CASC PROJECT

Computational Aspects of Statistical Confidentiality

31 December 2003

Proposal for the creation of a micro-data file for research for Business Surveys – Part 2

Internal report – not for reproduction or citation

Luisa Franconi
Giovanni Seri

ISTAT, PSM/C
Via C. Balbo, 16
00184, Roma
Italy

Deliverable No: 5-D5

1. Introduction: Proposals for Anonymising CIS3 microdata

Commission Regulation 831/2002 establishes the conditions for the release of anonymised microdata sets from the Community Innovation Survey. Anonymised microdata are "individual statistical records which have been **modified** in order to **minimise**, in accordance with common best practice, **the risk of identification of the statistical units to which they relate**".

This document briefly outlines two proposals to anonymise CIS3 microdata. The first one limit the information content without modifying the data, whereas the second adopts a perturbation of the microdata to minimise the risk of identification of a unit. In Section 3 we outline the first of such method and in Section 4 the second, microaggregation. We present results of the application of both methods to the Italian sample of CIS2 microdata and analyse the risk of disclosure as well as information loss associated with these methods.

2. Classification of CIS variables

Obviously, all the variables that may lead to a **direct** identification of a unit are to be removed from the file (e.g. name of the enterprise).

The **identifying variables**, those who may lead to the identification of a respondent, for the CIS data are: **principal economic activity**, **geographical area**, **number of employees**, and any **continuous variables** (such as turnover, exports, gross investment or innovation expenditures) that, because of its size, presents peculiar values for very large enterprises.

A characteristics of CIS data is the relatively little presence of continuous variables. Most of the survey specific variables are categorical and expressing personal evaluation (e.g. the whole section on Objective of Innovation, sources of information for innovation and so on). This implies that from them it is not possible to discriminate between small and large enterprises and therefore identify a unit. All these other variable are survey specific variables and, in general, are released as they are.

2.1 *Stratifying CIS3 data by economic activity and geographical area*

Main economic activity and **geographical area** are considered ***stratifying*** **variables** and, in general, they are released in **broad categories**. This will be true for any protection methods we want to apply because of the importance of these variables.

For the CIS data, in order to make units less identifiable, we suggest to recode the **geographical area** (**NUTS**) at **national level**. Moreover, we suggest to release the variable economic activity, identified by the **NACE** code, giving only the first **2 digits** (Division).

Each combination of 2 digit NACE code and country of residence represents a stratum. If a 2 digit NACE code contains too few enterprises it is aggregated with another one (e.g. a contiguous NACE code). In Italy divisions 10, 11, 13 and 14 should be aggregated together as well as 15 and 16.

The units at risk of identification are the large and most easily recognizable enterprises, and extreme care should be put to protect them. To make identification a difficult task, anonymisation techniques are applied that reduce the information content of the continuous variables. In the next section a non perturbative method is suggested for these variables whereas in section 4 a perturbative method is presented.

3. A Non Perturbative Method: top coding and turning absolute figures into relative values

The first proposal combines several methods that do not perturb the data in order to guarantee the confidentiality of the largest and most identifiable units while preserving the complete information without distortion of all the other units. In what follow we present the treatment the other identifying variables should undergo.

The variable **number of employees** is top coded. This means that all records presenting a value for the variable number of employees greater or equal to a given threshold are assigned this threshold instead of the real figure. It would be desirable to have a common value for all strata in a country or indeed for all the countries involved. However, it has to be noticed that if a NACE presents an enterprise that has been top coded then at least another enterprise has to be top coded as well otherwise it is possible to recover the exact value of all the numerical variables by differencing with published national tables. We do not consider here the possibility that one (top coded) enterprise can recover the value of the other (top coded) one as the file is for research purposes and it is given only to *bona fide* researchers.

As an example for the Italian sample (we used the CIS2 data NACE codes from 10 to 41) we selected a top coding at 1000, which mean that instead of releasing the real value, the value 1000 is given to all enterprises with 1000 or more employees. This top coding has affected 168 enterprises on a sample size of 5256 (3.2%). The number of NACE where there is only one suppression is 5. For these we have selected the top coding at lower level increasing to 186 the number of enterprises affected by the protection. Under a clerical review of each single stratum the threshold varies from 300 to 3000 and the number of enterprises affected by the top coding is less than 90 (1.7% of the sample size).

Once the dimension of the enterprise is protected, all the continuous variables are expressed as relative to this dimension. For example, the variable **turnover** is released as turnover per employee. This allows the user to compute perfectly the value of the turnover for all the enterprises which are not top coded. For the enterprises which are top coded the user has at least the **exact** value of turnover per employee and multiplying this value times the top coded value of the number of employees can reach a lower bound for the turnover of large enterprises. Moreover, in many studies and micro-econometric models only relative values are of interest.

The same procedure is applied to the variable **export** and to all the numerical variables that relate to different types of **expenditure**. Alternatively, these kind of figures can be released as percentage of the turnover (e.g. export/turnover).

3.1 Risk evaluation

What we want to achieve while releasing anonymised microdata is to offer the user the maximum information while preserving confidentiality of respondents. Confidentiality is kept if it is not possible to identify a unit in the sample i.e. to link a record in the file to an enterprise in the population. This link might be possible if either
(i) there are very few units in a certain stratum
or
(ii) a unit presents unusual values for one or more variables.

The case (i) is detected by checking the size of each stratum.

Unusual values, case (ii), are dangerous as they may reveal the size of the enterprise and large enterprises are more identifiable. However, expressing all the continuous variables as relative values, the only variable that can be used to identify an enterprise is the number of employees. If, inside a stratum, the values of the number of employees are not very dissimilar then, it is difficult to identify an enterprise. If extreme values are present then top coding is applied.

In a sense the measure of disclosure risk we are implying by means of this collection of protection methods relies on the avoidance of extreme cases. This means that controls have to be made on the variable number of employees in order to guarantee that after top coding no outlying observations are still present. Automatic check can be improved on the basis of statistical rules as, for example, 'an enterprise is an outlier if its observed number of employees differs from the median value more than 1.5 times the interquartile range'. Otherwise, expert clerical review maybe more accurate.

A valuable feature of this method is that there is no need to delete any observations from the data set, as absolute figure are not released.

3.2Data validity

The method is very elementary. At the same time it leaves the whole of the information unchanged for units which are not at risk while being effective for large enterprises. Using a top coding value at 1000 for the Italian sample of the CIS only 186 enterprises see a reduction of the information content of the variable number of employees i.e. 3,5% of the sample.

3.3Implementation

The method is straightforward to implement in any statistical package and does not require specific knowledge.

In Table 1. we summarize the type of protection proposed for the structural variables and all the other continuous variables.

Table 1: Protection proposed for main CIS variables; non perturbative method.

CODE	Question	Action to be taken (in bold)
ID	Name of enterprise	**To be deleted in individual records**
NUTS	Nuts level 2	**National level**
NACE	Main activity	**2 digit NACE Rev. 1 – If too small further aggregation with contiguous NACE**
GP	Is your enterprise independent or part of an enterprise **g**rou**p** ?	Left unchanged
CHG_1	Your enterprise was established	Left unchanged
CHG_2	Turnover increased due to merger with another enterprise or part of it	Left unchanged
CHG_3	Turnover decreased due to sale or closure of part of the enterprise	Left unchanged
EMP	Number of **emp**loyees	**Top coding**
EMPC	**C**hange in number of **emp**loyees 1994-1996	Left unchanged (as already expressed as %)
TURN	**Turn**over in 1996	**Turnover per employees**
TURNC	**C**hange of **turn**over 1994-1996	Left unchanged (as already expressed as %)
EXP	**Exp**ort in 1996	**Export/turnover**

EXPC	Change of **exp**ort 1994-1996	Left unchanged (as already expressed as %)
RTOT	Total expenditure	**Expenditure/turnover**
	Other expenditures	**Other expenditure/Total expenditure**

4. A Perturbative Method: microaggregation

The basic idea of microaggregation is to cluster units into small aggregates or groups of size at least three (Deafays and Nanopoulos, 1993).

As before we apply this protection method to each stratum. We present the application of the microaggregation method (as it is implemented in μ-Argus, see Domingo-Ferrer and Mateo-Sanz, 2002) to a specific stratum to get an idea of the possible results.

As regards the continuous variables 'export' and the various items of 'expenditure' many responses are either missing or '0'. This causes too high an impact on the microaggregated data. That is why we adopt a similar reasoning as in the previous section and suggest to release continuous identifying variables as a direct or indirect transformation of the variable number of employees and/or turnover. These two latter variables only are involved in the microaggregation process.

4.1Risk evaluation

Microaggregation guarantees an high level of data protection as it produces groups of at least three identical units with regard to the identifying variables involved in the microaggregation process. Protection of the other identifying variables is a consequence of the transformation into ratio or percentage. Identification of enterprises is difficult because data are perturbed and at least three enterprises present the same identifying values.

4.2Data validity

Concerning the number of employees and turnover the method maintains the totals for all the strata, does not guarantee variances and covariances and reduces the variability. Identifying variables computed as a transformation of the 'original' values of the number of employees and turnover do not need to be perturbed. However, inverse transformation can only be made on the basis of the microaggregated values and the result will be consequently distorted.

4.3Implementation

This method is currently being implemented in the software μ-Argus as part of the European project CASC (Computational Aspect of Statistical Confidentiality). The software is available free of charge from http://neon.vb.cbs.nl/casc/.

5. Hybrid proposal

To improve quality of data with respect to microaggregation and the protection level of released data with respect to the top coding based method we suggest a third proposal compromising the previous two. The idea is to modify the first method by releasing the weighted average of the extreme values (at least the largest three enterprises) of the number of employees in each stratum. With respect to the first proposal we improve the information content for the enterprises affected by top coding. In fact, the average is a better approximation of the true value than the top coding threshold. Moreover, the total number of employees is maintained. On the other hand, the number of enterprises affected by the method necessarily increase (at least three for each stratum). This means that the protection level increases as well and it can be tailored by the group size and the number of groups in a stratum (for example for a given NACE the eight largest enterprises can be aggregated in two groups of four). Even the dominance rule can be easily incorporated in the method if required. Detection of outliers is avoided and consequently the implementation should be simpler.

In a sense this proposal can be seen as a microaggregation applied to the *top* (highest) values for the *pivot* variable number of employees, in the stratum. Whereas the other variables are protected by transformation of the pivot variable.

References

Defays, D. and Anwar, M. N. (1998) Masking microdata using micro-aggregation. *Journal of Official Statistics*, **14**, 449–461.

Domingo-Ferrer, J. and Mateo-Sanz, J. (2002) "Practical data-oriented microaggregation for statistical disclosure control". *IEEE Transactions on Knowledge and Data Engineering*, **14**, 189-201**.**