Poster: Measuring Biases in a Data Broker's Coverage

Levi Kaplan
Northeastern University
kaplan.l@northeastern.edu

Alan Mislove
Northeastern University
amislove@ccs.neu.edu

Piotr Sapiezynski
Northeastern University
p.sapiezynski@northeastern.edu

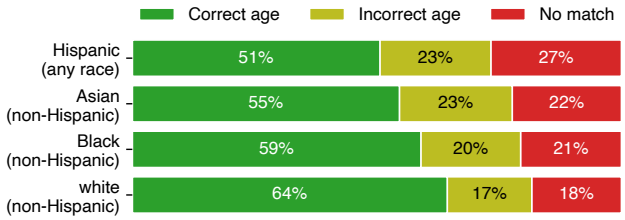

Figure 1: Differences in coverage and data accuracy among adult North Carolina population of different races. One in two Hispanic or one in three white voters are not represented or are represented inaccurately.

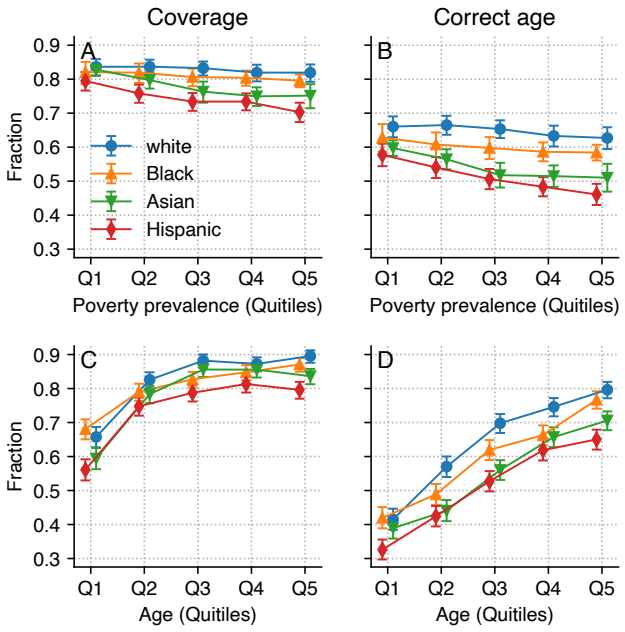

Figure 2: Coverage and accuracy varies not only with race, but also with the poverty levels in the ZIP code where individuals reside (A, B) and their age (C, D). Error bars represent the 99% confidence interval and plots are shifted along the x-axis for readability.

1 ABSTRACT

In the absence of a standardized form of identification in the United States, various businesses and organizations turn to the data broker industry to confirm the identity of, or perform background checks on, their clients. Often, potentially life-changing decisions depend on a successful and accurate match between the client's identity and data broker records; for example, decisions about housing, credit, employment, and—more recently—even access to vaccines against a global pandemic, can all be based in part on information from data brokers. However, the data brokers provide little transparency and it is notoriously difficult for researchers to study these companies at scale. In this work, we develop a measurement methodology to understand the coverage of one such data broker: Experian. We demonstrate that Experian's coverage of adults in North Carolina by is not only far from perfect, but is also worse for individuals who are more likely to be in historically disadvantaged groups. Our results indicate that younger populations as well as ethnic minorities and those living in lower income areas are less likely to be present in data broker databases, and even if they are, their data is more likely to be inaccurate than for white individuals and those living in more wealthy locations. These biases can potentially further exacerbate real-life societal divides along ethnic and economic lines, as they make access to essential life opportunities even more difficult for the most vulnerable populations.

We selected four samples of 8,930 individuals with four different self-reported races from voter records in North Carolina, created PII lists, and asked Experian to "append" their birth year.

Our measurements show that there are stark differences in data quality along the lines of race, ethnicity, age, and economic status. For example, the data we purchased on white non-Hispanic Americans was 25% more likely to be accurate than that on Hispanic Americans of any race. Further, only 32% of Hispanic voters below 26 were correctly represented by Experian, compared to 65% of those above 54.

Finally, we perform a logistic regression on this data to disambiguate how the factors of race and ethnicity, age, gender, and poverty contribute to the problem. We show that the racial differences persist even when we control for age, gender, and poverty.

Table 1: Dependent variable: Risk of wrong or missing data

Feature	Coefficient	std. err.	Risk change
Intercept	-0.405^{***}	±0.025	40.0%
Poverty in Zipcode	0.099^{***}	±0.011	+10.4%
Age	-0.381^{***}	±0.011	-31.7%
Female	-0.049^{*}	±0.022	-4.8%
Hispanic	0.602^{***}	±0.031	+82.6%
Asian	0.438^{***}	±0.031	+54.9%
Black	0.226^{***}	±0.032	+25.3%
AUC ROC = 0.626			

*p<0.05; **p<0.01; ***p<0.001