Working Paper No. 23 (Summary)

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (vi): Software tools for statistical disclosure control

MASSC: A NEW DATA MASK FOR LIMITING STATISTICAL INFORMATION LOSS AND DISCLOSURE

Invited paper

Submitted by Research Triangle Institute International, United States[1]

[1] Prepared by A.C. Singh (asingh@rti.org), F. Yu, and G.H. Dunteman.

MASSC: A new data mask for limiting statistical information loss and disclosure

A.C. Singh, F. Yu, and G.H. Dunteman
Statistics Research Division
RTI International, NC 27709

ABSTRACT

We propose a method termed 'MASSC' for ensuring statistical disclosure limitation (SDL) of categorical or continuous micro data, while maintaining the analytical quality of the micro data. The new SDL methodology exploits the analogy between (1) taking a sample (instead of a census,) along with some adjustments, including imputation, for missing information, and (2) releasing a subset, instead of the original data set, along with some adjustments for records still at disclosure risk. Survey sampling reduces monetary cost in comparison to a census, but entails some loss of information. Similarly, releasing a subset reduces disclosure cost in comparison to the full database, but entails some loss of information. Thus, optimal survey sampling methods can be used for statistical disclosure limitation. The method includes partitioning the database into risk strata, optimal probabilistic substitution, optimal probabilistic subsampling, and optimal sampling weight calibration.

The proposed method uses a paradigm shift in the practice of disclosure limitation in that the original database itself is viewed as the population and the problem of disclosure by inside intruders is considered. (Inside intruders know the presence of their targets in the database in contrast to the outside intruders.) This new framework has two main benefits: one, it addresses the more difficult problem of protecting from inside intruders as a result of which it automatically protects against outside intruders, and second, it allows for quantification of both information loss and disclosure risk when disclosure treatment is performed by employing known random selection mechanisms for substitution and subsampling. Empirical results will be presented to illustrate computation of measures of information loss and the associated disclosure risk for a small data set.