Using First Name Information to Improve Race and Ethnicity Classification

Ioan Voicu

Office of the Comptroller of the Currency
Compliance Risk Analysis Division

This version: February 22, 2016

I thank Kostas Tzioumis and Chau Do for very helpful comments on a previous draft, and Peter Trubey for excellent research assistance. The opinions expressed in this paper are those of the author alone, and do not necessarily reflect the views of the Office of the Comptroller of the Currency or the US Department of the Treasury.
Using First Name Information to Improve Race and Ethnicity Classification

Abstract

This paper uses a recent first name list to improve on a previous Bayesian classifier, the Bayesian Improved Surname Geocoding (BISG) method, which combines surname and geography information to impute missing race and ethnicity. The proposed approach is validated using a large mortgage lending dataset for whom race and ethnicity are reported. The new approach results in improvements in accuracy and in coverage over BISG for all major ethno-racial categories. The largest improvements occur for non-Hispanic Blacks, a group for which the BISG performance is weakest. Additionally, when estimating disparities in mortgage pricing and underwriting among ethno-racial groups with regression models, the disparity estimates based on either BIFSG or BISG proxies are remarkably close to those based on actual race and ethnicity. Following evaluation, I demonstrate the application of BIFSG to the imputation of missing race and ethnicity in the Home Mortgage Disclosure Act (HMDA) data, and in the process, offer novel evidence that race and ethnicity are somewhat correlated with the incidence of missing race/ethnicity information.

Keywords: Race and ethnicity, first name, surname, geography

JEL Classifications: C81; J15
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, and socio-economic status. 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. For example, until recently many viewed the collection of race or 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 and ethnicity of applicants for non-mortgage products; and, although they are required to collect such information for mortgage applications, under the Home Mortgage Disclosure Act (HMDA) of 1975, a significant proportion of these applications are missing this information, primarily either because the applicants decline to provide it in mail, internet or telephone applications or because the reporting of this information for purchased loans is optional.1,2

Absent direct information on race and ethnicity, practitioners and researchers have turned to methods of estimating these demographics indirectly, based on other parameters 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 parameters are associated with a specific race or ethnicity. Several indirect methods for estimating race and ethnicity have been proposed, some based on surname information, some on

---

1 The HMDA mandate has sparked a flurry of regulatory examinations and academic research on the field of discrimination since 1984.
2 In face-to-face applications, if a customer declines to provide race and ethnicity, the loan officer is required to record this data based on his or her self-assessment of the customer or the customer's surname. However, this requirement does not apply to mail, telephone or internet applications, although the loan officer is required to ask for this information (Federal Financial Institutions Examination Council, 2013, Appendix B).
geographic location, and others on a combination of surname and geographic location or surname and first name.

Surname-based methods typically infer race and ethnicity by matching the relevant surnames with well-established dictionaries of Hispanic or Asian surnames (e.g., Perkins, 1993; Lauderdale and Kestenbaum, 2000) or comprehensive list of surnames and their associated race and ethnicity prevalences\(^3\) compiled by national statistical authorities. For example, the more recent US-based studies have often used a surname list released by the U.S. Census Bureau in 2007 based on the Decennial Census 2000 and referred to hereafter as the census surname list.

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 and 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 ethno-racial 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 surnames are relatively non-distinctive: and the latter has little ability to identify Hispanics or NH Asians because these groups are not spatially segregated.\(^4\) For this reason, hybrid approaches have been suggested that attempt to improve the accuracy of race and 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

---

\(^3\) By “prevalence” I refer to the prevalence of one group over the other, as measured by the percentage of people with a given surname that belong to a specific ethno-racial group.

\(^4\) See, for example, Fiscella and Freemont (2006) and Elliott et al. (2009) for a more detailed description of the advantages and drawbacks of each approach.
maximum functions in Coldman et al. (1988) or naïve 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 ethno-racial category using the census surname list, and then applies the naïve form of the 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) performed evaluations of BISG using health plan enrollment data and mortgage data, and found that it predicts race and ethnicity, especially NH Black and NH White, more accurately than the surname-only and geography-only approaches. However, the method has also been criticized by Baines and Courchane (2014) that, despite its superiority relative to other alternatives, it is still subject to significant bias and estimation error, and results in overstated disparities in mortgage lending outcomes. Additionally, Baines and Courchane (2014) raised concerns about the sensitivity of the BISG performance with respect to the race and ethnicity distribution in the test population.

First name information has been underutilized in prior research on race and ethnicity proxies, largely due to the lack of comprehensive tabulations covering a wide range of first names and their associated race and ethnicity prevalences.² Tzioumis (2015) 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 and ethnicity, henceforth the first name list. In this paper, I use this first name list to improve on the existing hybrid approaches. Specifically, I

---
² 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 been yet validated on external, individual-level data.
use this additional information to enhance the naïve Bayes classifier proposed by Elliott et al. (2009), 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 to three conditional attributes—a naïve Bayesian updating formula which updates the surname-based probabilities of membership in one of six ethno-racial categories with the first name and geographic location proportions for these same six groups. Given that first name demographic information is a good predictor of actual race and ethnicity as documented in Tzioumis (2015), 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.6

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 and ethnicity estimates that they produce match the self-reported race and ethnicity for the same individuals. The evaluation is based on a large mortgage lending dataset comprising information from several large lenders with national reach.

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

To preview the main results, I find that, consistent with expectations, the new approach offers improvements, in terms of accuracy and coverage, over BISG for all major ethno-racial

---

6 Geographic information is almost always available, so the vast majority of missing values occur for either first name or surname.
categories. The improvements are most substantial for the group for which BISG is least accurate—NH Blacks. Moreover, the superiority of BIFSG could become even stronger as new, more comprehensive first name lists become publicly available. Additionally, I find that when estimating disparities in mortgage pricing and underwriting decisions among the various ethnoracial groups with reasonably well-specified regression models, the disparity estimates based on either BIFSG or BISG proxies are remarkably close to those based on actual race and ethnicity. This finding should significantly alleviate related concerns raised in Baines and Courchane (2014). Finally, the evaluation shows that using discrete classification schemes based on either the BIFSG or BISG probabilities may result in smaller biases in the estimation of disparities in mortgage lending outcomes than when using continuous probabilities. The results of the BIFSG application offer evidence that the incidence of missing race and ethnicity information is generally higher among minorities than among NH Whites, with NH Blacks being the most affected group.

The remainder of the paper 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 and ethnicity information in the HMDA data. Section 5 concludes the paper.

2. Method

The new method to estimate race and 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 methods to evaluate the accuracy of the new approach compared to BISG.

2.1 Data

2.1.1 Surname Data

For surname demographic information, I use the census surname list. This list includes all surnames occurring 100 or more times in the Census 2000 (over 150,000 surnames, covering about 90 percent of the US population), along with demographic information showing the percentage of people with a given surname that belongs to one of six groups: Hispanic, NH Black, NH White, NH Asian/Pacific Islander (API), NH American Indian/Alaskan Native (AIAN), and NH Multiracial.

2.1.2 First Name Data

This paper is the first to use the newly available list of first names compiled by Tzioumis (2015). This list draws information from a large pool of recent mortgage application data, and includes 4,250 first names with information on their respective counts and proportions across six mutually exclusive and collectively exhaustive ethno-racial categories that are consistent with the categories used in the census surname list. For the great majority of cases, the first names’ demographic information is calculated using more than 30 observations. The coverage of this

---

7 This list is publicly available at [http://www.census.gov/genealogy/www/data/2000surnames/index.html](http://www.census.gov/genealogy/www/data/2000surnames/index.html).

8 For confidentiality purposes, the Census Bureau has suppressed exact counts for ethno-racial 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. Similarly 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 non-zero counts.

9 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 US population (e.g., Eleftherios, Slobodan, Tomislav, Xiaoping).
first name list is very similar to that of the census surname list; it captures approximately 86 percent of the US population, based on the 1990 Decennial Census information on first name frequencies. As noted in Tzioumis (2015), since this list is not based on census data for the entire US 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). Therefore, its use is particularly conducive in the context of fair lending.

2.1.3 Geo-demographic Data

I use census-block-group-level data on counts and proportions across the six aforementioned ethno-racial 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).10

2.1.4 Mortgage Data

In order to demonstrate the calculation of BIFSG proxies and assess their accuracy, I extract information from four distinct proprietary databases of mortgage transactions, each

---

10 Fiscella and Fremont (2006) provide a good comparison of the various geographic area definitions typically used in geocoding approaches. Specifically, zip codes generally encompass relatively large geographic areas, often including 30,000 or more people with widely varying socio-demographic profiles; census tracts are smaller, with an average population of about 4,200, and tend to be more homogenous although it is not unusual for the same tract to include both wealthy and poor neighborhoods; and census block groups average about 1,400 people and typically are very homogenous.
belonging to a different large lender. The information includes applicant first and last names, address, self-reported race and 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). One dataset contains 2012 application data, two contain 2013 data, and one contains 2014 data. I combine these four datasets, partly for data confidentiality purposes and partly to improve the representativeness of the test population. It is worth noting that the requirements for collection and reporting of information on race and ethnicity are consistent across the four datasets since the respective lenders comply with HMDA requirements. Additionally, HMDA race and ethnicity classifications allow exact replication of the six groups from the census surname list and the first name list.

2.2 Implementation of the BIFSG Algorithm

Constructing the BIFSG proxies for race and ethnicity requires six steps. These steps involve cleaning the names and geocoding the addresses from the test mortgage data; then merging the names and geocoded addresses from the test data with the two name lists (surnames and first names) and the census geo-demographic data, respectively; then constructing the

---

11 I obtained these data in the context of the OCC’s supervisory authority. According to definitions by the Office of the Comptroller of the Currency (OCC), a large lender is one with more than $25 billion in assets. Importantly, these data are different from those used by Tzioumis (2015) to develop the first name database. For confidentiality purposes, I cannot identify the lenders used in the analysis or present separate statistics by lender.

12 For dual applications of most lenders, while the mortgage data includes race and ethnicity for both applicant and co-applicant, only the applicant’s name and address are available.

13 The reported address is the address of the primary applicant for about 80 percent of the applications with non-missing race and ethnicity information, and is the address of the mortgaged property for the remaining of those applications. The substitution of property address for applicant address should not distort results of the evaluation given the small proportion of affected applications and that the ethno-racial composition of the neighborhood around the mortgaged property is very similar to that of the neighborhood of the applicant address (for example, in a random sample of 10,000 applications having both addresses available, the address-driven differences in the mean geography-based proportions used as inputs in the BIFSG are very small and statistically insignificant as indicated by paired t-tests, with the exception of NH API and NH Multiracial proportions. Even for these two exceptions, however, the relative (percentage) differences are small. Moreover, it is also not evident on theoretical grounds that one address is more reflective of the applicant’s race and ethnicity than the other.
probabilities that are inputs in the BIFSG formula; and, finally, apply the BIFSG formula to compute the proxies. These steps are described in more detail below.

1. Mortgage applications with missing values for race and ethnicity and dual applications for which race and ethnicity of the applicant and co-applicant are not identical are dropped from the test data. As a result, the sample size is reduced from 524,962 observations to 279,404.

2. Applicants’ names are standardized. Slightly different steps are taken depending on the specific format in which the names are recorded in each lender-specific dataset. Nonetheless, after identifying the name formats, the process typically consists of edits to convert all strings to uppercase, remove punctuation and special characters, titles, and suffixes (e.g., III, IV, Sr., Jr., Dr.), remove blanks between common last name prefixes and last names (e.g., VAN HALLEN becomes VANHALLEN), set one-letter names as missing, or if applicable, replace one-letter first names with middle names, and parse compound and hyphenated first and last names into separate variables.

3. Applicants’ standardized surnames and first names from the test data are matched with the corresponding name lists. Following the CFPB (2014) approach, for applicants with compound first or last names, the matching is attempted first based on the first word of the compound name, and if the first word fails to find a match, the second word is then tried. For one-word first names, if the matching is not successful, the middle name is then tried (if available). As shown in Table 1, the match rates for first names and surnames are very similar and quite high – about 88 percent.

---

14 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 (primary) applicant, it is possible that this is not always the case.

15 I also verified that dropping applications with missing race/ethnicity also eliminated applicant names corresponding to companies and trusts (e.g., those that include strings such as “Inc.”, LLC”, or “Trust”).

16 A more detailed description of the name cleaning process is available upon request from the author.
4. Addresses are assigned to specific geographic areas—census block groups—through geocoding of the exact address, using the Arc GIS mapping software.\(^{17}\) Then, the adult population distribution by race and ethnicity for the geographic area of a given address as well as at the national level is obtained from the decennial census. If the address cannot be geocoded or the total adult population in the geography is zero, the geo-demographic information is considered missing. As shown in Table 1, we are able to assign geo-demographic information to 96 percent of the applications.

5. 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 test data that matches the census surname list, the probability that a person belongs to a given racial or 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 and ethnicity, which is readily available in the census 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 test data that matches the first name list, the probability of a person’s having that first name, given the person’s race or ethnicity is approximated by the proportion of the population of the given race and ethnicity who bear the respective first name. This proportion is derived from the full

\(^{17}\) To ensure match accuracy, the ArcGIS matching score parameter was set to 0.80.
sample used by Tzioumis (2015) to develop the first name list.\footnote{Specifically, the denominator of the proportion is derived from the size of the full sample used by Tzioumis (2015), including all mortgage applicants with valid first names regardless of how common the name is—not just the sub-set 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, along with the list of most common first names.} 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 geocoded address with non-missing geo-demographic information, I compute the proportion of the U.S. adult population for each race and ethnicity residing in the geographic area in which that address is located (e.g., the proportion of all Hispanics in the US that reside in the census block group associated with the given address). For applications with missing geo-demographic information, the geography-based probabilities are considered missing.

6. The BIFSG algorithm is built based on a naïve Bayesian updating formula which updates the surname-based probabilities computed in step (5a) with the first-name-based and geography-based probabilities from steps (5b) and (5c), 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:

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

where: \(p(r|s, f, g)\) is the updated (posterior) probability of being of race and ethnicity \(r\), given surname \(s\), first name \(f\), and geographic area \(g\); \(p(r|s)\) is the probability that a person is of race and ethnicity \(r\), given that the person has surname \(s\), as computed in step (5a); \(p(f|r)\) is the
probability that a person has first name $f$, given that the person is of race and ethnicity $r$, as computed in step (5b); $p(g|r)$ is the probability that a person resides in geographic area $g$, given that the person is of race and ethnicity $r$, as computed in step (5c); and the summation in the denominator occurs over the six race and ethnicity categories defined previously.\footnote{The Bayesian updating formula can be expressed using different permutations of the three attributes, e.g., $p(r|f) \cdot p(s|r) \cdot p(g|r)$. In preliminary work, I experimented with all possible permutations and the performance was very similar across alternatives, with the one presented in the paper resulting in marginally better overall performance.} Details on the derivation of this formula are provided in the appendix.

The statistical validity of the BIFSG formula relies on the assumption of conditional independence among surnames, first names, and geographic areas, i.e., $p(g|r,s) = p(g|r)$ and $p(f|r,s,g) = p(f|r)$. In other words, this assumption implies that 1) the probability of residing in a given geographic area, given person’s race and ethnicity, does not vary by surname and 2) the probability of having a given first name, given person’s race and ethnicity, does not vary by surname or geographic area. These assumptions cannot be tested with the available data, due to very small sample sizes for the various combinations of name, geography, race and ethnicity that are required for testing.\footnote{For the same reason, it is also not feasible to relax these assumptions in the Bayesian formula.} However, Domingos and Pazzani (1997) show that the naïve 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 non-missing. 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 non-missing values, and using surname-only (SO), first-name-only (FO), or geography-only (GO) probabilities if a single attribute has non-missing values. For brevity sake, I do not cover the performance of the various algorithm extensions or related weighting schemes in this paper. However, it is worth noticing that if BIFSG is more accurate than BISG when all attributes are non-missing, the BIFSG will also maintain its lead when “extended” with BISG if the first name is missing.

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

In the statistical analysis of variation in mortgage outcomes by race and ethnicity, BIFSG probabilities can be used to proxy for race and ethnicity either directly (i.e., using the predicted probabilities per se), or indirectly, (i.e., creating a binary measure based on a threshold rule). Either one of these two approaches has its own strengths and weaknesses, and it is not obvious that one is superior to the other, despite some research suggesting that using continuous probabilities should be the preferred choice (e.g., Elliott et al., 2008, CFPB, 2014). For example, McCaffrey and Elliott (2008) show that the loss of efficiency from modelling with discrete

---

21 The BISG-like formula takes on one of the following forms, depending on which two attributes are non-missing: $p(r|s,g) = \frac{p(r|s) \cdot p(g|r)}{\sum_{f=1}^{p(r|s)} p(g|r)}$; $p(r,s,f) = \frac{p(r|s) \cdot p(f|r)}{\sum_{g=1}^{p(r|s)} p(f|r)}$; $(r|f,g) = \frac{p(r|f) \cdot p(g|r)}{\sum_{s=1}^{p(r|f)} p(g|r)}$. Again, the Bayesian updating formula can be expressed using different permutations of the two attributes, e.g., $p(r|g) \cdot p(f|r)$ instead of $p(r|f) \cdot p(g|r)$. In preliminary work, I experimented with all possible permutations and the performance was very similar across alternatives, with the ones shown above resulting in marginally better overall performance.

22 The SO probability is computed like in step (5a) above; the FO probability is given by the proportion of all people with a specific first name who report being of a given race or 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 or ethnicity.

23 To the extent that methods which 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 which are assigned proxies based on one or two attributes.

24 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 and ethnicity imputed based on BISG.

25 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 non-missing attribute) is more accurate than a similarly extended BISG. Selected results from this preliminary research are available upon request from the author.
classifications is larger than from modelling with continuous probabilities, however, 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 and ethnicity) is an important area of further research. For another example, it is well known that the use of probabilities for classification purposes may result in biased estimates because of classification errors, however, so does modeling with probabilities if the probabilities are biased. The latter is a distinct possibility in the context of the naïve 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). Additionally, Baines and Courchane (2014) shows significant biases in APR disparity estimates from the direct use of BISG probabilities in regression models.

Statistical considerations aside, discrete classifications have the advantage of allowing identification of individuals of a specific race and ethnicity, which can be very valuable in certain areas. For example, in the context of fair lending, comparative file reviews used for monitoring purposes and restitution schemes used to settle discrimination lawsuits would be difficult to accomplish without classification.

2.4 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 and 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 and ethnicity estimates that they produce match the self-reported race and ethnicity for the same individuals. Given that proxies can be defined either using directly the BISFG and BISG probabilities 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 and 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 and ethnicity classes; and 3) by evaluating the biases that proxies may cause in estimating disparities in 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 three sets of metrics: the Pearson correlation between the proxy probability and the self-reported race and ethnicity, which measures the extent to which applicants of a given race and ethnicity are assigned higher proxy probabilities of belonging to that race and ethnicity; distribution of the difference between the BIFSG and BISG probabilities
for each ethno-racial group; and a set of accuracy metrics—false negative rate and false positive rate, along with the underlying contingency matrix—which together measure the diagnostic power of discrete classification schemes based on the proxy probabilities.\textsuperscript{26} Related to the last set of metrics, false negatives include individuals that self-report belonging to a particular group, but whom the proxy method categorize into another group; and false positives are cases where the proxy method assigns an individual to a particular group, when in fact she belongs 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 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.\textsuperscript{27,28} The underlying contingency matrix includes the number of false positives, false negatives, true positives, and true negatives for each ethno-racial group, as proportions of the total sample size. Individuals whom the proxy method cannot assign to any

\textsuperscript{26} In preliminary research, I also calculated the Area Under the Receiver Operating Characteristic Curve (AUC), as another measure of the ability of the proxy probability to accurately sort applicants into the self-reported ethno-racial categories (i.e., the likelihood that the proxy probability of being in a given ethno-racial category, e.g., NH Black, will rank a randomly chosen NH Black applicant higher than a randomly chosen applicant of different race and ethnicity). I found that 1) the AUC statistics for all four major ethno-racial 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 ethno-racial group. The AUC statistics are available upon request. I chose not to include them in the paper because my data is highly skewed towards NH Whites (see Table 2), 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 (see, for example, Davis and Goadrich, 2006).

\textsuperscript{27} The false negative rate is equal to 1 minus the true positive rate (or 1 minus recall), where true positives are cases where the proxy method assigns an individual to the group she actually belongs to; and the false positive rate, as computed here, is equal to 1 minus precision.

\textsuperscript{28} 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 imbalanced datasets, it may lead to erroneous conclusions and thus it is not appropriate to use (e.g., see Lin and Chen, 2012). For example, if there were 95 NH Whites, two NH Blacks and three Hispanics, the classifier could easily be biased into classifying everybody as NH White. The accuracy rate would then be 95 percent for NH Whites (and overall, for all classes combined), 98 percent for NH Blacks and 97 percent for Hispanics, but in practice the classifier would have a 100 percent recognition rate for NH Whites and a 0 percent recognition rate for each minority class.
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.

To enhance 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 and ethnicity and the self-reported attributes, and correlation coefficients—across all ethno-racial 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 and ethnicity across applicants and in predicting individual race and ethnicity.29

To ensure that the comparative accuracy of BIFSG and BISG is not driven by the underlying sample, the evaluation samples exclude any observations that have missing values for either set of proxies.30 As can be inferred from Table 1, the rate of missing values is somewhat higher 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.31 However, the somewhat lower coverage of BIFSG is not concerning since, as previously

---

29 The relative efficiency in matching the actual ethno-racial distribution is computed as the ratio of average squared deviations from the actual distribution of race and ethnicity for BIFSG and BISG; and the relative efficiency in predicting individual race and 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 to BISG—or, equivalently, that method BIFSG is 100 percent more efficient than BISG—means that the accuracy of an analysis testing for differences in a given outcome among ethno-racial 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.

30 In other words, the evaluation samples include only applicants with demographic information for all three attributes (surname, first name, and geocoded address).

31 Specifically, as shown in Table 1, applications for which both the surname and geography can be matched with the relevant demographic data represent 84 percent of the total sample, and those for which first name, surname, and geography can all be matched with the demographic data represent 76 percent of the total. Thus, the BIFSG coverage is about 8 percentage points lower than the BISG coverage.
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 and ethnicity composition of the evaluation sample compared to that for the whole U.S. from the Decennial Census 2010. The sample includes 211,570 applications (nearly 76 percent of all applications with valid race and ethnicity), a sample size which is comparable to those in CFPB (2014) and Baines and Courchane (2014). Although the test 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 to 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. However, this is not of significant concern because these two groups account for less than 1 percent of the application population (see Table 2).

3.1 Distribution of Lending by Race and Ethnicity

Table 3 shows the distribution of mortgage applications by self-reported race and 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.

---

32 The results for these two groups are available upon request from the author. Although these results are not shown in the paper, they are included in the computation of the weighted average statistics.
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 NH Blacks, NH Whites, and NH APIs, and with a weighted average prevalence error (deviation from self-reported) which is 39 percent lower than that of BISG.\textsuperscript{33} Using the efficiency measure described in section 2.4, BIFSG is almost 170 percent 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 Elliott et al., 2008) relative to the GO method.\textsuperscript{34}

3.2 Predicting Individual Race and Ethnicity

3.2.1 Correlations between the Proxy Probability and Self-reported Race and Ethnicity

Table 4 displays the correlations between the self-reported race and ethnicity and the proxy probabilities generated by BIFSG and BISG. All differences in correlation coefficients between the two methods are statistically significant at the 5 percent level.\textsuperscript{35} The correlation between BIFSG probabilities and self-reported race and ethnicity ranges from 0.748 to 0.885, with a weighted average across all categories of 0.838. The correlation coefficients for the BISG method are all smaller than their BIFSG counterparts, ranging from 0.712 to 0.871, with a weighted average across all categories of 0.817. Thus, BIFSG results in 2.5 percent higher

\textsuperscript{33} The absolute difference in deviations between the two methods is 0.6 percentage points for NH Blacks and 1.3 percentage points for NH Whites, compared to only 0.1 percentage points for NH APIs and zero percentage points for Hispanics. The relative difference in deviations between the two methods is similarly high for NH Blacks (44 percent), NH Whites (40 percent), and NH APIs (46 percent). Differences in prevalences between the two methods for NH Blacks, NH Whites and NH APIs are statistically significant at the 5 percent level, as indicated by both pairwise t-tests and non-parametric sign tests.

\textsuperscript{34} Specifically, Elliott et al. (2008) find that BSG has 20 percent lower average deviation from self-reported and it is 56 percent more efficient than GO.

\textsuperscript{35} Statistical significance is determined by comparing the 95 percent confidence intervals for Fisher's Z transformation of the correlation coefficients (see Fisher, 1970, p. 199).
average correlation and 5.1 percent higher overall efficiency than BISG. The improvements associated with BIFSG are largest for NH Blacks (5 percent higher correlation and 10.3 percent higher efficiency) and NH Whites (2.6 percent higher correlation and 5.3 percent higher efficiency).

3.2.2 Distribution of the Difference between the BIFSG and BISG Probabilities

To gain further insights into the sorting quality of the BIFSG and BISG probabilities—in particular, the effect from including the first name information into the probability calculation—, I investigate the distribution of the difference between the two probabilities for each of the four major ethno-racial groups. For example, for each applicant whose self-reported race and 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.36

Before reviewing the statistics, it is helpful to go over an illustrative example. Consider a hypothetical NH Black applicant, Latoya Davis, living in a predominantly NH White census block group. As illustrated in Table 5, this applicant has a very large difference between the BIFSG and BISG probabilities: the BIFSG probability of being NH Black is 98.5 percent, whereas the corresponding BISG probability is only 15 percent. The reason for this large difference is the additional information that the first name contributes to the probability calculation under the BIFSG method. Specifically, while the surname, Davis, is predominantly white, with 64.7 percent of people with that surname being white and only 30.8 percent being black, the first name, Latoya, is predominantly black, with 91.4 percent of people with that first name being black.

36 A positive difference means that the BIFSG probability is the larger of the two.
name being black. 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.

Figure 1 shows the distribution of the probability differences for each ethno-racial 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 percent of the applicants in this group have differences between the two proxy probabilities larger than 10 percentage points in absolute value. The Hispanic and NH API groups also have significant proportions of applicants with probability differences in this range—23 percent and 21.4 percent, respectively. For NH Whites, a smaller proportion (12.7 percent) 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 (20,254).

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, 23.1 percent of the NH Black applicants have a BIFSG probability of being NH Black which is larger than the corresponding BISG probability by more than 10 percentage points, but only 11.5 percent of the NH Blacks are assigned a BIFSG probability which is smaller than the BISG probability by more than 10 percentage points. Also, 17.6 percent of the Hispanic applicants are in the right tail compared to only 5.4 percent in the left tail, and 9.5 percent the NH White

---

37 BIFSG does not use the intra-first-name population shares (rather, it uses the share of the U.S. population of each ethno-racial group bearing the specific first name, shown in Table 5), however, I use them here (e.g., 91.4 percent) to better illustrate the point.
applicants are in the right tail whereas only 3.2 percent 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, and, therefore, BIFSG produces probabilities which, on average, are closer aligned with the actual race and ethnicity of an individual. This provides further evidence of the higher predictive ability of the BIFSG probabilities relative to the BISG probabilities.

3.2.3 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 true positives, false negatives and false positives (not including unclassified applicants), and the coverage is computed as the proportion of applicants of a given race and ethnicity that cannot be classified (the higher the proportion of unclassified applicants, the lower the coverage).

The evaluation is performed for two classification schemes that have been suggested in related research. One scheme, which is commonly used for Naïve Bayesian classifiers and is known as maximum a posteriori (MAP), classifies the application in the ethno-racial category corresponding to the largest proxy probability.38 The other scheme classifies the application in a given ethno-racial category if the proxy probability for that category is larger than a specified threshold, where the threshold values are set at 80 percent. Notably, MAP classifies all applicants, and thus has maximum coverage, whereas the threshold-based scheme does not classify applicants for which all six proxy probabilities are below the specified threshold. Nevertheless, as the discriminatory power of a classification methodology increases, the portion not classified by the threshold-based scheme decreases.

38 See, for example, Dai and Su (2014) for an application of the MAP classifier.
Table 6 reports the evaluation results. First, it is worth noting that the BIFSG proxies have similar or higher accuracy compared to BISG across classification schemes and ethno-racial categories. The accuracy improvements associated with BIFSG are somewhat larger when using MAP than when using the 80 percent threshold scheme, with the largest improvements occurring for NH Blacks based on MAP: a 3.7 percentage point decrease in the rate of false negatives, and a 4 percentage point decrease in the rate of false positives. NH Blacks also rank first in terms of the BIFSG-related improvements in the false negative rate under the 80 percent threshold classification. Another important finding is that, for the threshold-based scheme, BIFSG has significantly better coverage than BISG both overall and for each specific ethno-racial category. Specifically, the overall proportion of unclassified applicants is 10.8 percent for BIFSG and 14.7 percent 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 2.1 to 6.5 percentage points, with the largest improvement occurring for NH Blacks.\footnote{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 12.5 percent to 36.8 percent across groups, and amounting to nearly 27 percent overall.}

3.3 Estimating Disparities in Mortgage Lending Outcomes

To evaluate the biases that proxies may cause in estimating disparities in mortgage lending outcomes among ethno-racial groups, I use mortgage pricing and underwriting decisions as examples of outcomes. Specifically, I regress APR and a denied/approved indicator on race and ethnicity, first using actual race and ethnicity, and then using the BIFSG and BISG proxies—both as continuous probabilities and as discrete classifications. Unlike Baines and Courchane (2014) who estimate raw disparities, I estimate adjusted disparities using regression
models that control for important mortgage pricing and underwriting factors. The reason for this approach is that adjusted disparities are much more relevant than raw disparities for practical applications such as regulatory examinations of fair lending practices, where a wealth of lending data is typically available and used in statistical analyses to account for lending policies.\(^\text{40}\)

The APR regressions are estimated using OLS on a sample which includes only originated loans, whereas the denied/approved decision is modeled using logistic regression on a sample which includes originated loans as well as approved but not accepted and denied applications. Ethno-racial disparities in lending outcomes for a given proxy method are estimated relative to NH Whites, which are the omitted category and serve as a baseline for the comparisons. 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), a number of 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),\(^\text{41}\) the year-quarter when the rate was locked, and the state of property location, as well as a set of dummy variables that capture 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 the following explanatory variables: FICO score, combined loan-to-value-ratio, debt-to-income ratio, loan amount (in logarithmic form), and the

\(^{40}\) Biases in lending outcome disparities 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 (see, for example, Bound et al., 2000, and Ashenfelter and Card, 1999, p. 1291 and p. 1341). Additionally, raw disparities are plagued by omitted variable biases. For these reasons, a regression which includes controls for relevant determinants of lending outcomes is much more informative regarding biases from the proxies’ errors than a regression without such controls.

\(^{41}\) 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.
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.42

When assessing biases associated with continuous probabilities and the MAP classification, all regressions are run on the sample for which both BIFSG and BISG proxies can be computed. When evaluating biases for the 80 percent threshold classification scheme, the regression based on actual race and ethnicity is estimated on the sample for which the continuous proxy probabilities can be computed (same sample as above), whereas the regressions based on the proxy classifiers are run on the smaller samples of applications that can be classified. This approach ensures that I measure the full bias associated with this proxy classifier, which may be partly due to misclassification and partly due to incomplete coverage.

The relevant regression results are shown in Table 7. There are two noteworthy patterns that can be observed in this table. First, using discrete classification results in smaller biases in both the APR disparities and denial odds ratios than using continuous probabilities, with the MAP scheme generally producing the smallest biases. As an example, the BIFSG-induced bias in the denial odds ratio for NH Blacks is 0.170 when using continuous probabilities, but only 0.011 when using MAP. These findings are consistent with the Domingos and Pazzani (1997) analysis which shows that the naïve Bayesian classifier performs well in terms of classification accuracy but not necessarily in terms of the accuracy of estimated probabilities. They also

---

42 However, the extent of missing data is small – 5 percent of the observations in the underwriting sample and 11 percent of those in the pricing sample have missing values for at least one variable – and regression analyses which exclude observations with missing values produce similar results with those presented in the paper. These alternative results are available upon request from the author.
suggest that coverage-related biases for the 80 percent threshold scheme may outweigh the accuracy-related advantage that this method has over MAP, leading to larger overall biases for the threshold classification.\textsuperscript{43}

Second, and most importantly, the biases from both BIFSG and BISG proxies are generally very small—less than 3 basis points (bps) for APR disparities and, with few exceptions,\textsuperscript{44} less than 0.1 for denial odds ratios—and, consequently, differences in these biases between the two methods are trivial—up to 1.2 bps for APR disparities and less than 0.1 for denial odds ratios. It is also worth noting that in the cases with the largest bias differences between the two methods—1.2 bps for APR disparities for Hispanics, using continuous probabilities and the 80 percent threshold classification, and 0.081 for denial odds ratios for NH Blacks, using the 80 percent threshold classification—BIFSG produces the smaller bias.

In additional analyses which are available upon request, I also compute the biases from proxies relative to estimates based on actual race and ethnicity which are obtained from a sample that includes all observations for which actual race and ethnicity are available, and obtain results which are almost identical to those presented here. These additional results suggest that the loss of coverage from dropping observations for which BIFSG cannot be computed does not introduce significant biases in estimation.

\section*{4. Application: Imputation of Race and Ethnicity in HMDA Data}

This section demonstrates the application of BIFSG to the imputation of missing race and ethnicity information in the HMDA data. As mentioned in the Introduction, a significant

\textsuperscript{43} This is especially true for BIFSG where in all cases, the biases from the threshold classification are larger or equal to those from MAP.

\textsuperscript{44} The exceptions are the biases in denial odds ratios for NH Black based on continuous probabilities and the 80 percent threshold classification, which range from 0.169 to 0.260.
proportion of the mortgage applications in the HMDA data are missing race and/or ethnicity. For example, in HMDA 2014, about 22 percent of all applications and nearly 15 percent of the applications processed by the reporting institution (excluding purchased loans) are missing this information. Most empirical analyses using these data, including fair lending examinations, simply eliminate from the population applications that lack race and ethnicity information, under the assumption that this information is missing at random and therefore introduces no sample selection problems into the analysis. However, the only two studies that closely examined the trends in missing race and ethnicity in HMDA, Dietrich (2002) and Huck (2000), found that this information could be missing for systematic reasons and therefore may introduce bias and efficiency problems into fair lending exams. The use of race and ethnicity proxies offers an easy way to check and, if necessary, alleviate such problems, and therefore represents a worthwhile application of BIFSG. For example, a fair lending analysis for a financial institution could start with checking whether the incidence of missing information on race and ethnicity is correlated with the relevant lending outcomes (e.g., approval rates, pricing) and with the race and ethnicity imputed based on proxies. If these correlations are negligible, then the fair lending analysis can focus only on the sample with non-missing race and ethnicity. However, if these correlations are significant, then one should also include in the analysis the applications with imputed race and ethnicity information.

In this application, I randomly select 10,000 applications from the sample with available race and ethnicity used in the BIFSG evaluation, and 10,000 applications for which actual race and ethnicity are missing but name and geographic information is available (so that the BIFSG probabilities can be computed).\textsuperscript{45} Purchased loans are not included in the application because

\textsuperscript{45} 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”
they do not reflect the lender’s policies and thus are not used in fair lending exams. For each of the selected applications, I compute the BIFSG proxy probabilities and the associated MAP classifications, and then compare the ethno-racial prevalences based on these proxies in the sample for which actual race and 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.\(^46\) If such correlation exists, it will compound sample selection problems specific to the racial estimates which may arise in fair lending analyses if the incidence of missing race and ethnicity is also correlated with the relevant lending outcomes.

The application results are summarized in Table 8. First, notice that in the sample with non-missing race and ethnicity, prevalences based on either probabilities or MAP classifications are very similar to those based on actual race and ethnicity—a result which 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 and ethnicity being lower among NH Whites than among minorities. For example, according to the probability-based prevalences, the proportion of NH Whites is 71.7 percent in the sample with non-missing race and ethnicity and 69.7 percent in the sample with missing race and ethnicity. Similarly, the MAP-based prevalences indicate that NH Whites account for 74.6 percent of the applicants who report their race and ethnicity and for 72.7 percent of those with missing race and ethnicity. Differences in

\(^{46}\) While Huck (2000) and Dietrich (2002) also analyzed this correlation, they imputed missing race and ethnicity largely based on applicants’ geographic location—an approach which has been shown to produce much less accurate predictions than the Bayesian methods discussed in this paper.
prevalences between the two samples are statistically significant at the 5 percent level for all groups. Among minorities, NH Blacks have the highest incidence of missing race and ethnicity, as indicated by the relatively large difference between their prevalences in the two samples (e.g., 9 percent in the sample with missing race and ethnicity compared to 6.8 percent in the sample which is not missing this information, when using probability-based prevalences). Interestingly, although minorities overall have higher rates of missing race and ethnicity than NH Whites, Hispanics are less likely to miss this information than NH Whites.

5. Conclusions

Previous indirect methods to estimate race and ethnicity have underutilized first name information due to the lack of comprehensive lists of first names and their associated race and ethnicity distributions. In this paper, I propose an enhanced Bayesian method—the Bayesian First Name Surname Geocoding method (BIFSG)—for predicting race and ethnicity, using a new first name list developed by Tzioumis (2015). The new method improves on the existing BISG naïve-Bayesian algorithm by considering first name information, along with surname and geographic information. Using mortgage lending data and applying a wide array of metrics to evaluate the performance of the new method, I demonstrate that BIFSG results in non-trivial

47 The statistical significance of the difference in the probability-based prevalence of a given group (i.e., the mean probability of being in a given group across all sampled applicants) between the two samples 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 non-parametric test, and I find that, indeed, this is the case for each group. The statistical significance of the difference in the MAP-based prevalence of a given group between the two samples is determined using chi-square tests.

48 More specifically, this ranking can be inferred from the ratio of the prevalence of a group in the sample with missing race and ethnicity to that in the sample with non-missing race and ethnicity. For given sample sizes, the higher that ratio, the higher the incidence of missing information, and NH Blacks have the largest ratio among all groups. For example, according to the probability-based prevalences, the aforementioned ratio for NH Blacks is 1.32 (9 percent divided by 6.8 percent), whereas for the other groups it varies between 0.92 and 1.1.

49 For example, the aforementioned ratio of probability-based prevalences is 0.92 for Hispanics and 0.97 for NH Whites.
accuracy and coverage improvements over BISG. These improvements hold both for the continuous probabilities and the discrete classification schemes across ethno-racial groups. Importantly, the largest improvements associated with BIFSG occur for NH Blacks—the group for which BISG is least accurate.

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 bias in APR disparities and denial odds ratios from both BIFSG and BISG proxies are very small, with the BIFSG generally having a smaller bias. Additionally, I find that using discrete classification schemes results in smaller biases in the estimation of disparities in mortgage lending outcomes than when using continuous probabilities, with the MAP classification scheme generally producing the smallest bias across classification schemes. This finding provides additional support for the use of discrete classifications in the fair lending area, beyond the advantage that such schemes have over the continuous probability approach in terms of allowing identification of individuals of a specific race and ethnicity.

Following evaluation, I demonstrate the application of BIFSG to the imputation of missing race and ethnicity information in the HMDA data, and in the process, provide novel evidence that race and ethnicity are somewhat correlated with the incidence of missing race/ethnicity information. Specifically, I find that the incidence of missing race and ethnicity information is somewhat lower among NH Whites than among minorities, and that NH Blacks are the group most likely to miss this information.

Given its improvements over the previously most advanced method to estimate race and ethnicity, BIFSG represents an alternative worthwhile to consider 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 list) become available.

References

Ashenfelter, O. and D. Card. 1999. *Handbook of Labor Economics*, Volume 3, Part 1, Elsevier.

Baines, A. P. and M.J. Courchane. 2014. “Fair lending: Implications for the Indirect Auto Finance Market.” Study prepared for the American Financial Services Association.

Bound, J., C. Brown, and N. Mathiowetz. 2000: “Measurement Error in Survey Data,” Institute for Social Research, University of Michigan.

Coldman, A. J., T. Braun and R.P. Gallagher. 1988. “The classification of ethnic status using name information.” *Journal of Epidemiology and Community Health*, 42, 390-395.

Consumer Financial Protection Bureau. 2014. “Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment.” Washington, DC. Available online at http://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf

Dai, Yugang and H. Su. 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.

Davis, J. and M. Goadrich. 2006. “The relationship between precision-recall and ROC curves.” In *Proceedings of the 23rd International Conference on Machine Learning*, p. 233-240, New York, NY, USA.

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.

Domingos, P., and M. Pazzani. 1997. “Beyond independence: Conditions for the optimality of the simple Bayesian classifier.” *Machine Learning* 29:103–130.

Elliott, M., A. Fremont, P. Morrison, P. Pantoja, and N. Lurie. 2008. “A new method for estimating of race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity.” *Health Services Research*, 43(5p1), 1722-1736.

Elliott, M., P. Morrison, A. Fremont, D. McCaffrey, P. Pantoja, and N. Lurie. 2009. “Using the Census Bureau's surname list to improve estimates of race and ethnicity and associated disparities.” *Health Services and Outcomes Research Methodology*, 9(2), 69-83.

Federal Financial Institutions Examination Council, 2013. “A Guide to HMDA Reporting: Getting It Right!” Washington, D.C.: Federal Financial Institutions Examination Council.

Fiscella, K. and A. Freemont. 2006. “Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity.” *Health Services Research*, 41(4), Part I, 1482–1500.

Fisher, R.A. 1970. *Statistical Methods for Research Workers*. 14th ed. New York: Hafner.

Fremont, A. and Lurie, N. 2004. *The Role of Race and Ethnic Data Collection in Eliminating Health Disparities*. National Academies Press, Washington, DC.

Huck, P. 2000. “Home Mortgage Lending by Applicant Race/Ethnicity: Do HMDA Figures Provide a Distorted Picture.” *Consumer Issues Research Series-2000-3*. Chicago: Federal Reserve Bank of Chicago.

Krieger, N., J.T. Chen, P.D. Waterman, M.J. Soobader, S.V. Subramanian, and R. Carson. 2002. “Geocoding and Monitoring of US 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.

Lauderdale, D., and B. B. Kestenbaum. 2000. “Asian American Ethnic Identification by Surname.” Population and Development Review 19 (3): 283–300.

Lin W.J., and Chen J.J. 2013. Class-imbalanced classifiers for high-dimensional data.” Briefings in Bioinformatics 14(1):13-26

McCaffrey, D., and M. Elliott. 2007. “Power of Tests for a Dichotomous Independent Variable Measured with Error.” Health Services Research. DOI: 10.1111/j.1475-6773.2007.00810.x

Mateos, P. 2007. “An ontology of ethnicity based upon personal names: with implications of neighborhood profiling.” Unpublished Ph.D. Thesis, Department of Geography, University College London.

Perkins, R. C. 1993. “Evaluating the Passel-Word Spanish Surname List: 1990 Decennial Census Post Enumeration Survey Results.” U.S. Census Bureau, Population Division.

Tzioumis, K. 2015. Demographic aspects of first names. Unpublished manuscript. Available online at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2606163
Figure 1. Distribution of the Difference between the BIFSG and BISG Probabilities
Table 1. Match Rates for Names and Geography

| Match rate for... |        |
|-------------------|--------|
| first names       | 0.884  |
| surnames          | 0.879  |
| geography         | 0.957  |
| surname and...    | 0.842  |
| all three features| 0.757  |

Notes
1) Includes all applications with valid race/ethnicity from single applicants and dual applicants with same race/ethnicity.
2) Matches on geography exclude applications which could be geocoded but for which the census block group had no population.

Table 2: Reported Race and Ethnicity Composition

| Data Source (Census 2010) | Hispanic | NH Black | NH White | NH API | NH AIAN | NH Multiracial | N obs |
|---------------------------|----------|----------|----------|--------|---------|----------------|-------|
| National average          | 11.1%    | 11.3%    | 70.5%    | 7.0%   | 0.9%    | 0.8%           |       |
| Test Sample1              | 11.7%    | 6.6%     | 75.3%    | 5.9%   | 0.2%    | 0.4%           | 211,570|

Notes:
1) Includes all applications with non-missing name and geographic demographic information (for which both BIFSG and BISG proxies can be computed)

Table 3: Distribution of Loans by Race and Ethnicity: Proxies vs. Self-Report

| Method       | Hispanic | NH Black | NH White | NH API | Weighted Average | % Diff. in Average Deviation from Self-Report | Relative Efficiency of BIFSG vs. BISG |
|--------------|----------|----------|----------|--------|------------------|---------------------------------------------|----------------------------------------|
| Self-Report  | 11.7%    | 6.6%     | 75.3%    | 5.9%   | (0)              | -39.0%                                      | 168.6%                                |
| BIFSG        | 12.0%    | 7.3%     | 73.4%    | 5.7%   | 1.5%             |                                             |                                        |
| BISG         | 12.0%    | 7.9%     | 72.2%    | 6.2%   | 2.5%             |                                             |                                        |

Notes:
Differences in prevalences between BIFSG and BISG for NH Blacks, NH Whites and NH APIs are statistically significant at the 5 percent level, as indicated by both pairwise t-tests and non-parametric sign tests.

Table 4. Correlation between Proxy Probability and Self-Reported Race and Ethnicity

| Method | Correlation Coefficients | % Diff. in Average Correlation BIFSG - BISG | Relative Efficiency of BIFSG vs. BISG |
|--------|--------------------------|-------------------------------------------|----------------------------------------|
| BIFSG  | 0.885 0.748 0.841 0.874 0.838 |                                           |                                        |
| BISG   | 0.871 0.712 0.820 0.866 0.817 | 2.5%                                     | 5.1%                                   |

Notes:
Differences between the BIFSG and BISG correlations are statistically significant at the 5% level for all racial/ethnic groups.
Table 5. BIFSG and BISG Calculation Example

| Attributes | Hispanic | NH Black | NH White | NH API | NH AIAN | NH Multiracial |
|------------|---------|----------|----------|-------|--------|--------------|
| Surname: DAVIS | 1.6% | 30.8% | 64.7% | 0.4% | 0.8% | 1.7% |
| First name: LATOYA | 0.0022% | 0.0759% | 0.0002% | 0.0000% | 0.0000% | 0.0000% |
| Geography: Census Block Group X | 0.0006% | 0.0004% | 0.0011% | 0.0002% | 0.0001% | 0.0009% |
| BIFSG probability | 0.2% | 98.5% | 1.3% | 0.0% | 0.0% | 0.0% |
| BISG probability | 1.1% | 15.0% | 82.0% | 0.1% | 0.1% | 1.8% |

Table 6. Alternative Classification Methods and their Performance

| Classification Method | Proxy Method | Race/ Ethnicity | Total Applications in Group | Contingency Matrix (% of total classified applications) | % Unclassified | False Negative Rate | False Positive Rate |
|-----------------------|--------------|-----------------|----------------------------|-------------------------------------------------------|--------------|--------------------|--------------------|
|                       |              |                 |                            | Proxy=yes | Proxy=no | Proxy=yes | Proxy=no | Actual=yes | Actual=no | Actual=yes | Actual=no |
|                       |              |                 |                            | 10.4 | 86.8 | 1.5 | 1.3 | 0.0 | 10.7 | 12.6 |
|                       |              |                 |                            | 10.2 | 86.7 | 1.6 | 1.4 | 0.0 | 12.2 | 13.6 |
|                       |              | Hispanic        | 24685                      | 4.4 | 92.0 | 1.5 | 2.1 | 0.0 | 32.2 | 25.0 |
|                       |              | NH Black        | 13859                      | 4.2 | 91.7 | 1.7 | 2.3 | 0.0 | 35.8 | 29.0 |
|                       |              | NH White        | 159357                     | 72.4 | 20.7 | 3.9 | 2.9 | 0.0 | 3.9 | 5.2 |
|                       |              | NH White        | 159357                     | 72.0 | 20.3 | 4.4 | 3.3 | 0.0 | 4.4 | 5.7 |
|                       |              | NH API          | 12441                      | 5.0 | 93.5 | 0.7 | 0.9 | 0.0 | 15.1 | 11.5 |
|                       |              | NH API          | 12441                      | 5.0 | 93.3 | 0.8 | 0.9 | 0.0 | 15.1 | 13.7 |
|                       |              |                 |                            | 11.0 | 87.1 | 1.0 | 0.9 | 9.2 | 7.3 | 8.5 |
|                       |              |                 |                            | 10.7 | 87.3 | 1.0 | 1.0 | 14.6 | 8.4 | 8.8 |
|                       |              | Hispanic        | 24685                      | 3.6 | 94.8 | 0.5 | 1.1 | 35.9 | 23.2 | 11.3 |
|                       |              | NH Black        | 13859                      | 3.3 | 95.2 | 0.4 | 1.2 | 42.3 | 26.1 | 11.1 |
|                       |              | NH Black        | 13859                      | 75.9 | 20.2 | 2.5 | 1.5 | 8.4 | 1.9 | 3.1 |
|                       |              | NH White        | 159357                     | 76.2 | 19.7 | 2.6 | 1.5 | 12.0 | 1.9 | 3.3 |
|                       |              | NH White        | 159357                     | 5.1 | 93.9 | 0.5 | 0.5 | 14.8 | 9.7 | 8.4 |
|                       |              | NH API          | 12441                      | 5.2 | 93.8 | 0.5 | 0.5 | 16.9 | 9.2 | 8.7 |

Notes:
1) This is computed as share of total applications in the actual group. For the 80% classification, unclassified applications are excluded from total applications in the computation of the false negative rate.
2) This is computed as share of total applications with Proxy=yes (i.e., total applications which the proxy classifies in a given group)
Table 7. Comparison of Estimated Disparities in Pricing and Underwriting: Self-Report vs Proxies

A. Adjusted APR Disparities

| Method          | N obs | Hispanic       | NH Black       | NH API        |
|-----------------|-------|----------------|----------------|---------------|
|                 |       | Coef (bps) | Bias (bps) | Difference (bps) | Coef (bps) | Bias (bps) | Difference (bps) | Coef (bps) | Bias (bps) | Difference (bps) |
| self-report     | 122245 | 5.4      | 5.5        | -12.0     | 6.1        | 0.7         | -1.2            | 7.0        | 1.5         | -0.5            | -14.2       | 2.2         | 0.2            |
| BIFSG - prob    | 122245 | 7.3      | 1.9        | 5.4       | 5.2        | -0.2         | -0.2            | 7.5        | 2.0         | -14.0           | 0.0         | 2.0         | 0.2            |
| BIFSG - MAP     | 122245 | 4.9      | -0.5       | 0.4       | 5.1        | -0.4         | -12.2           | 6.0        | 0.5         | -0.6            | 13.8        | 1.8         | 0.8            |
| BIFSG - prob>=0.8 | 110016 | 5.7      | 0.3        | -1.2      | 6.0        | 0.5         | -0.6            | 6.6        | 1.1         | -13.0           | 0.0         | 1.0         | 0.0            |
| BIFSG - prob>0.8 | 105262 | 6.9      | 1.5        | 6.6       | 5.1        | -0.4         | 13.0            | 6.6        | 1.1         | -13.0           | 0.0         | 1.0         | 0.0            |

Notes:
- 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 percent level for all groups.
- The statistical significance of the difference in the probability-based prevalence is determined using t-tests with Satterthwaite approximation for unequal variances; and the statistical significance of the difference in the MAP-based prevalence is determined using chi-square tests.

B. Adjusted Denial Odds Ratios

| Method          | N obs | Hispanic       | NH Black       | NH API        |
|-----------------|-------|----------------|----------------|---------------|
|                 |       | Odds Ratio | Bias | Difference | Odds Ratio | Bias | Difference | Odds Ratio | Bias | Difference |
| self-report     | 173043 | 1.502     | 1.491       | 1.362       | 1.584     | 0.082   | -0.030      | 1.661     | 0.170      | 0.010        | 1.414     | 0.052      | 0.028        |
| BIFSG - prob    | 173043 | 1.614     | 1.651       | 1.386       | 1.584     | 0.082   | -0.030      | 1.661     | 0.170      | 0.010        | 1.414     | 0.052      | 0.028        |
| BIFSG - MAP     | 173043 | 1.481     | 1.480       | 1.344       | 1.481     | -0.021  | -0.011      | 1.480     | -0.011     | 0.000        | 1.362     | 0.000      | -0.017       |
| BIFSG - prob>0.8 | 154802 | 1.538     | 1.639       | 1.396       | 1.538     | 0.036   | 0.017       | 1.639     | 0.148      | -0.081       | 1.396     | 0.034      | 0.025        |
| BIFSG - prob>0.8 | 147919 | 1.521     | 1.719       | 1.371       | 1.521     | 0.019   | 0.017       | 1.719     | 0.228      | 0.009        | 1.371     | 0.009      | 0.009        |

Notes:
- Adjusted APR disparities are obtained from OLS regressions of APR on race and ethnicity indicators or probabilities, and controls for mortgage pricing factors.
- Adjusted denial odds ratios are obtained from logistic regressions of the denied/approved indicator on race and ethnicity indicators or probabilities and controls for mortgage underwriting factors.

Table 8. BIFSG Application: Distribution of Loans by Race and Ethnicity

| Ethno-racial Group | Self-Report Sample with race/ethnicity | Self-Report Sample without race/ethnicity | BIFSG-prob Sample with race/ethnicity | BIFSG-prob Sample without race/ethnicity | BIFSG-MAP Sample with race/ethnicity | BIFSG-MAP Sample without race/ethnicity |
|--------------------|---------------------------------------|------------------------------------------|---------------------------------------|------------------------------------------|-------------------------------------|------------------------------------------|
| Hispanic           | 13.3%                                 | 13.9%                                    | 12.7%                                 | 13.8%                                    | 12.6%                               | 12.4%                                    |
| NH Black           | 6.2%                                  | 6.8%                                     | 9.0%                                  | 5.5%                                     | 7.6%                                | 7.0%                                     |
| NH White           | 73.6%                                 | 71.7%                                    | 69.7%                                 | 74.6%                                    | 72.7%                               | 72.3%                                    |
| NH API             | 6.3%                                  | 6.1%                                     | 6.7%                                  | 6.0%                                     | 6.7%                                | 6.7%                                     |

Notes:
- Differences in prevalences between the two samples are statistically significant at the 5 percent level for all groups.
- The statistical significance of the difference in the probability-based prevalence is determined using t-tests with Satterthwaite approximation for unequal variances; and the statistical significance of the difference in the MAP-based prevalence is determined using chi-square tests.
Appendix

Derivation of the BIFSG Formula

Using the notation described in section 2.2, the conditional probability $p(r|s,f,g)$ can be written, applying the Bayes’ Rule and the chain rule, as follows:

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

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

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