JOINT PROGRAM IN SURVEY METHODOLOGY

May 3-5, 2021

An Online short course sponsored by the Joint Program in Survey Methodology

Title of the short course:

Data Linkage

Name and contact information of the instructor:

Dr. Partha Lahiri, Professor and Director, Joint Program in Survey Methodology and Professor, Department of Mathematics, University of Maryland, College Park, USA; E-mail: plahiri@umd.edu

Description of course:

The demand for statistics on a range of socio-economic, agricultural, health, transportation, and other topics is steadily increasing at a time when government agencies are desperately looking for ways to reduce costs to meet fixed budgetary requirements. A single data source may not be able to provide all the data required for estimating the statistics needed for many applications in survey and official statistics. However, information complied through different data linkage or integration techniques may be a good option for addressing a specific research question or for multi-purpose uses. For example, information from multiple data sources can be extracted for producing statistics of desired precision at a granular level, for a multivariate analysis when a single data source does not contain all variables of interest, for reducing different kinds of nonsampling errors in probability samples or self-selection biases in nonprobability samples, and other emerging problems.

The greater accessibility of administrative and Big Data and advances in technology are now providing new opportunities for researchers to solve a wide range of problems that would not be possible using a single data source. However, these databases are often unstructured and are available in disparate forms, making data linkages quite challenging. There is, therefore, a growing need to develop innovative statistical data linkage tools to link such complex multiple data sets. In the US federal statistical system, the need to innovate has been emphasized in the following report: National Academies of Sciences, Engineering, and Medicine. (2017), *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24652.

For over 75 years, survey statisticians have been using information from multiple data sources in solving a wide range of problems. One early example of combining surveys can be traced back to a 1943 Sankhya paper (www.jstor.org/stable/25047787). Over the years, we have witnessed tremendous progress in such research topics as small area estimation, probabilistic record

linkage, combining multiple surveys, multiple frame estimation, microsimulation, poststratification, all of which incorporate multiple data sources and can be brought under the broader umbrella of statistical data linkages. In a 2020 Sankhya B paper (doi 10.1007/s13571-020-00227-w), Professor J.N.K. Rao provides an excellent review of a selected subtopics of statistical data integration. Papers covering different data linkage methods can be found in a special issue of Statistics in Transition new series (Volume 21, Number 4, August 2020): https://sit.stat.gov.pl/SiT/SpecialIssue/August%202020/SIT_Special%20Issue_Vol%2021_No_4%20PRINT.pdf

In this short course, we will give a general overview of each topic in a somewhat nontechnical way and illustrate available packages for data analysis.

Proposed course length:

Three day

Course Text and Materials:

The course will be based on the presenter's lecture slides.

Target Audience and Prerequisites:

The course is intended for practitioners and should be accessible to graduate students and early career researchers. An undergraduate level course in mathematical statistics and applied regression analysis are required.

Course content

1. Introduction
2. Probabilistic record linkage
3. Linking Multiple Probability and/or Nonprobability Survey Databases
4. Small Area Data Analytics
5. Micro-simulation Methods
6. Multiple-Frame Surveys
7. Big Data
8. Other Applications

JOINT PROGRAM IN SURVEY METHODOLOGY

Presenters

Dr. Partha Lahiri is Professor and Director of the Joint Program in Survey Methodology (JPSM) and Professor of Department of Mathematics at the University of Maryland, College Park, and an Adjunct Research Professor of the Institute of Social Research, University of Michigan, Ann Arbor. Prior to coming to Maryland, Dr. Lahiri was the Milton Mohr Distinguished Professor of Statistics at the University of Nebraska-Lincoln. His research interests include statistical data linkage. Dr. Lahiri's research has been widely published in leading journals such as the *Journal of the American Statistical Association, Annals of Statistics, Biometrika* and *Survey Methodology*. Dr. Lahiri has served on a number of advisory committees, including the U.S. Census Advisory committee and U.S. National Academy panel. Over the years Dr. Lahiri advised various local and international organizations such as the United Nations Development Program, World Bank, Gallup Organization. Dr. Lahiri is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics and an elected member of the International Statistical Institute.

Agenda:

MONDAY: MAY 3, 2021

9:15 - 10:15 Introductions: Different Applications of Data Linkage Methods

10:15-10:30 Discussion

10:30 - 10:45 Morning Break.

10:45 - 11:45 Probabilistic Record Linkage

11:45 – 12:00 Discussion

12:00 - 1:00 Lunch Break

1:00 - 2:00 Linking Multiple Probability and/or Nonprobability Survey Databases

2:00 – 2:15 Discussion

TUESDAY: MAY 4, 2021

9:15 - 10:15 Small Area Data Analytics

10:15-10:30 Discussion

10:30 - 10:45 Morning Break.

10:45 - 11:45 Micro-simulation Methods

11:45 – 12:00 Discussion

12:00 - 1:00 Lunch Break

1:00 - 2:00 Multiple-Frame Surveys

2:00 – 2:15 Discussion

WEDNESDAY: MAY 5, 2021

9:15 - 10:15 Big Data

10:15-10:30 Discussion

10:30 - 10:45 Morning Break

10:45 - 11:45 Other Applications

11:45 – 12:30 Discussion