

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**COMMISSION OF THE EUROPEAN
COMMUNITIES**

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

**Joint ECE/Eurostat Work Session on
Statistical Data Confidentiality**
(Thessaloniki, Greece, 8-10 March 1999)

Working Paper No. 5 (Summary)
English only

Topic (i) - New applications of disclosure control methods

**DISCLOSURE CONTROL METHODS IN THE PUBLIC RELEASE OF A MICRODATA FILE
OF SMALL BUSINESSES**

Submitted by Statistics Canada¹

Contributed paper

Summary

1. On 19 March 1997, the Microdata Release Committee of Statistics Canada approved the release of a public use microdata file (PUMF) of financial data of small businesses. This was a first for Statistics Canada as approval for the release of a *business* microdata file had never before been obtained. The difficulty has always been related to the continuous and skewed nature of financial data. It was felt that it was too easy to deduce the identification of a business from the uniqueness of its data. Still the demand for the release of business microdata files has remained.
2. In 1996 Industry Canada approached Statistics Canada. They wanted access to small businesses= data to improve the general knowledge and understanding of the financial structure of small businesses, to help establish a more level playing field in assessing loan risk for small businesses, and to provide data users with full flexibility in conducting custom analysis of small business performance. Since the data of small businesses do not contain the extreme outliers of the complete business sector, it was felt that Statistics Canada could attempt to develop a small business PUMF.
3. The attempt was successful. This paper describes the approach used to develop the microdata file and to measure its data quality. The most effective part of the disclosure control methodology is the variety of disclosure control methods used to create the file. This makes it extremely difficult for an intruder to untangle the perturbations.
4. The following is an outline of the disclosure control method. First, the proportion of units from the small business population that exist on the PUMF is controlled; that is, the PUMF is a sample. An intruder interested in a particular small business thus cannot be certain that it exists on the PUMF. Second, each datum of each record in the PUMF is perturbed in two ways, and the extreme data values of each field are perturbed a third way. Thus an intruder who makes an identificationCwhether correctly or incorrectlyCdoes

¹ Prepared by Stuart Pursey.

not have the exact data for the business as they have been altered in some way. Third, the proportion of PUMF records that can be correctly linked to a population file (using an effective method of record linkage) is controlled by further perturbation, where required. It is therefore difficult for an intruder to make correct links from the PUMF to the population file (where identification information resides). Fourth, some records are so unusual that nothing can be done to protect these records while maintaining reasonable data quality. These unique records are not included in the PUMF.

5. The paper also discusses an approach to the analysis of data quality. The analysis may address the original data (micro data analysis) or the statistics generated from the data (macro data analysis). Data quality is maintained if the distance between an unmodified datum or statistic (before disclosure control) and the modified datum or statistic (after disclosure control) is small. Statistical measures to describe and summarize this distance are discussed. The impact of each step of the disclosure control process can be examined as well.

6. The analysis showed that unincorporated small businesses maintained much more data quality than incorporated small businesses. This was expected since the data of incorporated businesses require much more perturbation than those of unincorporated business to maintain their confidentiality. Although it is possible to isolate certain financial variables that are quite adversely affected by the disclosure control process relative to other variables, generally it is not possible to isolate particular variables and industries that fared particularly well or not well. This is encouraging as we want a disclosure control method that is equally 'good' and 'bad' to industries and variables. Still there is a fair amount of variability in quality by variable and by industry. It would be better if the disclosure control method provided a more consistent and predictable drop in quality. Overall the quality of data for unincorporated small businesses is good to excellent and for incorporated small businesses it is poor to good.