Rounding methods for protecting EU-aggregates

Sarah Giessing[1], Anco Hundepool[2] and Jordi Castro[3]

[1] Statistisches Bundesamt, 65180 Wiesbaden, Germany, Email: Sarah.Giessing@destatis.de
[2] Statistics Netherlands, Division of Methodology and Quality, P.O. Box 4000, 2270 JM Voorburg, The Netherlands, Email: ahnl@cbs.nl
[3] Universitat Politecnica de Catalunya, Department of Statistics and Operations Research, Jordi Girona 1–3, 08034 Barcelona, Catalonia, Spain, Email: jordi.castro@upc.edu

Abstract. In the European Statistical System the statistical information is collected by the National Statistical Institutes (NSIs). The NSIs produce aggregate tables at the national level. They are also responsible for proper protection of these tables and hence they have to keep certain cells confidential, suppressing them from publications. Eurostat produces statistical information at the EU-level. However, the national suppressions hamper very much the publication of EU-aggregates although it is often only a few smaller countries having to keep their contribution to the EU-total confidential.
This paper reports on a research-project that aims for making more EU aggregates available whilst at the same time guaranteeing the national suppressed figures to remain confidential.

Keywords. EU-aggregates, Controlled rounding, Controlled tabular adjustment, Cell-suppression, Interval protection.

1 Introduction

The NSIs in Europe collect a lot of statistical information and publish many statistical tables at the national level or below. They are also responsible to take care of the confidentiality aspects of their publications. In quantitative tables this implies often that several cells have to be suppressed due to confidentiality reasons. Cell suppression is the traditional way of protecting a statistical table. See for example the CENEX-SDC handbook (Hundepool et al, 2006).

The NSIs also deliver data to Eurostat. Eurostat aggregates the national data to tables at the European level. In this paper we study the tables from the production statistics (Prodcom) and SBS (Structural Business Statistics). These tables are broken down by geography (down to the member state level) and in the case of the SBS data by a hierarchical NACE classification.

Confidentiality charters have been agreed with the Member States for the data collected in the respective frameworks of Prodcom and SBS Regulation. These charters describe amongst other issues when an EU-aggregate can be published, given the national published and sometimes suppressed cells. In many cases these rules prevent publication of EU-aggregates. If, for instance, only one country is confidential, the EU-aggregate must not be published, because otherwise this confidential value could be computed by taking the difference between the EU-aggregate and the non-confidential member state figures which is a typical instance of "disclosure by differencing".

Using certain constraints on the cell values of the tables which are known independent from the publication (like f.i. non-negativity of cell values) it is possible to compute *feasibility intervals* (a minimum and a maximum bound for the set of feasible values) for each suppressed cell of a publication, for instance by solving two linear programming (LP)-problems per cell. Any user of the publication would in principle be

able to perform such an analysis. A table is protected properly, if all the feasibility intervals satisfy certain requirements, e.g. if they create a certain amount of uncertainty about the true cell value. For discussion of these requirements see (CENEX-SDC handbook, section 4.2.2). Technically these requirements can be expressed as *protection intervals* which must be covered by the feasibility intervals.

As the confidential national cell (often a cell of a smaller country) frequently makes only a marginal contribution to the EU-aggregate, the corresponding protection interval although perhaps rather large at the national level is often only marginal at the EU-level. So, small confidential contributions with relative small protection interval impede the publication of much larger EU-aggregates. It should therefore be possible to 'save' the EU-aggregate by introducing a relatively small amount of uncertainty into it. This can be achieved by replacing the true value of aggregates by approximations like for instance rounded versions of the true value, or by replacing them by intervals or by adding some random perturbation to the aggregates. Approximations have to be determined as to provide sufficient protection to confidential aggregates. This implies for instance that the bounds of rounding intervals must be at safe distance from the true value of a confidential aggregate.

Publication of approximations makes sense of course only, if users understand well the difference (in terms of reliability) between the true and the approximated values. For general purpose data such as the European SBS aggregates, rounding approaches seem to be appealing because rounded figures are easy to interpret even by a naïve user.
In the remainder of this paper we will describe the solutions proposed for the Prodcom and SBS tables. Section 2 proposes controlled rounding for the protection of Prodcom data, whereas section 3 suggests another rounding method for the SBS data.

2 Rounding method for the Prodcom tables

The Prodcom tables have a rather simple structure. European Prodcom aggregates are reported only at the lowest level of the NACE hierarchy and therefore there are no higher level NACE-aggregates. Technically this reduces the large Prodcom table to a large set of smaller tables at the lowest NACE classification. Only some hierarchy in the geography has to be taken into account. For the tables from before 2003 the EU25 is broken down by EU15 and EU10, while for the more recent years EU27 is broken down by EU25 and EU2 (=Romania + Bulgaria).

The member states do the confidentiality protection for their tables themselves and decide which cells have to be suppressed. Because the higher level NACE codes are not published there is no additive relation between aggregates. Therefore, only primary suppressions have to be assigned. When transmitting the data to Eurostat, the member states also provide Eurostat with the cell values of confidential cells, but flag them as confidential. Also they provide the nature of this confidentiality. This can be an unsafe cell due to too few contributions (frequency rule) or due to a violation of a dominance or p% rule threshold. In case of a frequency unsafe cell the member states also report the number of respondents while for a dominance unsafe cell the percentage of the contribution of the largest or largest 2 contributors is given.

Although Eurostat cannot publish these unsafe cells at the member state level, it can use this information to compute the EU-aggregates. And if no member state information is

confidential or a sufficient number of member states is confidential, the EU-aggregate can still be published according to the rules of the Prodcom Confidentiality charter.

For those situations where the EU aggregate cannot be published, we propose a rounding procedure. Recently a controlled rounding procedure (c.f. Salazar-Gonzalez et al., 2006) developed on behalf of ONS was included in the statistical disclosure control software τ-ARGUS. Unlike traditional deterministic or probabilistic rounding methods, this controlled rounding method is able to guarantee that the special protection requirements of tabulations of establishment data are satisfied. For a given table with sensitive cells, the method computes the closest rounded table that is additive subject to certain constraints. These constraints ensure that the rounding interval for any confidential cell covers the corresponding protection interval.

The special procedure implemented for the Prodcom data first decides on the minimal rounding base, given the protection intervals of the confidential member states. These protection intervals are computed on the basis of the additional information of the unsafe cells supplied by the member states. Then the τ-ARGUS rounding procedure is applied. Sometimes the initial rounding base may not provide enough protection and then the rounding base will be increased.

Initially we had in mind to restrict the procedure to rounding bases as a powers of 10 (10, 100, 1000, …); procedure 1. However sometimes this resulted in rather large rounding bases and larger information loss. So a more refined series was then adopted (10, 20,…,90, 100, 200, 300,…,900, 1000, 2000,…); procedure 2. This led to solutions with enough protection, but less information loss. Only in the publication it requires a bit more explanation.

Of course the rounding procedure cannot hide the already published national safe figures. Before applying the rounding procedure, these safe, published cells have been merged into one cell. The exact value of these cell combinations has been considered to be known (as information available to a possible intruder) when stating the protection of the table as controlled rounding problem.

Rounding base (% of EU-total)	Frequency	
	Proc.1	Proc.2
0 -< 1	23	70
1-< 5	190	277
5-<10	82	109
10-<20	92	58
20-<50	107	15
Over 50	39	4

Table 1: Distribution of the rounding bases used

As can be seen from table 1 in many cases a solution can be found with only limited information loss. In the majority of cases the rounding base is less than 10 % of the EU total. Cases where the rounding base is larger than 50 % of a EU-total are a rare exception. As we cannot modify the already published tables the result of this procedure is, that the required protection interval is wide enough, but sometimes a bit shifted. Nevertheless the size of the interval guarantees enough protection.

3 Rounding method for the SBS tables

Because of its more complex data structures, the rounding procedure proposed for the Prodcom case cannot be expected to work well in the SBS case. Unlike in the Prodcom case, there is a detailed hierarchical relation between SBS aggregates, because they are published on the EU-level at 5 different levels of the NACE classification. The τ-ARGUS controlled rounding method rounds all aggregates of a table to multiples of one rounding base. While for the protection of large confidential aggregates at high NACE levels a large rounding base would have to be chosen, this kind of rounding would lead to too much information loss on the lower NACE levels.

In the following we propose a rounding procedure, which – just like controlled rounding –is able to guarantee that the specific protection requirements of magnitude tables from business surveys are satisfied, i.e. it provides enough uncertainty round each primary unsafe cell. The procedure comprises several tasks which will be explained in 3.1. The following section 3.2 outlines an alternative methodology based on interval protection. Section 3.3 reports some test results. Finally, section 3.4 describes some ideas for future work.

3.1 Rounding Procedure based on Restricted CTA

We first compute protection intervals, and bounds on the cell values assumed to be general knowledge, taking care in particular of those EU-aggregates where the confidential cluster consists of one or two member states with only one single contributor (so called '*singletons*') in either of these two states. In those cases we must avoid the risk for instance that one of two singleton companies can use special knowledge (e.g. of its own contribution) to undo the protection provided (by the rounding) to the other singleton company.

We then apply Restricted Controlled Tabular Adjustment proposed in (Castro and Giessing, 2006) to compute an *adjusted table* that contains some true, original and some approximate ('adjusted') values with the following properties: The adjusted table is, according to some suitable measure of distance, the closest additive table to the original table satisfying the following constraints:

- the adjusted values of *all* confidential cells are safely (considering the protection levels assigned as explained above) away from their original values,
- the adjusted values are within a certain range, i.e. for a variable with only non-negative values adjusted values also have to be non-negative, and
- adjusted values for member state aggregates flagged as published must be identical to the original value.

Note, that the last constraint makes the procedure what we call a *restricted* CTA procedure.

The next step of the procedure is to compute rounded approximations for those cells on the EU-level that were subject to an adjustment in the RCTA step. We chose a suitable rounding base for each cell separately from the series (10, 20, …, 90, 100, 200, …, 900, 1000, 2000, …, 9000…). For each adjusted value we determine the rounding base b to be the smallest in the series which is larger than the distance between the true and the adjusted value. This property guarantees that it is possible to find a multiple $m*b$ of the

base, where the rounding interval [(m-1)*b+1; (m+1)*b-1] covers both, the true, and the adjusted value.

So far, our procedure now guarantees *sliding* protection but not in all cases *upper* protection for the confidential aggregates: As explained in the CENEX-SDC Handbook, 4.2.2 users of a table with suppressions can always compute a feasibility interval for any particular suppressed cell, i.e. they can derive upper and lower bounds for its true value. We assume now that for this kind of analysis users who attempt to compute feasibility intervals for confidential member state level cells take into account the rounding intervals for the rounded EU-level cells as *a priori* bounds. Our procedure so far guarantees that either the upper or the lower feasibility bound that can be computed in this way for a confidential member state level cell will be safely away from the true value. Assume now the member state level cell was declared confidential because of dominance, e.g. because the true cell value is an upper bound for the contribution of the dominant respondent that is considered too close. If now in the case of this cell the lower feasibility bound is safely away from the true cell value, but the upper feasibility bound is not, this means that - like the true cell value - the upper feasibility bound is an upper bound for this respondent contribution which is too close (this is our definition of 'not safely away'). Note that this is only a problem, if the reported variable takes only non-negative values, because otherwise it may happen that individual respondent contributions are larger than the cell value.

We can solve this disclosure risk problem for variables with non-negative values heuristically by extending the procedure in the following way: We first audit the rounding obtained so far by computing feasibility intervals considering the rounding intervals as just explained. If this audit establishes lack of upper protection we carry out an extra one-cell CTA procedure (one for each confidential cell lacking upper protection). One-cell CTA, targeted to one specific confidential cell, addresses upper protection of only this particular cell, i.e. it guarantees that the adjusted value of this cell will be larger than the true value, and safely away from it. The procedure is completed by another rounding step. This time we determine rounding bases b for the European level aggregates so that the corresponding rounding interval covers the smallest interval that contains the true value, the adjusted value from the RCTA procedure, and all the adjusted values from each of the one-cell CTA procedures. This extended procedure obviously guarantees sufficient upper protection of confidential member state level aggregates.

While the combination of restricted CTA and one-cell CTA guarantees sufficient upper protection of confidential member state cells, it does not guarantee that the set of intervals that is used to compute the rounding bases is the 'best' set of intervals satisfying our requirement. Optimal solutions for this problem could be achieved using a formulation as a large-scale linear optimization problem as outlined in the following section.

3.2 Outline of an interval protection methodology

We are given a table (i.e., a set of cells $a_i, i = 1,\ldots,n$, satisfying m linear relations $Aa = b, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$). Any set of values x satisfying $Ax = b, l \le x \le u$, is a valid table, $l \in \mathbb{R}^n, u \in \mathbb{R}^n$ being known a priori lower and upper bounds for cell values. For positive tables we have $l_i = 0, a_i = +\infty, i = 1,\ldots,n$, but the procedure outlined is also valid for general tables.

Our purpose is to compute the set of smallest intervals $\left[lb_h, ub_h\right]$ for cells $h \in \mathrm{H}$ (in our instance, H is the set of EU-level cells) instead of the real value $a_h \in \left[lb_h, ub_h\right]$, such that, from these intervals, no attacker can determine that $a_s \in \left(a_s - lpl_s, a_s + upl_s\right)$ for all sensitive cells $s \in \mathrm{S}$. Introducing two auxiliary vectors $x^{l,s} \in \mathbb{R}^n$ and $x^{u,s} \in \mathbb{R}^n$ to impose, respectively, the lower and upper protection requirement, this problem can be stated as follows:

$$\min \quad \sum_{i \in \mathrm{H}} w_i \left(ub_i - lb_i\right)$$

$$\text{s.t.} \quad \left.\begin{array}{rcccll} & & Ax^{l,s} & = & b & \\ & l \le & x^{l,s} & \le & u & \\ 0 \le lb_i \le & & x_i^{l,s} & \le & ub_i & i \in \mathrm{H} \\ & & x_s^{l,s} & \le & a_s - lpl_s & \\ & & & & & \\ & & Ax^{u,s} & = & b & \\ & l \le & x^{u,s} & \le & u & \\ 0 \le lb_i \le & & x_i^{u,s} & \le & ub_i & i \in \mathrm{H} \\ & & x_s^{u,s} & \ge & a_s + upl_s & \end{array}\right\} \forall\ s \in \mathrm{S}$$

where w_i is a weight for the information loss associated with cell a_i.

Indeed this problem, in theory, it is simpler than optimal CTA, so it may be more efficient and provide a better solution than the procedure based on CTA, plus a post-process with one-cell CTA for unprotected cells.

3.3 Test results

The structure of the SBS test tables is 2-dimensional (by NACE and by country). Because Eurostat does not publish an overall cross-sectoral total, each table corresponds to only one particular NACE sector. Rounded approximations had to be computed for tabulations of two different variables. Variable 1 takes non-negative values only while variable 2 may also take negative values.

The procedure of 3.1 was applied to these data (so far only to the variable 1 tabulations). Only in the case of sector D we had to carry out one-cell CTA post-processing.

In the following, we present results obtained for tabulations of variable 1 for NACE sectors C, D and E. As an indicator for the loss of information caused by rounding a particular cell we use the percentage of the rounding base (in terms of the true value of the cell). The largest perturbations we observed were about 17 % in sectors C and D, and about 0.6 % in sector E. Table 1 presents the number of EU-level cells by range of these percentages.

Rounding base (in % of the cell value)	NACE-sector			
	C	D	E	C-E
0 %	16	259	1	276
(0%, 2%]	13	104	4	121
(2% , 5%]	0	4	0	4
(5%,10%]	3	0	0	3
> 10%	5	1	0	6

Table 1 No. of EU-level cells by rounding base percentage ranges for NACE sectors C, D and E

Overall, in the three sectors C, D and E 276 cells remained unperturbed (rounding base percentage 0%). Nearly half as much (121) were perturbed by less than 2 %. Only a few cells got larger perturbations.
Table 2 presents the results with respect to the hierarchical level of the cells in the table. It shows the distribution of cells (no of cells in %) by ranges of the perturbation percentages and by NACE level.

NACE-level	Rounding base (in % of the cell value)				
	0	(0%, 2%]	(2% , 5%]	(5%,10%]	> 10%
	Sector C				
4-digit	50	25.00	-	12.50	12.50
3-digit	46.15	30.77	-	7.69	15.38
2-digit	20	60	-	-	20
sub-sector	50	50	-	-	-
sector	-	100	-	-	-
	Sector D				
4-digit	80.18	18.94	0.44	-	0.44
3-digit	56.31	41.75	1.94	-	-
2-digit	30.43	65.22	4.35	-	-
sub-sector	78.57	21.43	-	-	-
sector	100	-	-	-	-
	Sector E				
4-digit	-	100	-	-	-
3-digit	-	100	-	-	-
2-digit	50	50	-	-	-
sector	-	100	-	-	-

Table 2 No. of EU-level cells (in %) by rounding base percentage range and by hierarchical level for NACE sectors C, D and E

Table 2 shows that all throughout the sector and sub-sector level cells remained unperturbed or were perturbed by less than 2 %. Stronger perturbations were observed only on the lower levels of NACE sectors D and C. While in sector D perturbations

beyond 5 % were very rare, and were observed only on the 4-digit level, larger perturbations (i.e. more than 2 % of the cell value) were observed more frequently in the C-sector. In this sector, 20 to 25 % of the cells below the sub-sector level were perturbed by more than 5 % . This means, on the other hand, that even on the lower NACE levels of the C-sector about 80 to 85 % of the cells were perturbed by less than 5 %, which is quite a positive result for that sector with serious dominance problems where 297 of the 925 Member State cells are flagged confidential.

For the purpose of comparison we have also computed a cell suppression pattern for the sector D table using the τ-ARGUS modular optimization method for secondary cell suppression. In principle – to avoid certain risks of underprotection – the method should be applied to the full table, including the member state level cells. In practice, however, this is not feasible. The fact that the suppression pattern for the member state cells must not be changed leads to infeasibility problems. Therefore the original 2-dimensional (by NACE and member states) cell suppression problem was relaxed and turned into a 1-dimensional problem, addressing only the selection of secondary suppressions on the European level. Primary suppressions on the European level and the corresponding protection levels were identified on the basis of the rules of the SBS confidentiality charter. As a result we got 27 secondary suppressions protecting 25 primary suppressions, e.g. 52 suppressed cells. Obviously, the rounding affected a lot more cells (109). On the other hand, the information loss resulting from rounding a cell is certainly less than from suppressing that cell.

Table 3 below compares the cell suppression result for the D-sector tabulation with the rounding result using three alternative measures of information loss for rounded cells, and two for suppressed cells. For suppressed cells, the first information loss measure is a simple count of the suppressed cells. The second, more sophisticated measure is based on a computation of the feasibility interval for each of the suppressed cells. It considers the size of this interval as measure for the information loss. For the computation of the feasibility intervals we have taken into account as a lower *a priori* bound (i.e. a bound known to data users) for each suppressed EU-level cell the sum over the corresponding published (e.g. non-confidential) member state cells.

For rounded cells, the first measure is a simple count of the number of rounded cells, the second measure is a count of rounded cells where the rounding base exceeds 2 % of the cell value, and the third one considers the size of the rounding interval as information loss for a rounded cell.

NACE-level	Cell Suppression		Rounding		
	# sup-pressed	∑(size of feasibility intervals) (in tsd.)	# rounded	# rounded by more than 2%	2 ∑(size of rounding bases) (in tsd.)
4-digit	19	1920	45	1	581
3-digit	20	1961	45	2	579
2-digit	11	1431	16	1	405
sub-sector	2	18	3	0	12
total	52	5330	109	4	1578

Table 3 Information loss of the cell suppression result for NACE sector D tabulation compared to information loss of rounding result

Table 3 shows that much less cells were rounded by more than 2 % than suppressed (4 vs. 52, over all NACE levels). The impression that rounding outperforms cell suppression in this instance is also confirmed by the more sophisticated evaluation of feasibility interval sizes for suppressed cells (5330 tsd. in total), vs. rounding intervals for rounded cells (1578 tsd. in total).

3.4 Future work

There are some methodological aspects which have not yet been considered closely so far, but will need some attention in the future.

Linked tables: Eurostat also publishes tabulations of variables 1 and 2 by NACE and size class. These 3-dimensional tables have of course cells in common with the 2-dimensional tables we studied so far creating a linked-tables problem. Consequently, the current approach would have to be extended as to guarantee the use of identical rounded approximations for identical aggregates between tables.

One option to solve this problem could be joining linked tables into a single big 'table', and to solve the resulting large optimization problem. This is the only way to guarantee a feasible and good (or optimal) solution. It is not yet clear, however, how 'expensive' (in terms of computer resource requirements) and how efficient this solution would be.

Alternatively one could try an iterative so called 'coordinate descent' approach. In such an attempt we would first compute a solution for the first table, and then, considering these results, compute a solution for the second table. This will have to be repeated until some stage of convergence has been reached.

Related tables and time series: There are pairs of variables, for which Eurostat also intends to publish the ratio of the two. Computation of approximations of these ratios as ratio of the rounded approximations obtained by our procedure is of course straightforward, but has not yet been done for the test data sets. Afterwards, intervals for these ratios (given the rounding intervals of the corresponding enumerator and denominator indicator) will have to be computed and examined. If it turns out that those intervals are too large, i.e. the quality of the approximation for the ratio is too low, it could be considered to develop a more advanced procedure. The objective of the advanced procedure could for instance be increasing the likelihood, that if the rounded approximation of one indicator is smaller than its true value, the rounded approximation of the other indicator will also be smaller than its true value, e.g. that both approximations perturb the true value in the same direction. A similar approach could be attempted in order to improve the behaviour of the rounding method when applied to a time series of tabulations of a variable.

Interval protection methodology: Because of its advantages in terms of efficiency as mentioned above, it might be worth to fully develop the alternative interval protection methodology outlined in section 3.2.

4 Conclusions

We have proposed and tested rounding methodology for disclosure control of European aggregates of the ProdCom statistics and Structural Business Statistics (SBS). The procedure suggested for ProdCom data is based on the τ-ARGUS rounding procedure. Because of its more complex data structures, this procedure cannot be expected to work well in the SBS case. For the SBS data, we have therefore developed a rounding procedure based on restricted controlled tabular adjustment. For the special case of strictly non-negative tables we also outlined an alternative method based on interval protection instead of controlled tabular adjustment.

Both rounding procedures gave promising results. In the Prodcom case the majority of EU-level cells were perturbed by at most 10 % of the EU total. In the SBS case about 95 % of the cells remained unperturbed or were perturbed by at most 2 %. We also provided some evidence that the rounding in this case outperforms cell suppression.

With respect to the SBS data set, some aspects need further attention. Before the method can be used for production, the methodology has to be extended to be applicable to sets of linked tables. Another issue that should be addressed in future research is how to improve the behaviour of the method in a situation where we want to preserve to some extent the correlation between tabulations of different variables, or of tabulations of a time series of a variable. Finally, it might also be interesting to implement the alternative method based on interval protection and compare its behaviour to the current procedure.

Acknowledgements

This research was financed by the European Commission under two specific contracts (No. 22100.2006.002-2006-796 and No. 22100.2006.002-2006-795).

References

Castro, J., Giessing S. (2006). Testing variants of minimum distance controlled tabular adjustment, in *Monographs of Official Statistics*. Work session on Statistical Data Confidentiality, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, pp. 333-343

Hundepool., Anco,. Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Rainer Lenz, Jane Longhurst, Eric Schulte Nordholt, Giovanni Seri, Peter-Paul De Wolf (2006), *CENEX handbook on Statistical Disclosure Control,* CENEX-SDC project, http://neon.vb.cbs.nl/cenex/CENEX-SDC_Handbook.pdf

Hundepool, Anco, et al (2006), *τ-ARGUS manual Version 3.2*, Voorburg. The Netherlands, http://neon.vb.cbs.nl/casc/Software/TauManualV3.2.pdf

Salazar-Gonzales, J.J., Bycroft, C., Staggemeier, A.T. (2006). The controlled rounding implementation, in *Monographs of Official Statistics*. Work session on Statistical Data Confidentiality, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, pp. 303-308