

Data Confidentiality, Data Quality and Data Integration for Federal Databases

Alan F. Karr
National Institute of Statistical Sciences
Research Triangle Park, NC 27709-4006, USA

1 Objectives and Impact

The principal high-level goal of the research is to develop abstractions, theory and methodology and software tools that allow federal statistical agencies to disseminate useful information derived from confidential data but protect the privacy of data subjects—individuals and establishments.

Specific scientific objectives include:

- Problem formulations and scalable tools that accommodate both *disclosure risk* and *data/information utility*;
- Understanding *consequences of data integration* for data confidentiality, data quality and statistical inference;
- Creation of fundamental quantifications, usable models, scalable methods for *data quality*.

Impacts of the research include:

- Protecting government-collected data on individuals and establishments from increasingly severe threats to confidentiality;
- Protecting privacy of individuals and establishments;
- Preventing federal data warehouses from becoming data cemeteries;
- Assisting agencies in preparing for a “world without releasable microdata.”

2 Selected Accomplishments

The project is leading to a paradigm shift in statistical disclosure limitation. New techniques are based on scalable risk-utility formulations that enable agencies to balance disclosure protection against the utility of released information.

Geographical Aggregation. Algorithms and software for achieving disclosability by aggregating adjacent geographical units such as counties were developed using data provided by the National Agricultural Statistics Service (NASS). These enable release of information below the state level, which was previously thought infeasible.

Tabular Data. NISS table servers are the first successful implementation of query systems containing principled, scalable methods for dealing with query interaction. The project has also developed:

- Scalable methods and software to compute bounds on cell entries from released marginals;
- Scalable risk-utility formulations, which lead to optimal tabular releases maximizing data utility subject to a constraint on disclosure risk;

- Prototype table server software.

Data Swapping. Principal products of the research include:

- A complete risk-utility formulation for treating data swapping as a decision problem, with multiple measures utility/distortion and disclosure risk;
- The NISS Data Swapping Toolkit (DSTK): operational software for performing swapping large-scale studies to select the swap attributes and swap rate, as well as for visualization of the results. The DSTK is available at www.niss.org/software/dstk.html.
- A Web service implementation of data swapping, for which software is available at www.niss.org/WebServices/dg/WebSwap.html.

Remote Access Analysis Servers. The project has created fundamental abstractions—query space, answer space, disclosure risk measure and data utility measure—for systems that disseminate the results of statistical analyses of confidential data. *Regression servers* that optimally protect a sensitive variable have been developed, and software is being built.

Secure Regression. Techniques for secure multi-party computation have been used to implement secure linear regression on multiple, horizontally partitioned databases. Least squares estimators, standard errors and the coefficient of determination R^2 can be calculated in a way that minimizes risk to individual data elements. Diagnostics can be performed via by means of securely integrated synthetic residuals.

Corresponding techniques for vertically partitioned data and more complex partitioning, including missing data, are currently being developed.

3 Project Structure

The National Institute of Statistical Sciences (NISS) is lead institution for the project. University partners Carnegie Mellon University, Duke University, Purdue University and Southern Methodist University. Federal statistical agency partners are the Bureau of Labor Statistics (BLS), Bureau of Transportation Statistics (BTS), Census Bureau (Census), NASS and National Center for Education Statistics (NCES). All have provided both data and support.

Acknowledgements

This research was supported by NSF grant EIA-0131884 to NISS.