Working Paper No. 21
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

CELL SUPPRESSION IN EUROSTAT ON STRUCTURAL BUSINESS STATISTICS – AN EXAMPLE OF STATISTICAL DISCLOSURE CONTROL ON TABULAR DATA

Invited paper

Submitted by Eurostat[1]

[1] Prepared by Paul Feuvrier (paul.feuvrier@cec.eu.int) and Franca Faes-Cannito (franca.cannito@cec.eu.int).

Cell suppression in Eurostat on Structural Business Statistics - an example of Statistical Disclosure Control on tabular data

Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality
Luxembourg 7-9/4/2003

Paul Feuvrier
Eurostat Unit D2 Structural Business Statistics
Tel: (++352) 430 133 881
Paul.Feuvrier@cec.eu.int

Franca Faes-Cannito
Eurostat Unit D2 Structural Business Statistics
Tel: (++352) 430 133 394
Franca.Cannito@cec.eu.int

The opinions expressed in this document are those of the authors and do not necessarily reflect those of the European Commission.

Abstract

Statistical Disclosure Control is a key part of the data processing conducted in Eurostat by the Unit responsible for Structural Business Statistics. It has always been of high importance for it to perform a proper treatment creating the required uncertainty about the true value of the sensitive cells, while still preserving as much information in the table as possible. The paper shortly describes primary confidentiality rules currently implemented by Member States of the European Union in the field of annual business statistics. It also highlights Eurostat experience and expectations on various automated systems for Statistical Disclosure Control it has worked on: CIF (Confidentiality Interface), which is an interface of GHMITER, so far and Tau-Argus in the medium term.

Introduction

Annual Business Statistics consist of magnitude data (turnover value added, employment) combined with frequency data (number of enterprises, number of local units). Statistical confidentiality is a permanent problem in this area. First, because the data are broken down at a fine level of the classification (e.g. NACE Rev.1 4-digit level). Second, because the analysis often requires a breakdown of aggregates according to several dimensions at the same time (e.g. country, activity and size class). Third, because the distribution in business statistics is in most cases highly skewed, for instance in the car or the telecommunication industry. Dominant units often prevent the dissemination of the figure at national level (even if the aggregate can be disclosed at EU-15 level).

1 Confidentiality rules relevant for Structural Business Statistics

1.1 Primary confidentiality: (n,k) rule and p% rule

Bar some exceptions, National Statistical Institutes send data together with metadata to Eurostat, which performs both the primary and the secondary confidentiality on each data sets. Running primary confidentiality means implementing both the threshold and the dominance rule.

1.1.1 The (n,k) rule

As far as primary disclosure is concerned, the most traditional rule is the (n,k) rule, which can be described as follows:

"Regardless of the number of respondents in a cell, if a small number (n or fewer) of these respondents contribute to a large percentage (k percent or more) of the total cell value, the so-called n respondent, k percent rule of cell dominance defines this cell as sensitive."

1.1.2 The p% rule

Yet the (1,k) rule, say, does not consider the fact that the respondent with the second largest contribution can use his insider knowledge on his own contribution in order to obtain an estimate of the largest contribution, by subtracting his own contribution from the total aggregate value. Thus, some countries start to use the p% rule, where:

"A table cell is declared sensitive, if upper and lower estimates for the respondent's value are closer to the reported value than a given percentage p."

In the framework of the p% rule, one considers that the intruder is either the second largest contributor or more broadly a coalition of respondents pooling their data in an attempt to estimate the largest reported value.

Usually the second largest enterprise only is considered. However, one could assume for instance, as far as business statistics are concerned, that several affiliates of the same enterprise group can pool their data to estimate the largest contributor, in case it is not an affiliate of the group.

1.1.3 Is the p% rule to be promoted by Eurostat among Member States and Accession Countries as a primary confidentiality rule ?

It is widely recognised by researchers and specialists that the p% rule is more relevant than the (1,k) rule.

As for the (2,k) rule, it "incorporates" the p% one, provided that p = 100 * (100-k) / k. That is any cell sensitive according to the p% rule will be sensitive as well according to the (2,k) rule, but not the other way round. Some cells will be safe according to the p% rule but unsafe according to the corresponding (2,k) rule. To sum up, the (2,k) rule is to some respect more conservative than the p% rule.

Yet the (n,k) rule is still the most widely used in the area of Business Statistics among Member States and Accession Countries of the European Union. The technical format for SBS data transmission includes Metadata related to statistical confidentiality. Those metadata explicitly refer to the (n,k) rule, as the percentage dominance of the one or two enterprises which dominate the data is to be provided.

Eurostat might encourage the adoption of the p% rule by the Member States of the European Union in the medium or the long term.

2 Secondary confidentiality

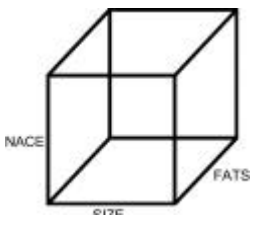
Statistical disclosure on SBS data in Eurostat has been mainly controlled through cell suppression, either by hand or with the help of an automated system.

The Eurostat Unit responsible for Structural Business Statistics has had a limited experience of perturbation-based methods so far.

2.1 Overlapping tables

A statistical disclosure control by hand is possible (though very time consuming) with less than three dimensions and against absolute disclosure only, i.e. without taking into account dominance.

Indeed, all SBS related data are broken down by activity, ie by NACE Rev.1. In the SBS size class series, data are broken down by activity and size class. In the FATS (FATS is for Foreign Affiliate Trade Statistics) series, data are broken down by activity and nationality of ownership (FATS). The instance below shows three dimensions.



The three tables (activity, activity * size class, activity * nationality of ownership, ie FATS) are intersecting through the NACE. Thus, the confidentiality treatment needs to be co-ordinated between the different tables. Otherwise one figure might be hidden in one table and not in another - and the protection would be broken. This operation, where confidentiality flags have to be synchronised between the tables on every intersecting dimension, is called table to table protection.

With such an instance (more than two dimensions and a necessary protection against approximate disclosure), one can assume that an automated system only can produce a protected data set where the body responsible for the data dissemination (International Organisations, National Statistical Institutes, …) is completely sure to withstand the risk of accidental disclosure.

2.2 SBS experience with GHMITER via CIF

Eurostat has been testing and using for statistical production GHMITER through CIF (Confidentiality Interface) since Autumn 2001. Indeed, the German engine met Eurostat requirements both in terms of treatment of dominance and table to table protection. GHMITER uses the rectangular parallelepiped method to smoothly protect data against absolute and approximate disclosure.

The hyper cube methods appeared to suit particularly well large tables or tables with more than two critical dimensions. Again, a protection of these tables by hand would be either impossible or very time consuming. To that respect, it has proved very relevant for Eurostat, as it was looking for a swift and efficient method.

Yet the "hyper cube criterion" is a sufficient but not necessary criterion for a safe suppression pattern. While testing and using the software, Eurostat pointed out significant over protection patterns. It seemed that, for particular sub tables it worked on, a better suppression pattern than the one suggested by GHMITER might have made a better solution.

2.3 Tau-Argus about to be tested on SBS data

A new disclosure package Tau-Argus (suitable for tabular data) should be tested in Eurostat in April 2003.

Eurostat relies on this new system to produce less over protective data sets, while still smoothly implementing the confidentiality rules. The strong point of Tau-Argus should be the calculation of an optimal set, where the hyper cube method is confronted to other techniques, e.g. the Hitas method, suppressing cells in hierarchical tables just like GHMITER.

While looking forward to testing Tau-Argus, Eurostat remains interested in ongoing works on SDC, in particular those conducted on the hyper cube method .

Not only Eurostat, but also many departments responsible for annual business statistics in many National Statistical Institutes proved increasingly interested in using a smooth software for SDC on tabular data.

2.4 Some requirements for an automated system for cell suppression relevant for annual business statistics

We document on annex some requirements provided to the CASC team (Computational Aspects for Statistical Confidentiality) in December 2002. These requirements concern modules that are currently not directly operational with CIF as well as some expected flexibility.

It has to be mentioned that, bar subsection A4 of the Annex, Tau-Argus should meet all those requirements.

Conclusion

The Eurostat Unit responsible for Structural Business Statistics will continue to test new systems for Statistical Disclosure Control in the future, in close collaboration with Member States. It looks forward, while still producing safe patterns, to limiting over protection problems. In addition, it might suggest the Member States a possible switch into more relevant primary confidentiality rules in the medium term. As far as confidentiality problems leading to the suppression of European totals are concerned, it has to be mentioned that the enlargement of the European Union should dramatically improve the situation, as the probability that a single country is confidential will automatically decrease.

ANNEX - Some requirements for an automated system for cell suppression relevant for annual business statistics

A1 - Waivers - possibility to disseminate singletons

The minimum number of respondents rule should be flexible to the following respect: some singletons are not confidential for those countries that managed, further to Eurostat requirements, to negotiate with the enterprises the dissemination of the cell. Current GH MITER does not accept such a situation and we would like to hear on progress to this respect with Tau-Argus.

A2 - More direct use of SBS metadata related to dominance rate

As far as dominance rules are concerned, Eurostat would like to use the information available in the SBS data set as such. Indeed, each dominant cell is associated to a dominance rate. We would like the software to take into account this dominance rate as such.

Let us take an example with the (1,k) rule.

Let assume one unit only dominates the cell. This corresponds to SBS flag B: one enterprise dominates the data.

The difference vis-à-vis the situation where a singleton has to be protected (SBS flag A) is that not the whole cell has to be protected with the given interval protection but p % of the cell only, that is the relevant dominant enterprise which represents x1 = p * x.

As has been previously emphasised, this metadata refers to the (n,k) rule and would need to be redefined, should the p% rule become a Eurostat standard.

A3 - A single treatment for several variables

A given table is to be protected for several variables (e.g. turnover, value added, employment, investment). Indeed, for many countries, the protection is to be conducted variable by variable (while for a second group of countries, the protection conducted for a group of variables, usually two or three, is applied to all other variables). For instance, the SBS series 2A with a breakdown by activity contains 28 variables each of them being theoretically to be protected an independent way. We would like the run a single treatment for all variables ! In the current version of CIF, the work needs to be done 28 times.

This should ideally be combined with the table to table protection. The series by activity and size class 2D contains 11 variables common to the series 2A.
To sum up, we should ideally be in a position to conduct in a single treatment the table to table protection on the common 11 variables together with the independent treatment on the 17 remaining ones of the series by activity only.

A4 - Table to table protection when the different data sets are inconsistent with each other

The table to table protection with GH MITER works smoothly only if both the number of units and the variable of the common dimension are exactly the same. This is unfortunately not always the case with actual SBS data ! For instance, the SBS series by activity is to be processed together with the one by activity and size class. It may happen that the data differ from 2% or 3% for a few cells, the other cells being consistent. Eurostat would like the software to accept this situation, that is we would like the software to conduct a table to table protection even in the consistency on the common dimension is not 100% perfect.

A5 - Confidentiality treatment when the number of statistical units is not available for each cell

The D2 production team currently works on a three dimension Business Statistics table. One of the dimension is a breakdown of turnover by product.

In the aggregated data set, each activity ACT 1 ACT2, etc… is broken down by product PA PB PC, etc… At the end we get the following Chart (this is a fictive example):

Turnover	Total	PA	PB	PC	PD
Total	59 000	11 000	20 000	6 000	22 000
ACT 1	15 000	10 000	5 000	0	0
ACT 2	15 000	0	10 000	5 000	0
ACT 3	25 000	0	2 000	1 000	22 000
ACT 4	4 000	1 000	3 000	0	0

The dimension "product" is indeed particular. As a matter of fact, each enterprise can theoretically produce all products, even if its production may be 0 for some products.
Thus, the difference of the dimension product vis-à-vis a classical dimension is that each cell theoretically represents all statistical units.

Number of statistical units	Total	PA	PB	PC	PD
Total	100	100	100	100	100
ACT 1	50	50	50	50	50
ACT 2	30	30	30	30	30
ACT 3	5	5	5	5	5
ACT 4	15	15	15	15	15

In the first chart, some cells may be primary confidential (with a dominance rate) and have to be protected. The problem is that it is currently not possible for us to conduct this treatment with our automated system. Each cell has indeed in CIF / GH MITER to correspond to a number of statistical units detailed in the input table. The total number of statistical units over each dimension has to correspond to the total of the dimension (otherwise the system is blocked). This is not possible for the dimension product !