Abstract

Paper will be available shortly; Please email CenHRSinfo@umich.edu to receive a copy when released

Finding Needles in Haystacks: Multiple-Imputation Record Linkage Using Machine Learning*

John M. Abowd[†] [‡] Joelle Abramowitz[§] Margaret C. Levenstein[§] Kristin McCue[†]
Dhiren Patki[¶] Trivellore Raghunathan[§] [‖] Ann M. Rodgers[§]
Matthew D. Shapiro[**§] [††] Nada Wasi[‡‡] Dawn Zinsser[§]

July 2021

Abstract

This paper considers the problem of record linkage between a household-level survey and an establishment-level frame in the absence of unique identifiers. Linkage between frames in this setting is challenging because the distribution of employment across establishments is highly asymmetric. To address these difficulties, this paper develops a probabilistic record linkage methodology that combines supervised machine learning (ML) with multiple imputation (MI). ML allows for improved match prediction accuracy, while MI propagates linkage uncertainty into subsequent analyses. This ML-MI methodology is applied to link survey respondents in the Health and Retirement Study to their workplaces in the Census Business Register. The linked data reveal new evidence that nonsampling errors in household survey data are systematically correlated with respondents' workplace characteristics.

Keywords: Machine learning; multiple imputation; probabilistic record linkage; survey data; administrative data.

*Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau, the Federal Reserve Bank of Boston, the principals of the Board of Governors, or the Federal Reserve System. We thank Jamie Fogel, Dyanne Vaught, and Sara Zobl for research assistance. All results have been reviewed to ensure that no confidential information is disclosed (release number CBDRB-FY21-CED006-0019). This research is supported by the Alfred P. Sloan Foundation through the CenHRS project at the University of Michigan with additional support from the Michigan Node of the NSF-Census Research Network (NCRN) under NSF SES 1131500.

[†]U.S. Census Bureau
[‡]Labor Dynamics Institute, Cornell University
[§]Institute for Social Research, University of Michigan
[¶]Federal Reserve Bank of Boston
[‖]Department of Biostatistics, University of Michigan
[**]Department of Economics, University of Michigan
[††]NBER
[‡‡]Puey Ungphakorn Institute for Economic Research, Bank of Thailand