Race and Ethnicity in the 2020 US Census & Beyond: Machine Learning, Coding Models, and Privacy Concerns

A Featured Colloquium
By
Math-Stat Department (American University) & US Census Bureau

Date: Tuesday, April 2nd, 2024
Time: 2:00 to 4:00 pm
In-Person Location: Room 111, DMTI Building, East Campus
Zoom Link (off-campus): https://american.zoom.us/j/93790916053
(Meeting ID: 937 9091 6053; Tel#: +13017158592)

Alli Coritz
Senior Analyst, Racial Statistics Branch
Population Division

Magdaliz Alvarez Figueroa
Analyst, Ethnicity and Ancestry Branch
Population Division

Matt Spence
Senior Advisor, Special Population Statistics and Disclosure Avoidance
Population Division

Haley Hunter-Zinck
Data Scientist, Center for Optimization and Data Science Division

Contact: Nimai Mehta mehta@american.edu; Steve Casey scasey@american.edu

Race and Ethnicity in the 2020 US Census & Beyond: Machine Learning, Coding Models, and Privacy Concerns

Abstract: Data from the U.S. Census touch on many parts of American life, from political representation to business research. More than $2.8 trillion in federal funding was distributed in fiscal year 2021 to states, communities, tribal governments and other recipients using Census Bureau data in whole or in part.

The 2020 Census received more than 350 million detailed responses to the race and ethnicity questions — six times more than in the 2010 Census. These data are used widely across a variety of fields, including data science, the social sciences and more.

Subject matter experts from the U.S. Census Bureau will provide an overview of how these responses were coded in 2020, how data were protected in the 2020 Census using differential privacy, and ongoing research using machine learning to enhance existing coding processes. The presentation will cover three main areas:

- Improvements to the race and Hispanic or Latino origin (referred to as Hispanic origin) questions design, data processing and coding procedures. We will also provide insight into how race and ethnicity data were reported and coded in the 2020 Census.
- The 2020 Census used a new form of disclosure avoidance called differential privacy, which adds noise to published tables and statistics to prevent unintended disclosure of individual responses. We will discuss what differential privacy is, how it was applied to decennial data, the challenges we faced with implementing a wholly new method, and what plans we have for 2030.
- To standardize free-text survey responses for downstream data processing, responses are mapped to a standardized set of codes, often in the form of an ontology. While coding for downstream tabulation may be conducted manually, survey institutions frequently employ automated coding (auto-coding) procedures to reduce clerical workload and increase processing speed. Autocoding is especially challenging when free text responses are heterogeneous and the universe of possible codes is large, such as in the case of race and ethnicity coding of write-in responses to the Decennial Census. We investigate the feasibility of using classical machine learning and transformer models to map free text responses to a standardized ontology in order to enhance existing autocoding procedures.

United States® Census Bureau