33rd Annual Morris Hansen Lecture

Discussant Remarks

Discussant: Lisa B. Mirel

Speaker: Partha Lahiri

Title: *Combining Information from Multiple Data Sources Using Statistical Modeling and Methods*

NATIONAL CENTER FOR SCIENCE AND ENGINEERING STATISTICS
U.S. NATIONAL SCIENCE FOUNDATION

Discussant Remarks: Opening

- Thank you to Washington Statistical Society, Hansen Lecture Committee, Westat, USDA NASS, and Summit, LLC
- Congratulations to Partha Lahiri
- Continuing to honor Morris Hansen
  - Speaker is chosen who meets outstanding merit to cover emerging topics and diversity of disciplines to inform survey practice
  - Stand on shoulders of greatness
- Today's lecture key message: The successful integration of multiple data sources is reliant on robust statistical methods to inform official statistics

Motivation

- Federal Statistical System is facing challenges
  - Declining response rates
  - Increasing costs to conduct surveys
  - Demand for more timely granular statistics
  - Lack of standardized practices for working with non-survey data
- Opportunities
  - Link data, combine probability and non-probability data
  - Utilize non-survey design data (e.g., administrative records, geospatial data, social media data)
  - Link data to improve small area estimation
  - Develop standardized practices for data integration

Critical Question

- How can we integrate data securely and ensure sufficient quality for official statistics?

Connections to survey statistics

- Data integration can inform gaps in surveys
  - Improve content and coverage
  - Permit longitudinal analyses
  - Improve quality of estimates
- However, we still need to balance quality and confidentiality

Methodological questions

- How do we assess the suitability of non survey/nonprobability data for linkages?
  - Coverage
  - Measurement error
- How do we ensure privacy and confidentiality when linking multiple sources?
- How do we integrate innovative methodological tools to stand behind official statistics?
- How can we capture linkage errors especially with privacy enhancing technology?

Implementation questions

- How can statistical agencies maintain public trust when using integrated sources?
- Should we be striving toward consistency in assessment methods?
- Can we develop a shared service toolkit to conduct data quality assessments of our integrated data?
- What governance mechanisms are needed to support large scale data integration?

Need for infrastructure to support data integration: National Secure Data Service

- Government-wide initiative to strengthen secure data access and linkages
- Facilitate statistical activities to support data-informed decision making
- Connect research communities to share innovative methods and results (e.g., code, frameworks, publications)
- Use of emerging technologies

NSDS Demonstration Project: Core Components for Data Integration

- Implementing privacy enhancing tools to link data
- Establishing a data linkage program framework
- Assessing the use of alternative data for survey integration efforts
- Evaluating alternate data collection frames and estimation techniques

NSDS: Problem and Solution Approach

PROBLEM

Linking data with attribute records can increase disclosure risk

SOLUTION

Privacy Preserving Record Linkage (PPRL) and Secure Multiparty Computation (SMPC) tools can mitigate these risk by encrypting data prior to linking data

PROBLEM

Linking data without a clear framework can ignore legal authorities and ethical responsibilities

SOLUTION

Systematically create a framework for a data linkage program that includes standardized approaches that can be used by others

NSDS: Problem and Solution Approach (cont.)

PROBLEM

Need for efficiency in surveys to reduce duplication while still maintaining quality

SOLUTION

Integrate multiple surveys into one and use data integration to support efficiencies

PROBLEM

There is a rapidly changing landscape of survey data collection and dissemination of federal statistics

SOLUTION

Develop alternative data collection frames and data dissemination methods using innovative tools

Concluding remarks

- NSDS projects and infrastructure builds support both the methodology and implementation frameworks to inform a data integration model
- Having a consolidated shared service platform allows for
  - Secure data access and linkages
  - Successful and responsible data integration
  - Support of data driven decision making
- Data integration can rely on new technologies and data sources, while still maintaining the transparency, quality, objectivity, and independence that was previously applied survey data

Thank you!

NSDS website

Project descriptions, final reports, and lessons learned