Statistical Disclosure Control: New Directions and Challenges

Natalie Shlomo*
The Cathie Marsh Centre for Census and Survey Research, University of Manchester, United Kingdom Natalie.Shlomo@manchester.ac.uk

Protecting the confidentiality of statistical entities is a key factor in the dissemination of statistical outputs. Traditionally, statistical agencies disseminate outputs in the form of microdata arising from social surveys, census tables containing whole population counts and magnitude tables from business surveys or registers. For these types of outputs, there has been much research on how to quantify the risk, how to control the disclosure risks and how to assess the loss of information. However, with increasing demand for more access to statistical information, new strategies for disseminating statistical outputs are being researched and developed. These stategies include web-based applications, such as: flexible table generation where users specify and download their own tables of interest; and remote analysis platforms where users carry out statistical analysis through an interface on non-confidentialized microdata and only have access to the results of the analysis. In both applications, the outputs are generated from the original data and disclosure control techniques are applied 'on the fly' to the outputs prior to dissemination without the need for human intervention. Examples of automated disclosure control techniques include perturbation of counts generated in tables, or replacing scatterplots by grouping one of the continuous variables and disseminating sequential box-plots. Another dissemination strategy is on-site data enclaves where approved researchers can gain access to non-confidentialized microdata at a central location, and its extension where researchers can access the confidential data on their personal PCs through a remote connection to special servers which are closed to other networks. The outputs are stored on the server until they are manually checked for disclosure risks and returned to the researcher via email. Business microdata is not released due to their high sensitivity and skewness of the variables. For these datasets, new disclosure control techniques rely on the production and release of 'synthetic' data based on replicates of the original data drawn from advanced statistical models. All of these new strategies for disseminating statistical outputs introduce challenges to the disclosure risk assessment, on-line and 'on the fly' applications of statistical disclosure control techniques and the measurement of information loss.

Key Words: Flexible table generation, remote access, remote analysis, synthetic data