A Novel Methodology for Improving Applications of Modern Predictive Modeling Tools to Linked Data Sets Subject to Mismatch Error

In recent years, the rise of social media platforms such as Twitter/X has provided social scientists with a wealth of user-content data, and there has been renewed interest in the utility of administrative records for increasing survey efficiency. Combining social media data, administrative records, and survey data has the potential to produce a comprehensive source of information for social research. These data are often collected from multiple sources and combined by probabilistic record linkage. For the analysis of these linked data files, advanced machine learning techniques, such as random forests, boosting, and related ensemble methods, have become essential tools for survey methodologists and data scientists. There is, however, a potential pitfall in the widespread application of these techniques to linked data sets that needs more attention. Linkage errors such as mismatch and missed-match errors can distort the true relationships between variables and adversely alter the performance metrics routinely output by predictive modeling techniques, such as variable importance, confusion matrices, RMSE, etc. Thus, the actual predictive performance of these machine-learning techniques may not be realized. In this presentation, I will describe a new general methodology designed to adjust modern predictive modeling techniques for the presence of mismatch errors in linked data sets. The proposed approach, based on mixture modeling, is general enough to accommodate various predictive modeling techniques in a unified fashion. I evaluate the performance of the new methodology with simulations implemented in R. I will conclude with recommendations for future work in this area.

Brady T. West is a Research Professor in the Survey Methodology Program, located within the Survey Research Center at the Institute for Social Research on the University of Michigan-Ann Arbor (U-M) campus. He earned his PhD from the Michigan Program in Survey and Data Science in 2011. Before that, he received an MA in Applied Statistics from the U-M Statistics Department in 2002, being recognized as an Outstanding First-year Applied Masters student, and a BS in Statistics with Highest Honors and Highest Distinction from the U-M Statistics Department in 2001. His current research interests include the implications of measurement error in auxiliary variables and survey paradata for survey estimation, selection bias in surveys, responsive/adaptive survey design, interviewer effects, and multilevel regression models for clustered and longitudinal data. He is the lead author of a book comparing different statistical software packages in terms of their mixed-effects modeling procedures (Linear Mixed Models: A Practical Guide using Statistical Software, Third Edition, Chapman Hall/CRC Press, 2022), and he is a co-author of a second book entitled Applied Survey Data Analysis (with Steven Heeringa and Pat Berglund), the second edition of which was published by CRC Press in June 2017. He was elected as a Fellow of the American Statistical Association in 2022.