QÜESTIIÓ, vol. 21, 1 i 2, p. 221-231, 1997

PROTECTING MICRO-DATA BY MICRO-AGGREGATION: THE EXPERIENCE IN EUROSTAT

DANIEL DEFAYS*

Eurostat

A natural strategy to protect the confidentiality of individual data is to aggregate them at the lowest possible level. Some studies realised in Eurostat on this topic will be presented: properties of classifications in clusters of fixed sizes, micro-aggregation as a generic method to protect the confidentiality of individual data, application to the Community Innovation Survey. The work performed in Eurostat will be put in line with other projects conducted at European level on the topic of statistical confidentiality.

Keywords: Confidentiality, clustering, micro-aggregation.

* Daniel Defays. Recherche et développement, méthodes et analyses des données. Eurostat. L-2920 Luxembourg.

– Article rebut l'octubre de 1996.

– Acceptat el febrer de 1997.

1. STATISTICAL CONFIDENTIALITY WITHIN EUROSTAT

There is a great concern at national and international level on statistical confidentiality. How to provide statistical information of high quality without disclosing confidential data? The respect of privacy has led national authorities to develop ad-hoc legislation which among others prevented national statistical institutes to transmit confidential data to Eurostat. The effect on the quality and completeness of European Statistics was disastrous. In most cases, it was impossible to provide European aggregates because some national data were missing. Facing this serious problem, Eurostat initiated beginning of the nineties, different actions.

- In the legal sphere, it proposed to the Council of the Union a regulation which sets the condition through the Commission for data transmission from national administrations to Eurostat. The regulation was adopted in 1990.
- Eurostat launched in parallel international seminars on confidentiality which brought together statisticians, academics and other officials and set the milestones for international co-operation in that area. The first seminar was organised in 1992 in Dublin in co-operation with the ISI. The next one was held in 1994 in Luxembourg and in 1996 it will take place in Bled (Slovenia).
- Eurostat organisation was updated in order to take into account the new constraint coming from the transmission of confidential data by the Member States. Administrative principle and procedures were established and technical measures taken in order to control the respect of those principles and procedures.
- Methodological work was also encouraged. Through the fourth European framework programme on research and development, financial support was given to a multinational team for the development of a software to control statistical disclosure of both microdata and tubular data (Waal, Willenborg, 1995).
- Eurostat made an inventory of existing methods and started in 93 to explore new techniques to protect micro-data by using micro-aggregates. The development of these techniques is the subject of this paper.

2. THE LEGISLATIVE FRAMEWORK

As already written, statistical confidentiality at the European Union level is governed by a Council Regulation (Euratom/EEC, 1588/90). This regulation was a response to a triple need:

– need of safeguards against possibilities of abusing data;

– need to allow the Member States of the Union to adopt a less restrictive data transmission policy regarding Eurostat;

– need for more and better Community statistics.

It makes it possible to remove national legal obstacles to the flow of statistical data from the national authorities to Eurostat. In that regulation, there is no European definition of confidentiality; confidential statistical data are defined as data declared confidential by the Member States in line with national legislation or practices. In most cases, precise rules exist for aggregated data: any cell of a table either containing data relative to less than three units, or dominated by one (or in some countries two) unit should not be disclosed.

For individual data, national policies are more different. In some countries, any transmission of micro-data is forbidden, in other ones only the research community has access under specific condition; part of micro-data can be publicly accessible in some countries.. Under this regulation, the use of confidential data transmitted to Eurostat is very limited. Dissemination of micro-data to scientists for statistical purposes is not allowed.

The investigation of Eurostat in the field of micro-aggregation was a response to the conflicting needs to give a maximum of information to the scientific community without disclosing confidential data.

3. BRIEF HISTORY OF MICRO-AGGREGATION AT EUROSTAT

The first idea was to start from the definition of confidentiality for tabular data: only cells with at least a minimum number of units and no dominance are not confidential. Applied to micro-data, this meant aggregation of units three by three, the units to be aggregated being as similar as possible in order to avoid dominance by one of them. Therefore we made the proposal to replace individual data by averages of small aggregates, which can play the role of fictive individuals on which data analysis could be performed (Defays, Nanopoulos, 1993). The problem was then to partition the whole population in cluster of fixed sizes. The averages of the optimal clusters could be transmitted. It was shown that if the primary data are points in Rp, the optimal partitioning (minimisation of within-group variances is characterised by the following property: every pair of clusters is separated by a hyperplane perpendicular to the line joining their barycenters.

To find the optimal partitioning seems to be a difficult problem. He have proposed to improve an algorithm proposed by Uri Hanani to cluster a set of points under an

equal groups constraint (Hanani, 1979). The problem was presented as a particular case of multicriteria dynamic clustering. Unfortunately there is no guarantee that that method will always reach the optimum. Another problem is the quality of the micro-aggregated data. It is easy to see that if the variables are not all correlated, the groups will be heterogeneous and the method will transform drastically some of the variables.

The difficulties to get the exact solution and the fact that in some cases the data can be strongly perturbated led us to investigate alternative methods, but in the same spirit.

4. A NEW MICRO-AGGREGATION TECHNIQUE

In order to avoid the loss of information caused by aggregation in clusters of fixed size of the original units, it has been proposed that the different unidimensional variables be aggregated separately, by ranking the values assumed by these variables and by an aggregation in fixed size groups of contiguous values.

The basic idea comes an application developed in the U.S. Internal Revenue Service (Strudler).

To illustrate the point lets take data for a set of 100 fictitious units covering three variables and aggregate into groups of 3, 5, and 10 to show the likely structural changes under different group sizes.

In a first step, the units are sorted in ascending (or descending) order of variable 1 and- grouped k by k (where k in our case was 3, 5 and 10). The original variable 1 value for each unit is then replaced by the average for variable 1 of the corresponding group. In next step the units are again sorted, but by variable 2 this time. Groups of k are formed and the original variable 2 values are replaced by the averages of the corresponding groups. This procedure —sorting, grouping, replacement with average values— is repeated for the third variable, and a new file is created consisting of 100 surrogate observations. Figures 1a to 1d show the degree of perturbation to be expected under different group sizes.

Set in this three dimensional space it is hard to tell apart the different figures, and as such this is the essence of the method. The method acts more on outlying observations while leaving the majority of the data structure intact; this property is both interesting and useful from a statistical and confidentiality viewpoint. The masking of extreme values is a prerequisite of any method purporting to safeguard confidentiality, yet a method which destroys data structure also destroys the statistical properties of the data. The method proposed both decreases risk of disclosure and maintains relationships.

Figures 1a-1d: Data transformation under different size classes

A better picture of the data transformations can be seen by using two dimensional plots as shown below with the arrows in figure 2a indicating the corresponding axis. It is easier to focus on the three plots in the top right hand quadrant of each figure. The three plots in the bottom left hand quadrant are exact mirror images of the former plots and care needs to be taken in their interpretation since the axis are also mirror transformations. The axis increase vertically upwards and horizontally to the right and decrease in value vertically downwards and horizontally to the left, see figure 2b.

Figures 2a-2d: Data transformation matrices

The figures above show the «grid» structure imposed by the method which becomes more pronounced as k is increased. This grid structure provides a guarantee

of confidentiality by creating observations with identical values on a single given dimension.

Tables 1 and 2 below present a range of statistics for the three variables in our example. The statistics summarise what can be seen from the figures above especially the reduction in the variance as the number of *k* is increased, however the method is mean invariant and the degree to which the summary statistics are altered is minimal. The figures for correlations are also stable and near the original.

Table 1. *Summary Statistics*

Original Statistics

Variable	Mean	Std Dev	Minimum	Maximum
VAR 1	10.24	6.83	1.00	48.00
VAR 2	2338.70	671.64	1127.00	4500.00
VAR 3	375.48	306.49	.00	1338.00

Modified Statistics $(k = 3)$

Variable	Mean	Std Dev	Minimum	Maximum
VAR 1	10.24	6.62	2.25	35.00
VAR 2	2338.70	669.50	1223.50	4307.00
VAR 3	375.48	305.28	.00	1216.67

Modified Statistics $(k = 5)$

Variable	Mean	Std Dev	Minimum	Maximum
VAR 1	10.24	6.52	2.40	31.00
VAR 2	2338.70	667.36	1243.20	4115.40
VAR 3	375.48	302.78	.00	1085.20

Modified Statistics $(k = 10)$

Variable	Mean	Std Dev	Minimum	Maximum
VAR 1	10.24	6.21	3.10	25.00
VAR 2	2338.70	654.27	1332.5	3753.30
VAR 3	375.48	298.61	.10	948.50

Table 2. *Correlations Matrices*

Original Correlations

	VAR 1	VAR 2	VAR 3
VAR 1	1.0000	.5770	.1641
VAR 2	.5770	1.0000	–.2344
VAR 3.	1641	–.2344	1.0000

Modified Correlations $(k = 3)$

	VAR 1	VAR 2	VAR 3
VAR 1	1.0000	.5516	.1544
VAR 2	.5516	1.0000	–.2183
VAR 3	.1544	–.2183	1.0000

Modified Correlations $(k = 5)$

	VAR 1	VAR 2	VAR 3
VAR 1	1.0000	.5521	.1213
VAR 2	.5521	1.0000	–.2274
VAR 3	.1213	–.2274	1.0000

Modified Correlations $(k = 10)$

	VAR 1	VAR 2	VAR 3
VAR 1	1.0000	.5516	.1226
VAR 2	.5516	1.0000	–.2374
VAR 3	.1226	–.2374	1.0000

5. VARIANTS OF THE METHOD

Some generalisations of this method were then proposed to tackle other types of variables and to generalise the replacement of original values by averages.

a) Segmentation of the set of variables

The micro-aggregation method presented above has been referred to in the reference documents as «micro-aggregation method by individual ranking». This term underlines the necessarily separate treatment of the different individual variables which results in separate classifications and aggregations of units as illustrated above.

But what is regarded as an individual variable in this context? A set of p variables can be treated as a single multivariate variable, resulting in a single grouping, or as separate p variables, resulting in p groupings each.

More generally, the initial vector of variables can be segmented into a number of variables, multivariate or univariate, which we called segments. Each segment is treated separately.

b) Characterisation of the groups

For each segment formed, units are regrouped and within each group, the values of the units are replaced, when the variable is quantitative, by an average. This again can be generalised. First, if the variables are not numeric, one can use other central values that the average, for instance, the median or the mode. Whatever is the choice of the central value to be used, the method will cause a diminution of the original dispersion of the variable. The variance of the microaggregated values will always be lower than the original variance for instance. To counter balance this effect, one has proposed to replace in each group the original values by new ones, taken from a distribution which as close as possible to the original one (in the group). With that logic, the values of a cluster are not identical anymore, but the original distribution is replaced by another one with similar characteristics.

c) Definition of the groups

In the original methods, groups all have the same size. One could imagine to use different size groups according to the type of variable, its sensitivity. Even for a given variable, one could below and above a given threshold use different size constraints. For instance, for company statistics, small enterprises could be grouped three by three whereas larger enterprises could be aggregated in larger groups.

d) Measure of the homogeneity of a group

When the segments are defined by one dimensional variables, the notion of homogeneity of the groups is easy to define. In the multivariate case, the concept of underlying similarity in the definition of groups leaves room for interesting variants; homogeneity can be measured in different ways: within group-variances, entropy or measure based on any type of distance.

6. APPLICATION TO THE COMMUNITY INNOVATION SURVEY

The generalised method was applied to the processing of the Community Innovation Survey (CIS). This survey combined both quantitative and qualitative data, key data on the enterprise which might permit indirect identification and more neutral subjective assessments, simple questions or questions with a more complex structure. The objective was to put at the disposal of research teams working on behalf of the Commission on Innovation topics a maximum of information from the CIS. Given the restriction put on Eurostat dissemination policy by the above mentioned regulation, we were not allowed to disclose to the scientists the original data. We thus decided to use micro-aggregates. The 50.000 company data transmitted to Eurostat were micro-aggregated by country and sector of activity. The quality of the protection of the individual data brought by the method was then checked by the countries and the perturbated data were sent to a limited number of contractors which analysed them. Results of these analyses were reported upon during an international conference on Innovation and its measurement held in Luxembourg in May 1996. Eurostat services have started to check on the original data the robustness of the results established with the micro-aggregated data. The conclusions reached so far are very encouraging. The distortions introduced by the method do not seem to affect the structure of the variables and their relations.

7. CONCLUSIONS

The conflicts between the needs for access to statistical information and demands for confidentiality will not disappear in the coming years. More powerful methods than the existing ones will have to be developed to protect the privacy and at the same type to disclose as much information as possible. Micro-aggregation is an attempt in that direction, unfortunately an empirical one. We have established the need for an such a transformation of the data and its feasibility in a specific case. More evidence of its robustness will have to be given. Properties and conditions of applications of its variants will have to be explored. This challenge is part of the mission of the official statisticians which is to provide the user with a high-quality statistical information service. Eurostat is determined to contribute to the advancement of ideas and techniques which will tackle these issues.

ACKNOWLEDGEMENT. The work presented in this paper is a collective effort at Eurostat level in which the author with many other colleagues was involved.

REFERENCES

[1] **Anwar, M. N.** (1993). «Micro-aggregation - The small Aggregates Methods». *Internal report*. Luxembourg, Eurostat.

[2] **Defays, D.** & **Anwar, M. N.** (1995). «Micro-aggregation: A Generic Method». *Proceedings of the 94 International Seminar on Statistical Confidentiality*. Luxembourg. Office for Official Publications of the European Communities, 1995.

[3] **Defays, D.** & **Nanopoulos, Ph.** (1993). «The Small Aggregates Method». *Proceedings of the 92 Symposium on «Design and Analysis of Longitudinal Surveys»*.. Ottawa, Statistics Canada.

[4] **Hanani, U.** (1979). «Multicriteria dynamic clustering». *Research report,* **358**, IRIA, Rocquencourt.

[5] **Strudler, M., Lock, H.** & **Scheuren, F.** «Protection of Taxpayer Confidentiality with respect to the Tax Model». *Washington, U.S. Internal Revenue Service*.

[6] **Waal, A.G.** & **Willenborg, L.C.R.J.** (1995). «Development of ARGUS: Past, Present, Future». *Proceedings of the 94 International Seminar on Statistical Confidentiality*. Luxembourg, Office for Official Publications of the European Communities.