NISS | *The Statistics Community Serving the Nation*

Research Highlights

Dissemination of High-Quality Data while Protecting Data Confidentiality

The Triangle Census Research Network (TCRN), established by NISS and Duke University, develops broadly-applicable methodologies intended to transform and improve data dissemination practice in the federal statistical system. It focuses primarily on methods for (1) handling missing data and correcting values in large complex surveys, (2) disseminating public use data with high quality and acceptable disclosure risks, and (3) combining information from multiple data sources, including record linkage techniques.

Project goal: *Develop methodology to improve current practice for handling missing and faulty data and statistical disclosure limitation.*

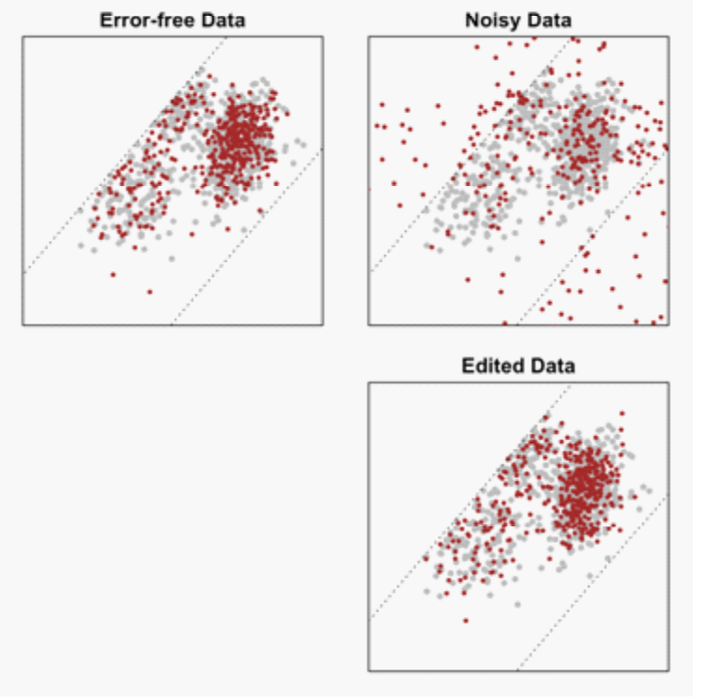
Edit-Imputation of Microdata

Publicly released microdata by federal agencies potentially support complex and rich secondary analyses that lead to deeper scientific understanding and more informed policy making; yet, data quality often suffers because data subjects are reluctant to respond to surveys or are prone to report data with errors. Statistical agencies frequently handle nonresponse with ad hoc approaches that tend not to work well in complex datasets with many variables—especially when data are released to the public—because they often ignore multivariate relationships. Similarly, editing faulty data is frequently based on heuristics rather than principled theory.

NISS and Duke researchers have developed a general framework for simultaneous imputation of missing data and editing of faulty data.. This framework adopts nonparametric Bayesian methodology to reflect complex distributional features of collected data. By integrating paradigms from statistics and operations research, the resulting edit-imputed data are guaranteed to satisfy logical constraints provided by domain-experts. The approach is applied to data from the Census of Manufactures, where we demonstrate the improvement of our procedures over existing approaches to edit-imputation

Data Confidentiality

Most federal agencies view disseminating data to the public for secondary analyses as a core mission; yet, concerns over data confidentiality make it increasingly difficult to do so. As threats to data confidentiality grow, federal agencies planning to produce public use data may be forced to release heavily redacted files. Many confidentiality protection strategies applied at high intensities result in severely reduced data quality. Even worse, analysts of secondary data have no way to determine how much their analysis has been compromised by the disclosure protection.



NISS and Duke researchers have extended the theory and methodology for releasing multiply imputed, synthetic datasets based on flexible, nonparametric Bayesian models. We generate synthetic data that preserve features of the joint distribution while respecting linear constraints among variables. Ongoing research topics include generating synthetic data for survey data with sampling weights, and developing the framework for computer systems that provide secondary analysts with feedback on the quality of inferences from heavily redacted data.

Research Team: The current NISS team members include Larry Cox (NISS), Saki Kinney (NISS), Hang Kim (NISS/Duke postdoc) and Alan Karr (RTI/NISS) and the Duke team is leaded by Jerry Reiter (Duke).

Funding Sponsor: The project is funded by National Science Foundation in partnership with the Census Bureau under the NSF-NCRN-MN grant mechanism (2011-2015).

Continued on back of page

Publications

1) H. J. Kim, A. F. Karr, and J. P. Reiter (forthcoming), "Statistical disclosure limitation in the presence of edit rules", Journal of Official Statistics.

2) H. J. Kim, J. P. Reiter, Q. Wang, L. H. Cox, and A. F. Karr, (2014) "Multiple imputation of missing or faulty values under linear constraints," Journal of Business and Economic Statistics, 32, 375 – 386.

3) A. F. Karr, (2014), "Why data availability is such a hard problem," Statistical Journal of the International Association for Official Statistics, 30, 101 – 107.

4) A. F. Karr and J. P. Reiter (2014), "Using statistics to protect privacy," in Privacy, Big Data, and the Public Good: Frameworks for Engagement, edited by J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, Cambridge University Press, 276 – 295.

5) L. Cox, (2014), "Enabling statistical analysis of suppressed tabular data," in Privacy in Statistical Databases, edited by J. Domingo-Ferrer, Lecture Notes in Computer Science 8744. Heidelberg: Springer, 1-10.

6) J. P. Reiter and S. K. Kinney (2012), "Inferentially valid partially synthetic data: Generating from posterior predictive distributions not necessary," Journal of Official Statistics, 28, 583 – 590.

* Refer to http://sites.duke.edu/tcrn/ for more papers produced by the TCRN project

Conference Presentations

"Bayesian Data Editing for Continuous Microdata", Joint Statistical Meetings, Boston, MA, August 2014

"Bayesian Data Editing for Continuous Microdata", Federal Committee on Statistical Methodology (FCSM) Research Conference, Washington, DC, Nov 2013

"Multiple Imputation Under the Edit Constraints", Joint Statistical Meetings, Montreal, Quebec, Aug 2013

Invited Presentations

"Bayesian Data Editing for Continuous Microdata", Invited speaker, U.S. Census Bureau, Center for Statistical Research Methodology (CSRM) Seminar, Washington, DC, April 2014

"Bayesian Data Editing for Continuous Microdata", SAMSI Computational Methods for Censuses and Surveys Workshop, Washington, DC, Jan 2014

"Multiple Imputation Under the Edit Constraints", U.S. Census Research Center for Economic Studies Seminar, Washington, DC, Apr 2013