

Combining information from multiple data sources using statistical modeling and methods

Professor Partha Lahiri

Joint Program in Survey Methodology and Department of Mathematics
University of Maryland College Park

Abstract

The demand for statistics on diverse topics—including socio-economic conditions, agriculture, health, and transportation—is on the rise, while governments and survey organizations have strived to address the increasing costs of conducting high-quality surveys. Along with technological advancements, the increasing accessibility of various data sources, including administrative records, geospatial data, social media data, and AI-generated data, presents researchers with new opportunities to produce improved estimates. In addition, this allows for the investigation of complex problems that would be challenging using only a single data source. Recently there has been a significant surge in statistical methodological research for diverse applications that is focused on combining information from multiple data sources.

In this presentation I will begin by briefly discussing the scope of statistical modeling to harness information from multiple data sources for precise estimates at a granular level, conducting multivariate analysis when a single data source lacks all relevant variables, reducing nonsampling errors in probability samples, mitigating self-selection biases in nonprobability samples, and addressing other emerging challenges. The remainder of my presentation will focus on recent statistical methodological developments for combining information from multiple data sources in the context of small area estimation for poverty mapping—a topic of significant interest to various national statistical offices and international agencies.