ESSNet S D C

A Network of Excellence in the European Statistical System in the field of
Statistical Disclosure Control

Protection of tables with negative values

Sarah Giessing

1 Introduction

In the current implementation, the SDC software package τ-ARGUS does not allow to carry out tabular SDC on a variable that takes both, positive and negative values. In this paper we propose methodology that could be implemented into the software in order to handle such data. The focus is on the suggestion of simple heuristic techniques that can be implemented into the software in the course and with the very limited resources of the ESSNET-SDC project.
In section 2.1 we will first describe methodology to assess cell sensitivity[1]. Section 2.2 will consider issues of secondary cell suppression in this context.

2 How to handle tables with positive and negative contributions to cell values?

In the following we suggest heuristic strategies to handle the problem of SDC for a tabulation of a variable taking both, positive and negative values. In section 2.1 we propose two alternative concepts for how to adapt the common concentration rules for assessing cell sensitivity to that case.
In section 2.2 we propose how to specify the secondary cell suppression problems for this case in such a way that it can be solved by current implementations of algorithms for secondary cell suppression available in τ-ARGUS.

2.1 Assessment of cell sensitivity

In the following we propose two alternative strategies for how to adapt the common concentration rules for assessing cell sensitivity to the case when disclosure risk has to be assessed for a variable that can take not only positive, but also negative values. The first strategy is the one proposed in the CENEX SDC-handbook. The second approach is based on the idea to protect – instead of the original table - a tabulation of a

[1] Note that this methodology is already mentioned in the CENEX SDC-handbook[].

transformed version of the variable. The transformed variable should take only positive values. The advantage of the latter approach is that the secondary cell suppression problem can be defined for the tabulation of the transformed (positive) variable in the usual way.

Strategy 1: Relaxing the parameters of the concentration rules

This strategy basically comes up to a reduction of the value of *p* (or increase of the parameter *k*, for the dominance-rule, resp.). It may even seem adequate to take that reduction to the extent of replacing a concentration rule by a minimum frequency rule. This recommendation is motivated by the following considerations:

A well known extension of the p %–rule is the so called prior-posterior (p,q)%-rule. With the extended rule, one can formally account for general knowledge about individual contributions assumed to be around prior to the publication, in particular that the second largest contributor can estimate the smaller contributions $X_R := \sum_{i>2} x_1$ to within q %.

Technically, any (p,q)-rule with q < 100 can also be expressed as $(p*,q*)$-rule, with $q*=100$, e.g. choose

$$p^* =: 100 \cdot p/q \qquad (1)$$

When contributions may take negative, as well as positive values, it makes sense to assume that bounds q_- for the relative deviation of *a priori* estimates $\hat{X}_R$ exceed 100 %. This can be expressed as $q_- = f*100$, with $f>1$.

According to (1) the *p* parameter p_f for the case of negative/positive data should be chosen as $p_f = 100*p/q_- = 100*p/(f*100) = p/f < p$. Hence for large *f* the p%-rule with corresponding parameter p_f is asymptotically equal to the minimum frequency rule.

The CENEX SDC-handbook (sec. 4.2.1) also points out that in the same way, any (*n*,*k*)-dominance rule could be adapted to account for *q* % relative *a priori* bounds. In such a case one would use

$$(100-k^*)/100 = (100-k)/q \qquad (2)$$

in order to determine a suitable parameter *k** for the adapted rule.

Because of (2) then, a parameter k_f suitable to account for *a priori* bounds $q_- = f*100$ can be determined by $(100-k_f)/100 = (100-k)/(100\cdot f)$. As $f>1$ this means that $k_f > k$. For large *f*, dominance rules with parameter k_f will be asymptotically equal to minimum frequency rules.

Hence, according to proposal, regarding the assessment of primary cell sensitivity, no extension of the ARGUS software is necessary. Any adaptations of the methodology as recommended above can be easily performed by the users of the software themselves when they select a suitable rule for primary suppression.

Strategy 2: Transformation of the microdata

For this approach, we assume that – because of some *a priori* information available to the users of the table – each individual observation x_j *(j=1,..,N)* of an original variable ***x*** taking positive and negative values is known to be at least of size *M*, where *M* is a large negative constant, i.e. $x_j \geq M$ for *j=1,…,N*.

Let now ***y*** denote a simple linear transformation of ***x***, i.e. $y_j := x_j + M$. Then ***y*** is a positive variable.

For any group C of *n* observations it holds:

$\sum_{i \in C} y_i = n \cdot M + \sum_{i \in C} x_i$. Thus, if we publish a cell value $X_C := \sum_{i \in C} x_i$ and users know (or can estimate) the number *n* of observations in C, they can easily compute the cell value for the transformed variable

(1) $Y_C := n \cdot M + X_C$

It makes sense then to define disclosure (and in particular: inferential disclosure) for variable ***x*** as equivalent to disclosure of the transformed variable ***y***. Because ***y*** is positive, the usual methods for assessing cell sensitivity (c.f. 4.2.1 of the CENEX-SDC handbook) apply.

Example: p %-rule

According to the p %-rule, a cell value Y_c is sensitive, if (and only if) $p/100 \cdot y_1 - (Y - y_1 - y_2) > 0$. This is equivalent to

$p/100 \cdot (x_1 + M) - ((X + n \cdot M) - (x_1 + M) - (x_2 + M)) > 0$ and (for n>2) to

$p/100 \cdot (x_1) - (X - x_1 - x_2) - M \cdot ((n-2) - p/100) > 0$.

Because of the negative term $M \cdot ((n-2) - p/100)$, even if a cell C involves only positive observations for variable ***x***, and if the cell value X_c would have been sensitive according to the p %-rule, according to the definition based on cell sensitivity of the transformed variable ***y***, it will often not be sensitive.

2.2 Issues of secondary cell suppression

Several alternative algorithms for the selection of secondary suppressions are available in τ-ARGUS, e.g. the Hypercube/GHMiter method, a Modular and a Full optimal solution, and a method based on a network flow algorithm. In the following, we discuss how to specify secondary cell suppression problems for tables containing both, positive and negative values, in such a way that they can be handled by the current software implementation of those algorithms.

Fischetti and Salazar (2000) give a mathematical formulation of the secondary cell suppression problem (CSP) as mixed integer linear programming problem. This formulation is a general one, in the sense that neither the cell values nor external *a priori* bounds on the cell values are required to be positive. For practical reasons however, the algorithm implemented to solve this problem (e.g. the "Full optimal solution" of τ-ARGUS), assumes all cell values to be positive. We have the same problem also with the implementation of the τ-ARGUS network flow algorithm, and of course with the Modular method (calling the "Full optimal" algorithm as a sub-routine) too. Changing these implementation might be feasible, theoretically. It is however out of scope for the limited resources of the ESSNET project. It should be mentioned in this context that also the 'clean' formulation of the CSP is often 'heuristic' in itself, because it involves fixing external *a priori* bounds on each cell value. And those will often have to be guessed.

Hence, the simple heuristic method we propose in the following may even be considered a fully appropriate approximation. The idea is a straightforward extension of strategy 2 of section 2.1. Instead of solving the secondary cell suppression problem for a tabulation of the original variable ***x*** taking positive and negative cell values, we solve it for the tabulation of the transformed variable ***y***. For this transformed (positive) variable protection levels can be computed as usual (c.f. example 5 of section 4.2.2 of the SDC-handbook for an example relating to the p %-rule).

As explained in 2.1 above, we assume that users given with a tabulation of ***x*** could estimate the corresponding tabulation of ***y*** using (1). They could then compute for any suppressed cell in the table of ***y*** a feasibility interval (assuming ***y*** to be positive). Let U denote the distance between a bound of a feasibility interval and the corresponding true, sensitive cell value $Y\ (=n\cdot M+X)$. If this distance satisfies the usual constraints imposed by the secondary cell suppression algorithm (like, e.g. for the p %-rule $U \geq p/100\cdot y_1-(Y-y_1-y_2)$) then it also satisfies the corresponding constraints formulated in terms of variable ***x*** (like, e.g. $U \geq p/100\cdot(x_1+M)-\big((X+n\cdot M)-(x_1+M)-(x_2+M)\big)$, c.f. example at the end of 2.1 above).

Note that as usual we do not need a full micro-data set to carry out disclosure control for a table according to this method. It is enough, if cell values X, the largest contributions (like, for the p %–rule x_1 and x_2), the number n of respondents for each cell, and the *a priori* bound M are supplied[2].

[2] If the largest contributions are not provided, then (as always) we can only protect against exact disclosure, or determine fixed percentages as protection ranges for sensitive cells. If also information on the number of respondents is lacking, the method can still be implemented in a similar way, considering the lowest level cells as individual

In order to avoid very large cell values (in the table of the transformed variable), which sometimes causes problems with the optimization routines in the LP based secondary cell suppression algorithms, it is suggested to round the ***y*** observations in advance of processing. A suitable rounding base should be chosen so that the smallest ***y*** value is still rounded to at least one (instead of zero).

3 Conclusion and final remarks

This report has proposed methodology to handle cell suppression problems of table that contain positive, as well as negative cell values. Using the suggested concept, cell suppression problems can be formulated that can be handled by each of the secondary cell suppression algorithms provided by τ-ARGUS.
With respect to the GHMiter/hypercube method, there is also an alternative option: the GHMiter program offers a variant to handle positive/negative tables. Control statements to run this variant are already implemented in τ-ARGUS. Note that this variant of GHMiter does only provide protection against exact disclosure. This is a sensible concept, considering the remark in section 2.1 (…it may even seem adequate to … replace …. a concentration rule by a minimum frequency rule…).

References

Fischetti, M, Salazar Gonzales, J.J. (2000), *'Models and Algorithms for Optimizing Cell Suppression Problem in Tabular Data with Linear Constraints'*, in Journal of the American Statistical Association, Vol. 95, pp 916

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte-Nordholt, E., Seri, G., de Wolf, P.P. (2006) *CENEX SDC Handbook on Statistical Disclosure Control*, version 1.01, available at http://neon.vb.cbs.nl/cenex/

respondents and thus adding M (instead of nM) to such a lowest level cell, and then completing the table (e.g. the higher level cells) consistently.