,  $p(g|r, s) = p(g|r)$  and  $p(f|r, s, g) = p(f|r)$ , (A6) simplifies to formula (1) in section 2.2, as follows:

$$\begin{aligned} p(r|s, f, g) &= \frac{p(s) \cdot p(r|s) \cdot p(g|r) \cdot p(f|r)}{\sum_{r=1}^6 p(s) \cdot p(r|s) \cdot p(g|r) \cdot p(f|r)} \\ &= \frac{p(r|s) \cdot p(g|r) \cdot p(f|r)}{\sum_{r=1}^6 p(r|s) \cdot p(g|r) \cdot p(f|r)} \end{aligned} \quad (A7)$$

## Appendix B: Explanation of the effect of first names on the sorting quality of the proxy probabilities

To gain insight into how the addition of first name information helps improve the sorting quality of the proxy probabilities, I investigate the distribution of the difference between the BIFSG and BISG probabilities for each of the four major racial/ethnic groups. For example, for each applicant whose self-reported race/ethnicity is NH Black, I compute the difference between the BIFSG probability of being NH Black and the BISG probability of being NH Black, and then analyze the frequency distribution of this difference.<sup>34</sup>

Before reviewing the statistics, it is helpful to go over an example. Consider a hypothetical NH Black applicant, Latoya Davis, living in a mostly NH White census block group. As illustrated in Table B1, this applicant has a very large difference between the BIFSG and BISG probabilities: the BIFSG probability of being NH Black is 98.5%, whereas the corresponding BISG probability is only 15.6%. The reason for this large difference is the additional information that the first name contributes

<sup>34</sup> A positive difference means that the BIFSG probability is the larger of the two.



to the probability calculation under the BIFSG method. Specifically, while the surname, Davis, is predominantly white, with 62.2% of people with that surname being white and only 31.6% being black, the first name, Latoya, is predominantly black, with 91.4% of people with that first name being black.<sup>35</sup> Since the BISG method does not use the first name information, it will assign a low probability of being black and a high probability of being white. By comparison, the BIFSG method upgrades the probability of being black significantly due to the addition of the demographic information associated with the first name.

Table B2 describes the distribution of the probability differences for each racial/ethnic group. Two results are worth noting. First, for most groups, a considerable proportion of the applicants have sizeable differences between the BIFSG and BISG probabilities. NH Blacks have the distribution with the heaviest tails—almost 35% of the applicants in this group have differences between the two proxy probabilities larger than 10% age points in absolute value. The Hispanic and NH API groups also have significant proportions of applicants with probability differences in this range—21.4% and 19.5%, respectively. For NH Whites, a smaller proportion (14.6%) of the NH White applicants are in the tails of the distribution; however, given the large NH White population, this proportion translates into a large applicant count (23,363).

Second, the distributions are asymmetric, with larger proportions of applicants in the right tail, where the BIFSG probability is greater than the BISG probability. The distribution asymmetry is particularly strong for NH Blacks, Hispanics, and NH Whites. Specifically, 23.4% of the NH Black applicants have a BIFSG probability of being NH Black that is larger than the corresponding BISG probability by more than 10%age points, but only 11.5% of the NH Blacks are assigned a BIFSG probability that is *smaller* than the BISG probability by more than 10%age points. Also, 16.2% of the Hispanic applicants are in the right tail compared with only 5.2% in the left tail, and 11.2% of the NH White applicants are in the right tail, whereas only 3.4% are in the left tail.

In summary, there are significantly more individuals for whom BIFSG increases the probability of belonging to the right group (relative to BISG) than for whom it reduces that probability. Therefore, BIFSG produces probabilities that, on average, are closer aligned with the actual race/ethnicity of an individual; hence, the higher predictive ability of the BIFSG probabilities relative to the BISG probabilities.

## Acknowledgments

I thank Konstantinos Tzioumis and Chau Do for very helpful comments on a previous draft, and Peter Trubey and Regina Villasmil for excellent research assistance. The opinions expressed in this article are those of the author alone, and do not necessarily reflect the views of the Office of the Comptroller of the Currency or the U.S. Department of the Treasury.

## Supplementary Materials

**Appendix SA1:** Evaluation Results for the 80% Threshold Classification Method.

See file Appendix\_SA1\_80pct\_results

**Appendix SA2:** Evaluation Results Based on Voter Registration Data for North Carolina.

See file Appendix\_SA2\_NC\_voters\_results

Supplemental data for this article can be accessed on the [publisher's website](#).

## References

- Adjaye-Gbewonyo, D., Bednarczyk, R. A., Davis, R. L., and Omer, S. B. (2014), "Using the Bayesian Improved Surname Geocoding Method (BISG) to Create a Working Classification of Race and Ethnicity in a Diverse Managed Care Population: A Validation Study," *Health Services Research*, 49 (1, Part I), 268–283. [2]
- Angrist, J. D., and Krueger, A. B. (1999), "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics* (Vol. 3, pp. 1277–1366), eds. O. Ashenfelter and D. Card, Amsterdam, The Netherlands: Elsevier B.V. [8]
- Baines, A. P., and Courchane, M. J. (2014), "Fair Lending: Implications for the Indirect Auto Finance Market," study prepared for the American Financial Services Association. [2,5,6,8]
- Bound, J., Brown, C., and Mathiowetz, N. (2000), "Measurement Error in Survey Data," PSC Research Report No. 00-450. 8 2000. [8]
- Coldman, A. J., Braun, T., and Gallagher, R. P. (1988), "The Classification of Ethnic Status using Name Information," *Journal of Epidemiology and Community Health*, 42, 390–395. [2]
- Consumer Financial Protection Bureau (2014), "Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity: A Methodology and Assessment," Washington, DC. Available at [http://files.consumerfinance.gov/f/201409\\_cfpb\\_report\\_proxy-methodology.pdf](http://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf) [2,5,6]
- Dai, Yugang, and Su, H. (2014), "The Naive Bayes Text Classification Algorithm Based on Rough Set in the Cloud Platform," *Journal of Chemical and Pharmaceutical Research*, 6 (7), 1636–1643. [7]
- Davis, J., and Goadrich, M. (2006), "The Relationship Between Precision-Recall and ROC Curves," in *Proceedings of the 23rd International Conference on Machine Learning*, 233–240. [5]
- Dietrich, J. (2002), "Missing Race Data in HMDA and the Implications for the Monitoring of Fair Lending Compliance," *Journal of Housing Research*, 13 (1), 51–84. [10]
- Domingos, P., and Pazzani, M. (1997), "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier," *Machine Learning*, 29, 103–130. [4,5,9]
- Elliott, M., Fremont, A., Morrison, P., Pantoja, P., and Lurie, N. (2008), "A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity," *Health Services Research*, 43 (5p1), 1722–1736. [5,7]
- Elliott, M., Morrison, P., Fremont, A., McCaffrey, D., Pantoja, P., and Lurie, N. (2009), "Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities," *Health Services and Outcomes Research Methodology*, 9(2), 69–83. [5]
- Enos, R. J. (2016), "What the Demolition of Public Housing Teaches Us about the Impact of Racial Threat on Political Behavior," *American Journal of Political Science*, 60(1), 123–142. [9]
- Fiscella, K., and Fremont, A. (2006), "Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity," *Health Services Research*, 41 (4p1), 1482–1500. [2,3]
- Fremont, A., and Lurie, N. (2004), *The Role of Race and Ethnic Data Collection in Eliminating Health Disparities*, Washington, D.C.: National Academies Press. [1]
- Grofman, B., and Garcia, J. (2014), "Using Spanish Surname to Estimate Hispanic Voting Population in Voting Rights Litigation: A Model of Context Effects Using Bayes' Theorem," *Election Law Journal*, 13(3), 375–393. [9]
- Huck, P. (2001), "Home Mortgage Lending by Applicant Race/Ethnicity: Do HMDA Figures Provide a Distorted Picture," *Housing Policy Debate*, 12(4), 719–736. [10]
- Krieger, N., Chen, J. T., Waterman, P. D., Soobader, M. J., Subramanian, S. V., and Carson, R. (2002), "Geocoding and Monitoring of US

<sup>35</sup> BIFSG does not use the intra-first-name population shares (rather, it uses the share of the U.S. population of each racial/ethnic group bearing the specific first name, shown in Table B1); however, I use them here (e.g., 91.4%) to better illustrate the point.

- Socioeconomic Inequalities in Mortality and Cancer Incidence: Does the Choice of Area-Based Measure and Geographic Level Matter?: The Public Health Disparities Geocoding Project,” *American Journal of Epidemiology*, 156, 471–82. [3]
- Lauderdale, D., and Kestenbaum, B. B. (2000), “Asian American Ethnic Identification by Surname,” *Population and Development Review*, 19(3), 283–300. [1]
- Lin, W. J., and Chen, J. J. (2013), “Class-Imbalanced Classifiers for High-Dimensional Data,” *Briefings in Bioinformatics*, 14(1), 13–26. [5]
- Martino, S., Weinick, R., Kanouse, D. E., Brown, J., Haviland, A. M., Goldstein, E., Adams, J. L., Hambarsoomian, K., and Elliott, M. N. (2013), “Reporting CAHPS and HEDIS Data by Race/Ethnicity for Medicare Beneficiaries,” *Health Services Research*, 48(2), 417–434. [2]
- Mateos, P. (2007), “An Ontology of Ethnicity Based Upon Personal Names: With Implications of Neighborhood Profiling,” unpublished Ph.D. dissertation, University College London, Dept. of Geography. [2]
- McCaffrey, D., and Elliott, M. (2008), “Power of Tests for a Dichotomous Independent Variable Measured with Error,” *Health Services*, 43, 1085–1101, DOI: [10.1111/j.1475-6773.2007.00810.x](https://doi.org/10.1111/j.1475-6773.2007.00810.x) [5]
- Perkins, R. C. (1993), “Evaluating the Passel-Word Spanish Surname List: 1990 Decennial Census Post Enumeration Survey Results,” U.S. Census Bureau, Population Division. Available at <https://www.census.gov/population/www/documentation/twps0004.html>. [1]
- Tzioumis, K. (2017), “Demographic Aspects of First Names,” *Scientific Data*, forthcoming. The first name list is available at: <https://dx.doi.org/10.7910/DVN/TYJKEZ> [2,3,10]