Working Paper No. 30 (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 (ii): New data release techniques

IMPLEMENTING STATISTICAL DISCLOSURE CONTROL FOR AGGREGATED DATA RELEASED VIA REMOTE ACCESS

Contributed paper

Submitted by the National Statistical Institute of Italy [1]

[1] Prepared by Giovanni Merola (gmmerola@istat.it).

Abstract

Implementing Statistical Disclosure Control for Aggregated Data Released via Remote Access

Giovanni Merola
Italian National Statistical Institute
Methods for Statistics Production
Roma, Italy
gmmerola@istat.it

Keywords: Remote access, tabular data, statistical disclosure control, on-line database

In this joint work with Dr. L. Franconi, we analyze different strategies for implementing different Statistical Disclosure Control methods to systems that release tables to users connected remotely. Dissemination is one of the missions of National Statistical Institutes (NSIs); it is a way for giving society useful information and it also a way of motivating respondents to answering surveys. If NSIs have the legal standing to collect and release information, they also have the obligation to protect confidentiality of respondents, which sets a constraint on the amount of information that can be released. This trade-off is addressed by Statistical Disclosure Control (SDC), which consists of a collection of techniques that make difficult matching confidential information with the identity of respondents from a set of data. SDC methods employ a variety of techniques that either alter or suppress some of the data released.

We classify the data released in three categories:

- **Microdata**: files with individual observations;
- **Tables**: total values carried by individuals that fall in given classifications;
- **Other statistics**: summaries of different types, for example: regression coefficients, relative indices, correlation coefficients, etc.

In this paper we do not consider the release of microdata files but only of aggregated values. In particular, we focus on the release of tables, although some of the conclusions that we draw can be applied to the release of other statistics, as well.

NSIs have a long tradition in publishing periodically printed reports of the data that they collect, which are either sold or distributed freely. However, the Internet is becoming a standard channel with which public institutions communicate with general public. Institutions expect people to look for information on the Web and Internet navigators expect to find it on the Web. Furthermore, Web-sites are an ideal tool for disseminating data, as they are cheap, flexible, easy to update and accessible by most users (Blakemore, 2001). In fact, most NSIs already provide on-line data in tabular form. Such Web-sites, to which we refer as Web-based Systems for Data Dissemination (WSDDs), require automated systems that release data upon request. WSDDs can be designed in many different ways, whether giving access to a predefined set of tables or allowing users to query tables, choosing from a set of available variables. Therefore, SDC methods must be applied to WSDDs according to their structure and flexibility as well as to the type, quality and level of detail of the information released. In many cases SDC must also be combined with electronic access control.

The implementation of SDC to WSDDs requires the definition of standard protection procedures, often applied in an automated way. We argue that SDC methods can be applied to WSDDs in two ways: *a priori*, that is, before releasing the tables, or *a posteriori*, that is, after the user has made his/her particular query. We refer to the former as PRE SDC and to the latter as POST SDC. According to this classification we comment on benefits and drawbacks of different SDC approaches.