CASC PROJECT

Computational Aspects of Statistical Confidentiality

June 2004

Methodological Background of Cell Suppression in τ-ARGUS: an Introduction for the Practitioner

Author: Sarah Giessing

Institute: Federal Statistical Office of Germany

Deliverable No: 5-D6-2

June 2004

Methodological Background of Cell Suppression in τ-ARGUS: an Introduction for the Practitioner

Deliverable 5-D6
Part 2

Sarah GIESSING,
Federal Statistical Office of Germany
65180 Wiesbaden
E-mail: sarah.giessing@statistik-bund.de

1 Introduction

At first sight, one might find it difficult to understand that information presented in tabular form presents a disclosure risk at all. After all, one might say that the information is presented in aggregate form. However, in many cases, cells of tables indeed relate to single or very few respondents only, and this implies a disclosure risk for the individual data of those respondents. If it has been established that a disclosure risk would be connected to the data set, users of τ-ARGUS can apply cell suppression in order to protect those data. Cell suppression comprises two steps: In the first step the disclosure risk connected to each individual cell of the tables is assessed. Concepts used in τ-ARGUS for assessment of disclosure risk for tabular data will be explained in section 2. Cells are suppressed when they reveal too much information on individual respondent data. In the second step, in order to prevent these so called "primary suppressions", or "sensitive" cells, from exact disclosure or from being closely estimable from the additive relationship between the cells of the table, additional cells (so called "secondary" or "complementary" suppressions) must be suppressed. This second step is called "secondary cell suppression". Section 3 considers aspects of secondary cell suppression in theory and practice, discussing issues of information loss, as well as table design aspects. Section 4 provides a brief overview on alternative algorithms for secondary cell suppression in τ-ARGUS.

2 Disclosure Risk

In this section the concepts for assessment and control of disclosure risk for magnitude tables of τ-ARGUS will be explained. Section 2.1 is concerned with the kind of disclosure that may happen when disclosure control on the cell level is lacking.

From the linear relations between published and suppressed cell values, data users could derive bounds for the suppressed cell entries, and thus estimates of respondent data for respondents contributing to confidential cells. This kind of disclosure risk will be referred to as 'table level disclosure risk' and is discussed in section 2.2 .

2.1 Sensitive Cells in Magnitude Tables

τ-ARGUS offers several *safety rules* (also referred to as '*sensitivity measures*', or '*sensitivity rules*') as measures to assess the disclosure risk connected with release of a certain aggregate (or cell) within a table. Choice of a particular safety rule is usually based on certain intruder scenarios

(which involve assumptions about additional knowledge available in public or to particular users of the data) and on some (intuitive) notion on the sensitivity of the variable involved.

- *Intruder scenarios:* With business data, it is usually assumed, that the "intruders", those who might be interested in disclosing individual respondent data may be "other players in the field", e.g. competitors of the respondent or other parties who are generally well informed on the situation in the part of the economy, to which the particular cell relates. It is assumed specifically, that the intruders are able to identify the largest contributors to a cell. The commonly applied sensitivity rules differ in the particular kind and precision of additional knowledge assumed to be around.
- *Notion on the sensitivity of the variable:* Some safety rules protect against exact disclosure of individual data only, while others, used only with magnitude data, go further, and protect the data from approximate disclosure.

An aggregate (or: cell in a table) that is indeed '*unsafe*', or '*sensitive*' according to the safety rule employed, is subject to what is called '*primary suppression*'.

Sensitivity Rules to prevent exact disclosure When it is enough to prevent exact disclosure of respondent data, users of τ–ARGUS specify the parameter *N* of a *minimum frequency rule*. A cell with at least as many respondents as this minimum frequency is considered safe. Normally this parameter will be set to 3, except when it is realistic to assume that groups of *N*-1 ($N > 2$) respondents contributing to the same cell pool their data to disclose the contribution of another respondent.

When it is not enough to prevent exact disclosure, but the risk of approximate disclosure must also be limited, a sensitivity rule to prevent approximate disclosure must be specified.

Sensitivity Rules to prevent approximate disclosure When a particular variable is deemed strongly confidential, preventing only exact disclosure may be judged inadequate. We may also wish to prevent an intruder from deducing too precise an estimate. This is a risk whenever an aggregate is predominated by very few contributions. τ-ARGUS offers two types of concentration rules to prevent this kind of disclosure: the 'N respondent, k percent'-*dominance rule*, where *N* refers to a number of respondents, and *k* is a percentage threshold for the total of the *N* largest contributions in relation to the cell total, and the so called *(p,q)-rule*. The following section clarifies the concept of safety rules using simple examples for illustration. In a more general way, the sensitivity rules given below can be represented as "upper linear sensitivity measures". For definition and mathematical properties of linear sensitivity measures in general see [2].

The simplest concentration rule is the (1,k)-rule:

(1,k)-rule

According to the (1,k)-rule, an aggregate is identified to be unsafe, whenever the largest contribution x_1 is greater than k percent of the total aggregate value X , e.g. when $x_1 > \frac{k}{100} X$.

This will make sure, that in any non-sensitive aggregate the largest contribution x_1 will be k% of the aggregate (e.g. table cell) value X at most:

An intruder, estimating the largest contribution to be $\hat{x}_1 = X$, will overestimate x_1 by at least (100-k) % of the aggregate value X, and by at least $100 \cdot \frac{100-k}{k}\%$ of x_1 itself, as can be proven as follows:

$$x_1 \le \frac{k}{100}X \Leftrightarrow \frac{\hat{x}_1}{x_1} \ge \frac{100}{k} \Leftrightarrow \frac{\hat{x}_1 - x_1}{x_1} \cdot 100 \ge \left(\frac{100}{k} - 1\right) \cdot 100$$

Example 1:

Application of the (1,90)-rule.

Let the total value of a table cell be $X = 100{,}000$.

Let the largest contribution be $x_1 = 90{,}000$.

90,000 ≤ (90/100)*100,000, so according to the (1,90)-rule the cell is safe – no risk of disclosure.

The upper estimate for the largest contribution $\hat{x}_1 = 100{,}000$, will overestimate x_1 by 11.1 % of x_1: $\frac{\hat{x}_1 - x_1}{x_1} \cdot 100 = \left(\frac{100}{k} - 1\right) \cdot 100 = 11.1$.

Often however, the second largest contributor is able to derive a much more precise upper estimate of the largest contribution, by subtracting his own contribution x_2 from the aggregate total X ($\hat{x}_1 = X - x_2$). An example is given below.

Example 2:

Application of the (1,90)-rule.

Let the total value of a table cell be $X = 100{,}000$.

Let the largest contribution be $x_1 = 50{,}000$.

Let the second largest contribution be $x_2 = 49{,}000$.

50,000 < (90/100)* 100,000 , so according to the (1,90)-rule the cell is safe – there seems to be no risk of disclosure. But $\hat{x}_1 = 100{,}000 - 49{,}000 = 51{,}000$,

hence $100 \cdot \frac{\hat{x}_1 - x_1}{x_1} = 100 \cdot \frac{51{,}000 - 50{,}000}{50{,}000} = 2$.

So here, the second largest contributor is able to derive an upper estimate for the largest contribution which overestimates the true value by 2 % only – quite a good estimate!

The example shows, that, as with the minimum number of respondents' rules in case of a minimum frequency with N=2, there is indeed a risk of (approximate) disclosure, because either of the two largest contributors would be able to derive a close upper estimate for the contribution of the other one by subtracting his or her own contribution from the aggregate total. One option to prevent this kind of disclosure risk is to use a (2,k)-dominance rule instead of the (1,k)-dominance rule. The (2,k)-dominance rule is based on the percentage of the two largest contributions in the cell instead of only the largest contribution, e.g. cells are considered unsafe when

$$x_1 + x_2 > \frac{k}{100}X \qquad (1)$$

As the (2,k)-dominance rule has a certain tendency for over-suppression – as will be explained below - we rather suggest the use of (minimum protection of) p %-rules, or, briefly, p %-rules, also offered by τ-ARGUS.

p%-rules

According to the p%-rule, an aggregate is sensitive, when the second largest respondent could estimate the largest contribution x_1 to within p percent of x_1 , e.g. when

$$\frac{(X - x_2) - x_1}{x_1} \cdot 100 < p \qquad (2)$$

This rule can be illustrated as follows: Assuming, that there are no coalitions of respondents, i.e. there are no intruders knowing more than one of the contributions, then the best upper estimate of any other contribution can be obtained by the second largest contributor, when he subtracts his own contribution x_2 from the aggregate total (e.g. cell value) X to estimate the largest contribution ($\hat{x}_1 = X - x_2$). Application of the p%-rule yields that this upper estimate will overestimate the true value by at least p % for any non-sensitive cell.

Comparison of p%-rule to (2,k)-rule

When both sides of relation (1) used for definition of the (2,k)-rule above are subtracted from X and then divided by X the result will be

$$\frac{(X - x_2) - x_1}{X} \cdot 100 < 1 - \frac{k}{100} \qquad (3)$$

In this formulation, the (2,k)-rule looks very similar to the formulation of the p%-rule given by (2). Both rules define an aggregate to be sensitive, when the estimate $\hat{x}_1 = X - x_2$ does not overestimate the true value of x_1 'sufficiently'. The difference between both rules is in how they determine this 'sufficiency'. According to the p %-rule, it is expressed as a rate of the true value x_1, while according to the (2,k)-rule, it is expressed as a rate of the aggregate total X. Considering this, the concept of the p%-rule seems to be more natural than that of the (2,k)-rule.

(2,k)-rules correspond to p%-rules in the following way: If *k* is set to $100 \cdot \frac{100}{100 + p}$ then any aggregate, which is safe according to the (2,k)-rule, is also safe according to the p%-rule. This can be proven as follows:

For an aggregate which is safe according to the (2,k)-rule with *k*=$100 \cdot \frac{100}{100 + p}$ (i) and (ii) will hold:

(i) $x_1 \le \frac{k}{100} \cdot X = \frac{100}{100 + p} \cdot X$, and

(ii) $\frac{(X - x_2) - x_1}{X} \ge 1 - \frac{k}{100} = 1 - \frac{100}{100 + p} = \frac{p}{100 + p}$ (c.f. (3)). Equal to (ii) is

(iii) $\frac{(X - x_2) - x_1}{x_1} \ge \frac{p}{100 + p} \cdot \frac{X}{x_1}$.

From (i) it follows that $\frac{p}{100 + p} \cdot \frac{X}{x_1} \ge \frac{p}{100}$.

And hence from (iii) : $\frac{(X - x_2) - x_1}{x_1} \geq \frac{p}{100 + p} \cdot \frac{X}{x_1} \geq \frac{p}{100}$.

On the other hand however, not any aggregate, which is safe according to the p %-rule, is also safe according to this (2,k)-rule. An example is given below (example 3). In these cases the aggregate could be published according to the p %-rule, but would have to be suppressed according to the (2,k)-rule, with $k = 100 \cdot \frac{100}{100 + p}$.

Based on the above explained idea, that the concept of the p %-rule is more natural than that of the (2,k)-rule, one might interpret this as a tendency for over-suppression in the (2,k)-rule.

Example 3:

Let p = 10

Then $k = 100 \cdot \frac{100}{100 + p} = 90.9$

Let the total value of a table cell be X = 110 000 .

Let the largest contribution be $x_1 =$ 52 000 .

Let the second largest contribution be $x_2 =$ 50 000 .

Then $\hat{x}_1 = X - x_2 = 110\,000 - 50\,000 = 60\,000$

$\Rightarrow 100 \cdot \frac{\hat{x}_1 - x_1}{x_1} = 100 \cdot \frac{60\,000 - 52\,000}{52\,000} = 15.4$,

e.g. the upper estimate $\hat{x}_1 = X - x_2$ will overestimate the true value by 15.4 %. So the aggregate is safe according to the p %-rule at p = 10.

On the other hand the two largest contributions are $x_1 + x_2 = 52\,000 + 50\,000 = 102\,000$.

As $102\,000 > \frac{100}{100 + p} \cdot X = \frac{100}{110} \cdot 110\,000 = 100\,000$ the aggregate is not safe according to the (2,k)-rule. The overestimation expressed as a rate of X is in this example

$100 \cdot \frac{\hat{x}_1 - x_1}{X} = 100 \cdot \frac{60\,000 - 52\,000}{110\,000} = 7.\overline{27}$ which is less than what would be required according to the (2,k)-rule in formulation (ii) above: $100 \cdot \left(1 - \frac{100}{100 + p}\right) = 100 \cdot \frac{10}{110} = 9.\overline{09}$.

Literature also discusses an extension of the p %–rules, the so called prior-posterior (*p,q*)%-rules. 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 $X_R := \sum_{i>2} x_i$ to within q %. An aggregate is

then considered unsafe when the second largest respondent could estimate the largest contribution x_1 to within p percent of x_1 , by subtracting her own contribution and this estimate E(X_R) from the cell total, e.g. when $\left| \left(X - x_2\right) - x_1 - E(X_R) \right| < \frac{p}{100} \cdot x_1$. Because $\left(X - x_2\right) - x_1 = X_R$, the left hand side is assumed to be less than $\frac{q}{100} \cdot X_R$. So the aggregate is considered to be sensitive when $X_R < \frac{p}{q} \cdot x_1$. Evidently, it is actually the ratio p/q which determines which cells are considered safe, or unsafe. So, any (p,q)-rule with $q < 100$ can also be expressed as (p*,q*)-rule, with $q^*=100$ (choose $p^* := 100^* p/q$).

When it is realistic to assume that groups of N ($N > 1$) respondents contributing to the same cell pool their data to disclose the contribution of another respondent, users of τ-ARGUS can change the third parameter N of the rule corresponding to this assumption. Note also, that setting parameter N to zero yields a rule equivalent to a (1,k)-dominance rule, where $k = \left(\frac{100}{p+100} \right) \cdot 100$.

For a more analytical discussion of sensitivity rules the interested reader is referred to [2].

2.2 Table level disclosure risk

When a table present totals or subtotals along with its 'inner' cells, there is a linear relationship between the cells of the table. Because of this linear relationship, if it has been established that a disclosure risk would be connected to the release of certain cells of a table, then other cells (so called 'complementary' or 'secondary' suppressions) must be suppressed in order to prevent a risk of disclosure on the table level.

Example 3 shows turnover in a hypothetical food production sector as the basis for subexamples showing the correspondence between primary cell-level sensitivity measures and table-level disclosure control. Example 3.1 begins the discussion by showing the potential problem of table-level disclosure risk when a minimum frequency rule with parameter N=3 is used to identify cell-level sensitivity.

Example 4: Table column „turnover" in the food production sector

Food production sector	turnover	number of respondents
Total	T	N_T
thereof		
bakers	15 000	122
butchers	25 000	95
millers	X	N_X
brewers	Y	N_Y
others	15 000	51

T denotes the overall turnover in the food production sector, X and Y the turnover of the millers and brewers respectively. N_T, N_X, N_Y denote the corresponding number of respondents.

Example 4.1:

Assume that a minimum frequency rule with parameter N=3 is employed for primary confidentiality.

Let the number of millers in the table of example 4 be $N_X = 1$.

Let the number of brewers in the table of example 4 be $N_Y = 3$.

Then the turnover of the millers, X, is unsafe, and must be suppressed. But if no other cell is suppressed, X can easily be recalculated through subtraction: T – 15 000 – 15 000 – 25 000 –Y – 0 (= X) . A second cell also has to be suppressed (Y for instance) to avoid disclosure. In the following we discuss what should be considered when making this choice of a complementary suppression:

Feasibility Interval

Making use of the linear relations between published and suppressed cell values in a table with suppressed entries, it is always possible for any particular suppressed cell of a table, to derive upper and lower bounds for its true value. This holds for either tables with non-negative values, and those tables containing negative values as well, when it is assumed that instead of zero, some other (possibly tight) lower bound for any cell is available to data users in advance of publication. The interval given by these bounds is called 'feasibility interval'. Example 5[1] illustrates the calculation of the feasibility interval in the case of a simple two-dimensional table where all cells may only assume non-negative values:

Example 5	1	2	Total
1	X_{11}	X_{12}	7
2	X_{21}	X_{22}	3
3	3	3	6
Total	9	7	16

For this table the following linear relations hold:

$$X_{11} + X_{12} = 7$$
$$X_{21} + X_{22} = 3$$
$$X_{11} + X_{21} = 6$$
$$X_{12} + X_{22} = 4$$
$$\text{with } X_{ij} \geq 0 \text{ for all (i, j)}$$

Solving this linear problem for a particular suppressed cell, for X_{11} for instance, subject to the constraints $X_{ij} \geq 0$ for all (i,j), yields $3 \leq X_{11} \leq 6$. So, the feasibility interval for X_{11} is [3;6] .

Using linear programming methodology, it is possible for any suppressed cell in a table to derive an upper bound ($X^{\max}$) and a lower bound ($X^{\min}$) for the set of feasible values. Feasible means here with respect to the linear relations between published and unpublished cell values given by the table and also with respect to some *a priori* constraints for the suppressed cell values such as

[1] ([4], Table 10, p 20)

the assumption that they may only assume non-negative values. In the example above, for cell (1,1) these bounds are $X_{11}^{\min} = 3$ and $X_{11}^{\max} = 6$.

A general, mathematical statement for the linear programming problem to compute the bounds of the feasibility interval is given by (c.f. [8]):

$$\text{Min } y_i \quad \text{,and} \quad \text{Max } y_i \quad \text{subject to}$$

$$\sum_{i \in I} m_{ij} y_i = b_j \quad , j \in J$$

$$lb_i \le y_i \le ub_i \quad , i \in P \cup S$$

$$y_i = a_i \quad , i \notin P \cup S$$

where the additive structure of the table is given by the set of linear equations $\left[\sum_{i \in I} m_{ij} y_i = b_j \quad , j \in J\right]$ (typically $b_j = 0$, and $m_{ij} \in \{-1,0,1\}$). I, P, and S denote the set of all cells, of the sensitive cells, and of the secondary suppressions, respectively, and ub_i, lb_i are constraints on the cell values a_i.

τ-ARGUS assumes $ub_i - a_i = a_i - lb_i = \frac{q}{100} \cdot a_i$, where q= 100 by default.

Protection Interval

A proper suppression procedure should ensure that no suppression pattern be considered feasible, unless the resulting bounds of the feasibility interval of any sensitive cell cannot be used to deduce bounds on an individual respondent contribution that are too close. To address this problem technically, safety bounds are determined for any primary suppression. We call the interval between these upper and lower bounds *'protection interval'*. In the following, we explain how to compute a suitable protection interval.

A cell union of suppressed cells in a row or column of a table with an unsuppressed total , such as for instance the union of two primary suppressions or of a primary suppression together with the complementary suppression, can of course always be calculated through subtraction. Let us reconsider example 4:

Example 4.2:

Assume a rule of 3 is employed for primary confidentiality.

Let the overall number of respondents in the table of example 4 be $N_T = 268$.

Let the number of millers in the table of example 4 be $N_X = 1$.

Let the number of brewers in the table of example 4 be $N_Y = 1$.

The table contains two confidential cells: The cell values X and Y must not be published. Note, that cells such as these, with only a single respondent contributing to the cell, are occasionally referred to as *Singletons* in the context of secondary cell suppression.

If we apply the rule of 3 to the union of these suppressed cells, whose value can be obtained by subtraction from the column total T: X + Y = T – 15 000 – 15 000 – 25 000, we find that

an additional cell must be suppressed, because the number of respondents to the cell union is $N_{X+Y} = N_X + N_Y = 2$. Otherwise, the miller might disclose the contribution of the brewer by subtracting his own contribution X from the value of the cell union X + Y and vice versa.

Example 4.2 is an instance of exact disclosure. Generally, if unions of suppressed cells in a table can be computed from the linear table context, those unions should be safe in order to avoid the risk also of approximate disclosure for any particular single contribution to the combination. This is especially important, of course, if the effort for such computation is low, as in the case of cell unions within the same row or column.

Example 4.3:

Assume now, that a (1,85)-rule is employed for primary confidentiality.

Let the total turnover in the food production sector in example 4 be T = 55 345.

Let the number of millers in the table of example 4 be $N_X = 3$.

Let the number of brewers in the table of example 4 be $N_Y = 3$.

Let the sequence of contributions of distinct respondents to X (turnover of millers) be x_1=300, x_2=20, x_3=10.

Let the contribution sequence of Y (turnover of brewers) be y_1=5, y_2=5, y_3=5.

It can then be easily verified, that X is sensitive according to the (1,85)-rule, while Y is not. However, in order to prevent the largest miller from disclosure, it would not be sufficient to suppress the turnover of brewers along with the turnover of millers. The turnover of brewers is too small to protect the contribution of the largest miller from approximate disclosure. More precisely, the turnover of the largest miller still dominates the combined turnover of millers and brewers $Z = X + Y$ (contribution sequence $z_1 = 300 > z_2 = 20 > z_3 = 10 > z_4 = 5 = z_5 = 5 = z_6 = 5$)

because $z_1 = 300 > \frac{85}{100}\left(\sum_{i=1}^{6} z_i\right) = \frac{85}{100} \cdot (300 + 20 + 10 + 5 + 5 + 5) = \frac{85}{100} \cdot 345 = 293{,}25$.

So, for sufficient protection of the primary suppression, one should either suppress another cell instead of the turnover of brewers, or should suppress an additional cell along with it.

It has been proven in general [1] that a necessary condition for the union of a sensitive cell with an arbitrary secondary suppressed cell to be non-sensitive is that the value of the secondary suppression Y exceed a given minimum size. If the minimum size condition is not fulfilled for a complementary suppression Y to a sensitive cell X, then the combination of X and Y will still be sensitive. This minimum size depends on the particular "degree of sensitivity" of the sensitive cell according to the specific sensitivity rule employed. The formulas given below give calculations of the minimum size for the sensitivity rules mentioned above. They specify the minimum size for an adequate non-sensitive secondary suppression to protect a sensitive cell with a total cell value X and N distinct respondents with contributions $x_1 \geq x_2 \geq \ldots \geq x_N$.

Table 1: Minimum size condition for feasible complements

Sensitivity rule	Minimum size for a feasible complement
(1,k)-rule	$\frac{100}{k} x_1 - X$

(n,k)-rule	$\frac{100}{k}(x_1 + x_2 + ... + x_n) - X$
p%-rule	$\frac{p}{100}x_1 - (X - x_1 - x_2)$
(p,q)-rule	$\frac{p}{q}x_1 - (X - x_1 - x_2)$

See [2] for more general formulation and evidence of the minimum size requirement.

The formulas of table 1 can be used to compute bounds for the protection interval: If the distance between upper bound of the feasibility interval and true value of a sensitive cell were below the minimum size for a feasible complement calculated according to the formulas of table 1, then this upper bound could be used to derive an estimate for single contributions of the sensitive cell that were too close according to the safety rule employed.

Example: The turnover of the millers in example 4.3[2] (X = 330 with a sequence of contributions of distinct respondents $x_1 = 300,\ x_2 = 20,\ x_3 = 10$) is confidential according to the (1,85)-rule. If the upper bound X^{max} for this confidential value is below $\frac{100}{85}x_1 = 352{,}94$, then it will be dominated by the largest single contribution x_1.

τ-ARGUS therefore adds the cell value X to the minimum size for a feasible complement in order to compute an upper bound of the protection interval, also referred to as upper protection level. Out of symmetry considerations a lower bound for the protection interval (= lower protection level) is usually computed by subtracting this minimum size from the cell value. The protection interval given by these bounds would normally, according to the primary confidentiality rule, be sufficient to protect the sensitive cell.

In two important cases, however, the minimum size requirement is not a sufficient criterion for the combined cell to be safe. It is not sufficient, if the complementary suppression is an unsafe cell itself, and it is not sufficient, when the same respondent can contribute to more than one component cell in a combined cell. In these cases, a combination of a sensitive cell and its complement may still be sensitive, even though the minimum size requirement is fulfilled for the complement. The most prominent case is that of two singleton cells. No matter how large the cell values, the combination will be unsafe. For further discussion of this problem, which is often referred to as 'multi-cell disclosure' see [2]. τ–ARGUS offers heuristic solutions to this problem which are explained in the τ–ARGUS manual.

3 Secondary Cell Suppression in Practice

The "Secondary Cell Suppression Problem" is how to apply complementary suppressions to a set of sensitive cells in such a way as to ensure that the complementary suppressions:

- create the required uncertainty about the true values of the sensitive cells, but
- preserve as much information in the table as possible.

---

[2]) in section 2.2

A suppression pattern is assumed here to 'create the required uncertainty about the true values of the sensitive cells' if for any sensitive cell the suppression interval contains the protection interval given by the upper and lower protection level computed according to the formulas given above.

But how to rate the information content of a table, or how to rate the loss of information due to a particular suppression pattern?

To find a good balance between protection of individual response data and provision of information – in other words, to take control of the loss of information that obviously can not be avoided completely because of the requirements of disclosure control - it is necessary to somehow rate the information content of data. [8] presents a mathematical model of the secondary cell suppression problem as linear programming problem. Information loss is expressed in this model as the sum of costs associated to the secondary suppressions. The idea of equating a minimum loss of information with the smallest number of suppressions is probably the most natural concept. This would be implemented technically by assigning identical costs to each cell. Yet experience has shown that this concept often yields a suppression pattern in which many large cells are suppressed, which is undesirable. So, apart from the option to assign identical costs, τ–ARGUS offers a choice of cost functions, based on cell frequencies, or cell values, or power transformations thereof. Note that several criteria, other than the numeric value, may also have an impact on a users perception of a particular cells importance, such as its situation within the table (totals and sub-totals are often rated as highly important), or its category (certain categories of variables are often considered to be of secondary importance).

How now to solve the secondary cell suppression problem? τ-ARGUS offers an algorithm based on complex optimisation models to find the optimal solution. For larger tables, however, the required computation times are prohibitive. Moreover, users may not always be happy with this solution because cost functions offered by the current implementation fail to reflect some of the issues which affect a users perception of a particular cells importance, such as its situation within the table (totals and sub-totals are often rated as highly important), or its category (certain categories of variables are often considered to be of secondary importance). Alternatively, ARGUS also offers methods based on heuristic approaches. See section 4 for comparison of the performance of these alternative methods.

Tables in the context of secondary cell suppression

For setting up the secondary cell suppression problem for a table, all the linear relations between published and unpublished values of the table have to be considered. This leads us to a crucial question: What is a table anyway? In the absence of confidentiality concerns, a statistician creates a table in order to show certain properties of a data set, or to enhance comparison between different variables. So a single table might literally mix apples and oranges. Secondly, statisticians may wish to present a number of those 'properties', publishing multiple tables from a particular data set. Where does one table end, and the next start? Is the ideal table one that fits nicely on a standard size-sheet of paper? With respect to secondary cell suppression, we have to think of tables in a different way:

Firstly we consider the data basis for the table. In a micro-data file suitable for τ–ARGUS, each record contains a number of key codes and a number of entries giving respondent data on a response variable. If more than one record may correspond to the same respondent, the file must also contain an identifying code, the so called 'holding indicator'. The key codes may be regarded as respondent data on some categorical 'explanatory' variables. They can be used to group the respondents according to certain criteria such as their economic activity, region, size class of turnover, legal form, or to categorize a response variable, like for instance fruit production into apples, pears, cherries, etc. . When we talk about the number of dimensions in a table, we usually mean the number of explanatory variables used to specify groups and categories. A cell in a table

exhibits the value of (one category of) the response variable for the group of respondents falling into the same category for each explanatory variable. In this sense, a table is defined by the set of explanatory and response variables.

Hierarchical and Linked tables

Data collected within government statistical systems must meet the requirements of many users, who differ widely in the particular interest they take in the data. Some may need community level data, while others need detailed data on a particular branch of the economy but no regional detail. As statisticians, we try to cope with this range of interest in our data, by providing the data at several levels of detail. We usually combine explanatory variables in multiple ways, when creating tables for publication. If two tables presenting data on the same response variable share categories of at least one explanatory variable, there will be cells which are presented in both tables – those tables are said to be *linked* by the cells they have in common. In order to offer a range of statistical detail, we use elaborate classification schemes to categorize respondents. Thus, a respondent will often belong to various categories of the same classification scheme - for instance a particular community within a particular county within a particular state - and may thus fall into four categories of the regional classification.

The structure between the categories of hierarchical variables also implies sub-structure for the table. When, in the following, we talk about sub-tables without substructure, we mean a table constructed in the following way:

> For any explanatory variable we pick one particular non-bottom-level category (the 'food production sector' for instance). Then we construct a 'sub-variable'. This sub-variable consists only of the category picked in the first step and those categories of the level below belonging to this category (bakers, butchers, etc.). After doing that for each explanatory variable the table specified through a set of these sub-variables is free from substructure then, and is a sub-table of the original one.

Any cell within the sub-table does also belong to the original table. Many cells of the original table will appear in more than one sub-table: The sub-tables are linked.

Of course, we must not protect any linked tables, or sub-tables separately. Otherwise it might happen that the same cell is suppressed in one table because it is used as secondary suppression, while within another table it remains unsuppressed. A user comparing the two tables would then be able to disclose confidential cells in the first table. A common approach is to protect tables separately, but note any complementary suppression belonging also to one of the other tables; suppress it in this table as well, and repeat the cell suppression procedure for this table. This approach is called a 'backtracking procedure'. Though within a backtracking process for a hierarchical table the cell-suppression procedure will usually be repeated several times for each sub-table, the number of computations required for the process will be much smaller than when the entire table is protected all at once.

It must however be stressed, that a backtracking procedure is not global according to the denotation in [2]. See [2] for discussion of problems related to non-global methods for secondary cell suppression.

4 Solving the Secondary Cell Suppression Problem

τ–ARGUS offers a variety of algorithms to find a valid suppression. It is up to the user to trade-off quality vs. quantity, that is to decide how much resources (computation time, costs for extra software etc.) he wants to spend in order to improve the quality of the output tables with

respect to information loss. The package offers a choice basically between four different approaches which we charcterise briefly in the following

OPTIMAL(see [8]) This method aims at the *optimal solution* of the cell suppression problem. A feasible solution is offered at an early stage of processing, which is then optimized successively. It is up to the user to stop execution before the optimal solution has been found, and accept the solution reached so far. The user can also choose the objective of optimization, i.e. choose between different measures of information loss. Note that the method relies on high performance, commercial OR solvers.

MODULAR(see [4]) The method subdivides hierarchical tables into sets of linked, unstructured tables. The cell suppression problem is solved for each subtable using the algorithm of the optimal method (see above). Backtracking of subtables avoids consistency problems when cells belonging to more than one subtable are selected as secondary suppressions.

NETWORK(see [3]) Algorithm based on network flows methodology. It aims at a heuristic solution of the CSP for 2-dimensional tables. Network flow heuristics are known to be highly efficient. The method is able to produce high quality solutions for large tables very quickly. τ-ARGUS offers an implementation applicable to 2-dimensional tables with hierarchical substructure in one dimension. A license for a commercial OR solver will not be required to run the algorithm.

HYPERCUBE(see [5,9]) The hypercube algorithm GHM*ITER* is a fast alternative to the OR based methods. This heuristic is able to provide a feasible solution even for extremely large, complex tables without consuming much computer resources. The user, however, has to put up with a certain tendency for over-suppression. The variant of the method offered by τ-ARGUS involves, like the modular method, backtracking of subtables.

We observed the performance of those secondary cell suppression algorithms on a set of small to moderate sized 2- and 3-dimensional hierarchical tables considering information loss (number and added value of secondary suppressions at certain hierarchical levels of the tables), and disclosure risk.

Comparison of the results (for details see [7]) proves that, in a situation where a user is interested in obtaining a suppression pattern for a single table with rather few, rather small secondary suppressions, preferably on the lower levels of the table, the best choice is to use the method ‘Modular’. For medium sized, 3-dimensional tables, long CPU times (compared to the hypercube method) are a nuisance, but quality of the results clearly justify the additional computational effort.

Results obtained by method ‘Optimal’ on the other hand were less convincing: firstly, the disclosure risk problem for singleton cells is not solved in the current implementation. Secondly, results depend strongly on the particular cost function employed. If the cost function does not fully reflect the users idea of a good suppression pattern, performance of method ‘Optimal’ will not be worth the additional computational effort (compared to method ‘Modular’) which is quite considerable for 3-dimensional tables with elaborate hierarchical structure.

However, auditing the tables revealed that – because of the backtracking involved - users of the modular, and the hypercube method indeed face some disclosure risk: In our tables, we found up to 9 percent of primary suppressions where protection lacked.

In a situation where multiple linked, or extremely large 3-, or more dimensional tables have to be protected, with the current version of τ–ARGUS, the user is confined to method ‘Hypercube’. For

suggestions how to improve the performance of this method for linked tables by specialized processing see [6].

5 Summary

This paper has introduced into fundamental concepts of secondary cell suppression. The software package τ-ARGUS for tabular data protection is based upon those concepts. Using simple examples, it has been illustrated how common rules for disclosure risk assessment on the cell level naturally lead to certain criteria for the validity of a suppression pattern. It is the objective of cell suppression in τ-ARGUS to determine secondary suppressions as to meet these requirements of a safe suppression pattern. In order to obtain a safe suppression pattern, however, users have to define tables properly. Finally, the paper has mentioned some results of a study comparing alternative algorithms for secondary cell suppression in τ–ARGUS.

References

[1] Cox, L. (1981), 'Linear Sensitivity Measures in Statistical Disclosure Control', *Journal of Planning and Inference*, 5, 153 - 164, 1981

[2] Cox, L. (2001), 'Disclosure Risk for Tabular Economic Data', In: *'Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies'* Doyle, Lane, Theeuwes, Zayatz (Eds), North-Holland

[3] Castro, J. (2004), 'Network Flows Heuristics for Complementary Cell Suppression: An Empirical Evaluation and Extensions', In: '*Privacy in Statistical Databases*' Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 3050)

[4] De Wolf, P.P. (2002), 'HiTaS: A Heustic Approach to Cell Suppression in Hierarchical Tables', In: '*Inference Control in Statistical Databases*' Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 2316)

[5] Giessing, S., Repsilber, D. (2002), 'Tools and Strategies to Protect Multiple Tables with the GHQUAR Cell Suppression Engine', In: '*Inference Control in Statistical Databases*' Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 2316)

[6] Giessing, S. (2003), 'Co-ordination of Cell Suppressions: Strategies for use of GHMITER', proceedings of the UN/ECE worksession on Statistical Confidentiality, Luxembourg, April 2003

[7] Giessing, S. (2004), 'Survey on methods for tabular data protection in ARGUS', In: '*Privacy in Statistical Databases*' Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 3050)

[8] Salazar Gonzalez, J.J (2002), 'Extending Cell Suppression to Protect Tabular Data Against Several Attackers', In: '*Inference Control in Statistical Databases*' Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 2316)

[9] Repsilber, D. (2002), 'Sicherung persönlicher Angaben in Tabellendaten' - in Statistische Analysen und Studien Nordrhein-Westfalen, Landesamt für Datenverarbeitung und Statistik NRW, Ausgabe 1/2002 (in German)