Our Perspectives:
Commentary on the economy & regulatory policies affecting financial companies

Barnett Sivon & Natter P.C.
ATTORNEYS at LAW
WASHINGTON, DC

How CFPB Determines Proxies for Race and Ethnicity*

Bob Barnett

October, 2014

The public and the lending industry agree that there should not be discrimination in lending against persons because they are in a specific racial, ethnic, gender or other group covered by the Equal Credit Opportunity Act. At the same time, the ECOA prohibits lenders from "enquiring about the race, color, religion, national origin, or sex of an applicant or any other person in connection with a credit transaction."[1] Putting aside the dilemma that presents to a lender, how can a lender, regulator or the public generally know if the law is being observed, when missing the operative data?

It is clear that if there is direct evidence that a particular individual has been intentionally discriminated against, public policy and the law have been violated — a lender's documents direct loan officers not to make loans to members of a specific religion because they have a lousy record in repaying loans, and a loan is not made to an applicant who is a member of that group and who has a 780 FICO score. It is not so clear when a lender decides not to direct any advertising and mailings for loans into a geographic area that is heavily populated with consumers who are a part of that religion but has a major campaign into geographic areas that effectively surround that area. There is no direct data that demonstrates intentional discrimination for any particular borrower.

A number of groups have utilized alternative methods for displaying data that will stand-in as a substitute for direct data. The Consumer Financial Protection Bureau has recently released a description of the methodology it uses in assisting it in determining race and ethnicity of groups when there has been no report in documents of that data.[2] That information is then used in determining if there has been unintentional prohibited discrimination in practices followed by a lender, even if the practices are facially neutral. While the authority for the regulators to use these data with a disparate impact theory of discrimination under certain of the laws remains unsettled by the Supreme Court,[3] regulators are assuming the authority lawfully exists and are vigorously enforcing the theory of disparate impact.

*The information contained in this newsletter does not constitute legal advice. This newsletter is intended for educational and informational purposes only.

[1] 12. C.F.R. §1002.5(b).

[2] "***Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity***," CFPB (Summer 2014). The Bureau also relies on proxies for sex based on data from the Social Security Administration, but that is not reflected in this current release.

[3] See the Fair Housing Act issues raised in a series of U.S. Supreme Court filings, the latest of which is Texas Department of Housing and Community Affairs v. The Inclusive Communities, Docket No.13-1371.

The methodology

The basics of the CFPB methodology are fairly simple. The Bureau uses the census data for 2000[4] and from that data determines for each surname[5] what proportion that number is to the total population. For example, assume there are 100 persons named Smith on the Census Bureau data, and an analysis shows that 10 of those are African American. That will establish the probability of African Americans in any Smith population on the average, in this case 10%. If the individual has a compound surname, the first name is used unless using it fails to locate a surname on the list, in which case the second name is used. About 10% of the surnames are not on the Census Bureau list, and those names are excluded from the fraction.

In addition to using surnames to develop a proxy for race and ethnicity, the Bureau also uses geocoding. In other words, using Census Bureau data, the Bureau can determine, by various geographical units of disaggregation, how many persons of each protected group is in each geographic unit. In other words, if there are 1000 persons in zip code 12345, and 10% of them are African Americans, the probability is 10% that any one person will be African American.

Both surname and geographical analysis assist the Bureau in determining the impact on a geographical unit of facially neutral practices. The Bureau, however, uses a more sophisticated analysis that combines both the surname data and the geographical data into a single proxy probability for race and ethnicity. That system is called the Bayesian Improved Surname Geocoding System, or BISG.

BISG integrates both the surname analysis and the geographic analysis to obtain a more precise calculation of the probabilities. There is a standard equation that will be of use if one wishes to engage in the analysis,[6] but basically what Bayesian analysis does is to ensure that the probability of something being or happening is not left to just one set of conditions but integrates multiple conditions into the reasoning.

For example, assume a runner had run the 100 meter dash 12 times and on 8 of those times, he finished with a time of less than 11 seconds. What is the probability that in his race today he will finish in a time of less than 11 seconds? If you said 66 2/3%, you have concluded the intuitive, but you are wrong. The correct answer is — I don't know.

The reason why you don't know is that there are many factors that can influence the probabilities of a runner running at a particular speed, the distance being only one of them. For example, suppose you asked just a couple of obvious questions — on which of those races was the wind blowing with the runners, and on which of the races was the wind blowing against the runners. If the evidence shows that on 7 of the races in which he ran less than 11 seconds, the wind was blowing with the runners and on the other 5 races the wind was blowing against the runners, that should affect the probabilities. If the wind on the day in question is blowing against the runners, the probabilities are more like 20%.[7]

Bayesian analysis takes those additional considerations into account. In the Bureau's explanation, they start with the surname probabilities and then apply the geocoding to those by way of the Bayesian analysis. Effectively, they are saying that the probabilities of the surname analysis correctly describing the protected groups existing in the analysis depends in part on the geographic area being considered, and with respect to that, correct analysis will adjust the probabilities to that situation.

[4] They plan to update that when newer data are available.
[5] The data is concealed when there are fewer than 5 surnames.
[6] See any Internet site on Bayes and his theory, e.g., ***Wikipedia article on Bayes' Theorem***.
[7] See "***Bayesian Statistics for Dummies***," Keven Boone.

This methodology is only designed to find proxies for race and ethnicity in cases where those data are not known through reporting or any other direct method. As the Bureau said in its release,

> Statistical analysis based on proxies for race and ethnicity is only one factor taken into account by [the Bureau divisions] in our fair lending review of non-mortgage credit products. This paper describes the methodology currently employed by [the divisions], but does not set forth a requirement for the way proxies should be constructed or used by institutions supervised and regulated by the CFPB.[8]

Bob Barnett is a partner with the law firm of ***Barnett Sivon & Natter, P.C.***

[8] "Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity," CFPB, Summer 2014.