Working Paper No. 36
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)**

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (vi): Software tools for statistical disclosure control

COORDINATION OF CELL SUPPRESSIONS: STRATEGIES FOR USE OF GHMITER

Contributed paper

Submitted by the Federal Statistical Office of Germany[1]

[1] Prepared by Sarah Giessing (sarah.giessing@statistik-bund.de).

Co-ordination of Cell Suppressions: strategies for use of GHM*ITER*

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

Abstract. The paper provides some background information on the method of the GHM*ITER* algorithm for secondary cell suppression in multiple linked tables. It proposes strategies how to use and control the program in an efficient way to co-ordinate cell suppressions in multiple linked tables, considering particularly the disclosure protection dilemma of a table production process across several statistical agencies. The paper reports on the results of empirical tests of the suggested strategies.

1 Introduction

GHM*ITER* is a software package for tabular data protection, offering to protect multiple hierarchical tables by secondary cell suppression using the GH*QUAR* hypercube method. For detailed description of the GH*QUAR* method see [12], [13], [14], [15]. An input data set to be processed by GHM*ITER* typically contains "primary suppressions": cells, which must not be published, because otherwise the published data of the table might be used to disclose individual respondent data contributing to those cells. In order to prevent these primary suppressions from exact, or interval disclosure from the linear 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".
The problem of finding an optimum set of suppressions is known as the 'secondary cell suppression problem'. It is computationally extremely hard to find exact, or close-to-optimum solutions for the secondary cell suppression problem for multiple, large hierarchical tables. The GHM*ITER* algorithm developed at the Statistical office of Northrhine-Westphalia/Germany offers a quick heuristic solution. Lately, the GHM*ITER* program has become more easily accessible and usable: It was one of the tasks of the EPROS[2] project CASC (Computational Aspects of Statistical Confidentiality, [10]) to integrate GHM*ITER* into the software τ-ARGUS [11] for tabular data protection. And while τ-ARGUS is a control centre for a variety of alternative algorithms for secondary cell suppression, on behalf of Eurostat a specialised user interface for GHM*ITER*, the CIF package [4], has been developed recently.

In the following sections the paper suggests strategies for application of the GHM*ITER* algorithm. Some of those strategies are already implemented in τ-ARGUS. In addition, the paper will propose some advanced approaches, suitable to improve results of the method when applied to sets of linked tables. The paper will focus particularly on the problem of co-ordinating suppression patterns between linked tables provided by different agencies. It will discuss models and strategies for such a co-operation, and present first empirical test results.

[2] European Plan for Research in Official Statistics

2 Background: Information loss and data security concept of GHM*ITER*

This section will briefly introduce into the method of the GHM*ITER* algorithm, focussing particularly on aspects of data security, and information loss concepts. It will explain in which way weighting options allow the user to influence the selection of secondary suppressions. Weighting options will be introduced that are consistent with the general information loss concept of GHM*ITER*.

The cell suppression algorithm implemented in the software package GHM*ITER* is based on a heuristic hypercube method. A hypercube method builds on the fact that a suppressed cell in a simple n-dimensional table without substructure cannot be disclosed exactly, if that cell is contained in a pattern of suppressed, nonzero cells, forming the corner points of a hypercube.

GHM*ITER* subdivides n-dimensional tables with hierarchical structure into a set of n-dimensional sub-tables without substructure. These sub-tables are then protected successively in an iterative procedure that starts from the highest level. Successively for each primary suppression in the current sub-table, all possible hypercubes with this cell as one of the corner points are constructed.
For each hypercube, a lower bound is calculated for the width of the suppression interval for the primary suppression, that would result from the suppression of all corner points of the particular hypercube (for details see sec. 2.1 below). If it turns out that the bound is sufficiently large, the hypercube becomes a feasible solution. For any of the feasible hypercubes, the loss of information associated with the suppression of its corner points is calculated. The particular hypercube that leads to minimum information loss is selected, and all its corner points are suppressed.
After all sub-tables have been protected once, the procedure is repeated in an iterative fashion. Within this procedure, when cells belonging to more than one sub-table are chosen as secondary suppressions in one of these sub-tables, in further processing they will be treated like sensitive cells in the other sub-tables they belong to. The same iterative approach is used for sets of linked tables.
It should be mentioned here that the 'hypercube criterion' is a sufficient but not a necessary criterion for a 'safe' suppression pattern. Thus, for particular subtables the 'best' suppression pattern may not be a set of hypercubes – in which case, of course, the hypercube method will miss the best solution and lead to some overprotection. Other simplifications of the heuristic approach that add to this tendency for over-suppression are the following: when assessing the feasibility of a hypercube to protect a specific target suppressions against interval disclosure, the method

- is not able to consider protection maybe already provided by other cell suppressions (suppressed cells that are not corner points of this hypercube) within the same sub-table,
- does not consider the sensitivity of multi-contributor primary suppressions properly, that is, it does not consider the protection already provided in advance of cell suppression through aggregation of these contributions,
- attempts to provide the same *relative* ambiguity to (eventually large) secondary suppressions that have been selected to protect cells in a linked sub-table, as if they were single-respondent primary suppressions, while actually it would be enough to provide the same *absolute* ambiguity as required by the corresponding primary suppressions.

2.1 Disclosure Control Aspects

Within this section, we will explain concepts used in GHQUAR to avoid interval disclosure and how to use them efficiently. When the sensitive cells have been identified as such according to a concentration rule for instance, which is often the case for economic data, then it is certainly not enough for secondary suppression to prevent exact disclosure only. The idea of a concentration rule is to prevent that users of statistical data are able to obtain close estimates for individual respondent data. In choosing the particular parameters of the

concentration rule, the disseminator defines the precision of estimates that would be too close. Considering this, the secondary suppressions must also be selected in such a way that users of a table published with suppressions are not able to derive estimates for individual respondent data that are any closer.
It is a well known fact that it is possible to derive upper and lower bounds for the true value of any particular suppressed entry of a table with non-negative, or similarly bounded, entries by using the linear relations between published and suppressed cell values in the table, and eventually some additional *a priori* constraints on the cell values, known to the user in advance of publication of the table, because the cell values are approximately known. The interval given by these bounds is called the 'suppression interval'. A lower bound for the width of the suppression interval resulting from suppression of the corner points of a hypercube $H = H_0 \bigcup H_1$ where H_0 and H_1 denote the even and odd corner points of the hypercube is given by:

(1) $\min(\min_{i \in H_0}(ub_i - a_i); \min_{i \in H_1}(a_i - lb_i)) + \min(\min_{i \in H_1}(ub_i - a_i); \min_{i \in H_0}(a_i - lb_i))$, where $\{(a_i, ub_i, lb_i); i \in H\}$ denote the cell values, and the upper and lower *a priori* bounds for the corner points of that hypercube (for discussion on even and odd corner points see f.i. [12]).

Users of GHQUAR are requested to state a minimum width for the suppression interval, that is, the size of a *sliding protection range* (denoted in the following as $\frac{spl}{a_p}$), as a ratio to the cell value a_p. When choosing a hypercube to protect a particular sensitive cell, GHQUAR computes the width of the suppression interval that would result from suppressing all its corner points, while eventually considering given *a priori* constraints for the cell values, and assuming that there are no other suppressions within this (sub-) table. The hypercube is considered to be feasible, only if the computed width exceeds the size of the sliding protection range stated by the user.

The upper bound of the suppression interval can be used to derive upper estimates for individual contributions to the suppressed cell. It has been proven ([1],[2]; for illustration and example see [7]) that these upper estimates will be too close according to the sensitivity rule employed by the disseminator, if the upper bound of the suppression interval for a sensitive cell is below a certain minimum feasible size which depends on the particular distribution of the individual contributions to this cell. (See ([6], Appendix) for formulas for minimum feasible upper bounds $a_p + upl$ relating to suitable upper protection levels *upl*, for the most common sensitivity rules.) Note that, in relation to the cell total, the feasible upper bound will be the larger, the stronger the concentration of the contributions in the cell, with a (relative) maximum for the worst case which are the single respondent cells. This fact can be used to derive (worst case) feasible upper bounds, that are unrelated to the particular distribution of the individual contributions in a cell, but depend on the cell value only. Such bounds can then be stated as a ratio to the cell value $\frac{a_p + upl}{a_p}$. See the appendix of [8] for formulas relating to the most common sensitivity measures.

In the following, we propose two alternative strategies to control GHQUAR as to ensure an upper protection level *upl*, that is, to ensure that the upper bound of the suppression interval for any sensitive cell exceeds $a_p + upl$.

Strategy 1: Choose $\frac{spl}{a_p} := \frac{upl}{a_p} + 1$ (2)

Strategy 1 is based on the assumption of zero lower *a priori* bounds, and no finite upper *a priori* bounds. Under this assumption, a lower bound for the upper bound of the protection

interval for the cell value a_p of an even corner point $p \in H_0$ is given by: $a_p + \min_{i \in H_1}(a_i)$, and the formula (1) for the lower bound for the width of the suppression interval simplifies to $\min_{i \in H_1}(a_i) + \min_{i \in H_0}(a_i)$.

The GHM*ITER* feasibility criterion thus ensures $\min_{i \in H_1}(a_i) + \min_{i \in H_0}(a_i) \geq \frac{spl}{a_p} a_p$ to hold, for any feasible hypercube H. Because of (2), this is equivalent to $\min_{i \in H_1}(a_i) + \min_{i \in H_0}(a_i) \geq \frac{upl}{a_p} + 1$, and because $\min_{i \in H_0}(a_i) \leq a_p$ it follows $a_p + \min_{i \in H_1}(a_i) \geq a_p + upl$. An upper protection level *upl* is thus ensured.

Strategy 2: Chose $\frac{spl}{a_p} := 2\frac{upl}{a_p}$ (3)

Strategy 2 is based on the assumption of symmetric *a priori* bounds (lb_i, ub_i),where $lb_i = (1\text{-}q)a_i$, $ub_i = (1+q)a_i$ $(0 \leq q \leq 1)$. Under this assumption, a lower bound for the upper bound of the protection interval for the cell value a_p of an even corner point $p \in H_0$ is given by: $a_p + q \min_{i \in H}(a_i)$, and (1) simplifies to $2 \cdot q \cdot \min_{i \in H}(a_i)$.

Under a symmetric *a priori* bounds assumption, the GHM*ITER* feasibility criterion thus ensures $2q \min_{i \in H}(a_i) \geq \frac{spl}{a_p} a_p$ to hold for any feasible hypercube H. Because of (3), this is equivalent to $2q \min_{i \in H_0}(a_i) \geq 2\frac{upl}{a_p} a_p$, and from there follows $a_p + \min_{i \in H}(a_i) \geq a_p + upl$. An upper protection level *upl* is thus ensured.

It is important to note that the resulting sliding protection range $\frac{spl}{a_p}$ is much smaller with strategy 2 than with strategy 1.

Empirical tests have confirmed that strategy 2 seems to be superior, resulting in suppression patterns with lower information loss (less cell suppressions and value loss). In practice however, it turned out that this strategy leads to a certain instability of GHM*ITER*. As mentioned above, the method is unable to 'add' the protection given by multiple hypercubes. In certain situations, particularly when *a priori* bounds are to be considered, it is not possible to provide sufficient protection to a particular sensitive cell (or secondary suppression) by suppression of one single hypercube. In such a case, GHM*ITER* would be unable to confirm that this cell has been protected properly, according to the user specification. The program would not end properly. In order to avoid this problem, a relaxation has been implemented lately. The new version is able to reduce the sliding protection range automatically, and individually, step by step, for those cells, the protection of which the program cannot confirm otherwise.

Another relaxation which should be mentioned in this context is the following: If we assign zero *a priori* bounds to a particular cell, and a positive sliding protection range has been specified, then this cell will be ineligible for secondary suppression, because any hypercube containing this cell will appear to give zero protection. We call this strategy '*freezing*' a cell. '*Frozen*' cells, however, may cause similar instability problems as mentioned just above. Therefore, an option to '*conditionally freeze*' cells is offered in the

new version. This option allows GHM*ITER* to reduce the sliding protection range even to zero, if there is no other solution.

2.2 Information Loss Aspects

For any hypercube satisfying the disclosure control conditions as described in section 2.1, GHQUAR computes the loss of information associated with the suppression of its corner points. The particular hypercube that leads to minimum information loss is selected, and all its corner points are suppressed.

The standard information loss measure for this selection procedure (in the following also referred to as 'costs' for suppressing a cell) is proportional to the logarithm of the cell value. This measure is used in combination with some heuristic approaches to distinct between several categories of cells, such as suppressed and (so far) unsuppressed cells, zero cells, and single contributor cells. The costs for suppressing a hypercube that contains at least one unsuppressed cell for instance, will at any rate exceed the costs for a hypercube containing only cells that are already suppressed.

There is also an option to make the algorithm avoid the suppression of cells in the margins of subtables. When this option is activated, cells in the margins of a subtable are assigned high additional costs. It is strongly recommended to use this option – not alone because the information content of cells in margins may be higher, but also because these cells create links between the subtables of a complex table. Any suppressed cell in a margin of one subtable may produce additional suppressions in another subtable.

2.2.1 Advanced weighting options

Weighting options are supposed to be used to make GHM*ITER* prefer certain cells to remain unsuppressed, or on the contrary, to make certain cells be used as complements first. In order to make application of the weighting option easier, and more effective, so called 'preference facilities' have been implemented. The user classifies particular cells, or groups of cells, into *preference categories* by assigning corresponding preference codes to the cells. GHM*ITER* will then increase, or, on the contrary, decrease the costs for suppressing a cell. Those modified costs will be derived according to one of several linear cost transformation functions to compute increased or decreased costs.

Preference-categories

1: do not suppress, as top priority
2: do not suppress, use identical costs for all preference '2' cells
3: do not suppress, costs still depending on the cell value
4: no preference
5: prefer as secondary suppression , costs still depending on the cell value
6: prefer as secondary suppression, use identical costs for all preference '6' cells
7: prefer as secondary suppression, as top priority

As mentioned in 2.1 above, whenever there is a choice of different hypercube-patterns that would all protect a particular target suppression, GHM*ITER* will chose the one with the smallest total value of the cost variable. As a standard, the value of the cost variable is derived as a logarithmic transformation of the cell value. Cell costs for unsuppressed 'inner' cells of a subtable vary from 1 to 99 999 .

When preference categories have been assigned, the interval [1;99 999] is split into several sub-intervals I_1 to I_7 with interval bounds a_i, b_i, where $a_i \le b_i$ for $i = 1$ to 6, $b_i = a_{i-1}$ for $i \in \{1,3,4,5,6\}$, $b_1 = 99\,999$, and $a_7 = 1$. Each of the categories 1 to 7 correspond to one particular sub-interval. Original cell costs of cells belonging to category i are transformed by suitable linear functions $f_i : I \to I_i$. The parameters of these functions are determined in such a way that $f_i(1) = a_i$ and $f_i(99\,999) = b_i$. The split of the interval [1;99 999]

depends on whether either of the 'top priority' categories 1 or 7, or none of them, are in use. It is not allowed to use both 'priority' preference categories 1, and 7 within the same application. If either of these two categories is in use, the result is in an interval-split, where there is some space between two of the
sub-intervals, e.g. if $a_7 < b_7$ then $a_1 = b_1 = b_2$ and $b_6 < a_7$, and *vice versa*
if $a_1 < b_1$ then $a_7 = b_7 = b_6$ and $b_1 < a_2$. The bounds are determined as to achieve the following objective: The total (modified) costs for any hypercube consisting of inner cells only, and containing a 'priority' category cell must either be larger (category 1) or smaller (category 7) than the total costs for any other candidate hypercube consisting of inner cells only. The length of the sub-intervals corresponding to categories 1,3,4,5, and 7 is linearly related to the percentage of cells in the corresponding preference category. Category 2, or 6 cells will all be projected on a single value a_2 or a_6 ., i.e. $a_2 = b_2$ and $a_6 = b_6$. Thus, in the simple case, without category 1, and 7 cells, modified costs for category 2 cells will be 99 999, and 1 for category 6 cells.

3 Co-ordination of cell suppression between tables

Due to technological advance in the information age, it is much easier nowadays for users of statistical data to compare and analyze suppression patterns in multiple tables. This increases the disclosure risk for linked tables immensely. Using proper procedures for co-ordination of suppression patterns in multiple linked tables is therefore becoming an issue of growing importance. This does impose not only methodological and technical, but also organizational problems in a situation where those tables are provided by different statistical agencies.

This section will present and discuss ideas to co-ordinate suppressions between tables on the background of facilities provided by GHM*ITER*. In sec. 3.1 we suggest and discuss strategies to apply GHM*ITER* to a set of linked tables, where all tables are provided by the same agency. In sec. 3.2 we look for a way out of organizational problems that we are faced with in a situation where multiple linked tables are provided by different statistical agencies. We mention two alternative models for potential co-operation between those agencies in order to co-ordinate secondary suppressions. Corresponding strategies to use GHM*ITER* for secondary suppression will be explained using a practical example for illustration. For this example, we have tested the suggested strategy empirically, using data from German business tax statistics. In sec. 3.3 we report on the results of this empirical study.

3.1 Strategy for an integrated approach to apply GHM*ITER* to a set of linked tables

As explained in section 2, GHM*ITER* has a strong tendency for overprotection. There is no guarantee at all that the resulting suppression pattern is the 'best' possible pattern. Moreover, especially when applied to sets of large, linked tables, GHM*ITER* tends to suppress rather many of the large values and/or high level cells. But these are the cells which are rated highly important by the statisticians, and should therefore not be used as secondary suppressions, if anyway possible. In the following, a strategy is proposed to make GHM*ITER* avoid suppression of large, high level cells to the extent possible:

In a first step, apply GHM*ITER* to high level (tabular) data only. Then, step by step, apply GHM*ITER* to larger sets of more detailed tables, while using a control option mentioned in sec. 2.1: We '*conditionally freeze*' cells that have not been suppressed in previous steps. These cells will then not be used as secondary suppressions where this is anyway possible without violating the requirement of protection against exact disclosure.

Obviously, there is some risk that the resulting suppression pattern does not protect all sensitive cells against interval disclosure properly. The resulting suppression pattern requires some post-processing. So, as final step of the procedure, we suggest to use the output of the last GHM*ITER* run as input for an exact method to solve the secondary cell suppression

problem, like the Fischetti/Salazar algorithm [9] for multiple tables to be delivered along with the final version of τ-ARGUS.

The proposed strategy will be of practical use of course only in a situation, where the original set of multiple tables is too large for direct use of the exact method, which requires a much higher computational effort than the simple heuristic of GHM*ITER*. The suggested strategy is expected to work well, if it turns out that in the GHM*ITER* output there are only a few sensitive cells left without proper protection. In that case, the exact method will have to assign only a few additional secondary suppressions. We anticipate that this will be possible with a much lower computational effort as compared to an application of the exact method to the original, unprotected set of tables.

Note, that we do not recommend to control GHM*ITER to protect against exact disclosure only*. We would expect that in this case not only a few, but nearly all sensitive cells would require additional protection against interval disclosure, with negative consequences for both, the computational effort required for the post-processing step, and the overall number of secondary suppressions required.

For illustration of the proposed GHM*ITER* processing, we consider the following example. The strategy proposed above was applied to a set of real life tables from the German investment survey. Unfortunately, the post-processing has not be tested because software for the exact method is not yet available. We hope to continue this work once it is implemented.

In our instance, the goal is to protect 3 overlapping tables of investment data.

Table dimensions are:

TYPE (type of expenditure), NACE-6[3], NACE-4, SC-E (Size Class Employees), SC-T (Size Class Turnover).

Tables are specified as follows:

Table 1: TYPE x NACE-4,

Table 2: TYPE x NACE-2 x SC-E, and

Table 3: TYPE x NACE-2 x SC-T .

Direct application of GHM*ITER* to this set of tables to protect against interval disclosure according to the level of protection specified by the statisticians of investment survey division resulted in a highly undesirable suppression pattern, particularly involving two important top level cells of table 1, namely the total of the variable TYPE (that is the total investment) for NACE sections C 'Mining', and D 'Production industries'.

After some trials, we finally obtained a 3-step strategy for application of GHM*ITER*, which resulted in a much more attractive suppression pattern.

Step 1: Apply GHM*ITER* to table 1 alone.

Step 2: Apply GHM*ITER* jointly to table 1 and subtables of table 2 and table 3, where

Subtable of table 2 is given as: NACE-0 x SC-E, and

Subtable of table 3 is given as: NACE-0 x SC-T,

while using the option to '*conditionally freeze*' those cells of table 1, which remained unsuppressed in step 1.

Step 3: Apply GHM*ITER* jointly to table 1, 2 and 3,

while using the option to '*conditionally freeze*' those cells, which remained unsuppressed inthe previous steps.

As compared to the initial 'direct' approach, this approach helped to save in table 1 100 % of the previously 10 suppressed cells on the top level of TYPE in combination with 2-digit NACE level and above. Over all NACE levels, the approach saved 50 % of the 36 previously suppressed top level of TYPE cells.

[3] With NACE-n (n = 2, 4, 5, 6) we denote common hierarchical (so called 'n-digit') levels of the NACE classification of the economy. NACE-0 denotes the level of the main economic sectors ('0'-digits level).

3.2 Strategies under models for co-operation

Using proper procedures for co-ordination of suppression patterns in multiple linked tables is becoming an issue of growing importance. In a situation where those tables are provided by different statistical agencies, certain difficulties are connected to this requirement. A typical case is the situation of EUROSTAT and the European National Statistical Institutes (NSI's). Tables published by Eurostat may virtually coincide with the total level of 15 corresponding tables on the national level, or actually overlap with those tables, for instance when a variable like 'COUNTRY' is one of the dimensions of a EU-level table. Germanys official statistical institutes face a similar problem. From the community (or German national) perspective, it would be ideal to protect those overlapping tables jointly. In practice, the dilemma is that the data become available in some countries/states earlier than in others. These agencies want to publish their data as soon as possible, before the community (or German national) data are even complete, in particular before it is possible to do a joint disclosure protection processing for the community. When finally the disclosure protection processing on the community level begins, many cells on the national level below are already published, and thus not available for secondary suppression. This has a badly damaging effect on the amount of information (e.g. the number of cells) that can be published on the community level. In the following, we will denote this problem as *'disclosure protection dilemma of a table production process across several statistical agencies'*.

This section describes an empirical study to test some strategies anticipated to help co-ordinate the selection of secondary suppressions between agencies in the process of disclosure protection of periodical (annual, in our instance) survey tabular data.

Our test setting consists of four independent sets of linked tables from German business tax statistics. Each set consists of 4 tables, presenting data on turnover, business taxes, etc. .Table dimensions are NACE-4, NACE-2, STATE, LF (legal form), and a dimension TYPE, which varies with each of the four sets.
Table 1: NACE-2 x STATE x TYPE,
Table 2: NACE-2 x LF x TYPE,
Table 3: NACE-4 x LF,
Table 4: NACE-2 x STATE x LF

It is important to note that for our scenario, we assume that table 2 and table 3 data is not intended to be published on the country level. We compare the performance of four strategies A, B1, B2, and C outlined below.

Strategy A is a simple application of GHM*ITER* as to minimize loss of information on the *national level*. This strategy is an option only under the following model of co-operation between the institutes: no information is published on the lower (country, or state) level in advance of the joint disclosure control processing.
The basic idea underlying strategies B1, and B2 is to use the advanced control of information loss facilities described in section 2.2.1 to prefer those cells for secondary suppression that are selected as secondary suppressions when strategy A is applied to the data of the *previous period*. Strategy C is based on the same procedure, but this time we do not make use of the advanced control facilities for the information loss. Under the B and C strategies, tables on the country or state level could be published in advance of the disclosure control processing on the top (European, or German national) level.

Strategy A: Apply GHM*ITER* jointly to Tables 1 to 4, using data of the *actual period.*

Strategies A1, A2, B:
Step 1: Apply strategy 1 to data of the *previous period.* (Step 1 is not involved in strategy B)
Step 2: Apply GHM*ITER* individually to the 16 state level projections of tables 1 and 4 (joint application to the 2 tables), using data of the *actual period*,
assign preference category 5 to secondary suppressions of step 1 **(Strategy B1)**
assign preference category 6 to secondary suppressions of step 1 **(Strategy B2)**
without use of advanced control of information loss facilities **(Strategy C)**
Step 3: Apply GHM*ITER* to Tables 1 to 4, using data of the *actual period*, and using the option to '*conditionally freeze'* those cells, which remained unsuppressed in step 2.

3.3 Empirical Results

We observed that both strategies B1 and B2 (= use of preference categories 5, or 6) in fact increased the likelihood for previous period secondary suppressions to be picked as secondary suppressions in the tabulations of the actual period. Whereas strategy C (= no utilisation of the 'historic' suppression pattern) yielded a rate of 42 % of historic secondary suppressions selected again in the actual data tables (14 010 of 33 210), strategy B1 resulted in a rate of 49 %, and strategy B2 in a rate of 52 % re-used historic secondary suppressions in tables 1 and 4. Strategy B2 appears to be a more powerful tool, when the aim is to re-use historic suppressions to the extent possible.

In Step 3, even though we used the option to '*conditionally freeze'* those cells, which remained unsuppressed in step 2, GHM*ITER* could not fully avoid them as suppressions, in order to obtain a suppression pattern that appeared to be feasible. Tables 1 shows that here again, strategies B1 and B2 gave better results. The rates of cells suppressed in step 3 although they had been declared 'publishable' in the previous step 2 are smaller. The more powerful strategy B2 performed best with 'only' 43 violations of the 'rule' not use frozen cells as secondary suppressions.

Table 1: Suppressed publishable cells

Strategy	No of publishable cells after step 2	No of publishable cells after step 2, suppressed in step 3	Percentage of publishable cells after step 2, suppressed in step 3
C	49782	82	0.16
B1	44964	58	0.13
B2	44048	43	0.10

Table 2 below compares the number of suppressions finally obtained in the tables with state and national level data, tables 1 and 4, after the final step 3 of the B and C strategies to the number of suppressions resulting from the simple strategy A .
Looking at this table, we find that from the point of view of the national statistical institute, strategy A may seem to appear as the best strategy: 1 076 national level suppressions result from strategy A, whereas the worst strategy (from the 'national' point of view), strategy B2, suppresses 5 338 cells on the national level.

Taking the states perspective however, leads to a very much different judgement: Now strategy A appears to be the worst strategy. It results in 25 726 secondary suppressions on the state level, whereas the best performing strategy under this perspective, strategy C, results in 19 004 secondaries on the state level only.

If we do not distinguish between the two levels, strategy C is still the best performer, with 23 965 secondary suppressions as compared to 26 802 resulting from the worst strategy A.

From every one of these perspectives, strategy B2, the 'stronger' tool (with respect to re-use of historic suppressions), also appears to be the more damaging one as compared to strategy B1. Note that this is the reason why the empirical study was not extended to involve an even stronger strategy B3 (= use of preference category 7, instead of categories 5 and 6). The conclusion is that both B strategies seem to be useful when the aim is to re-use historic cell suppressions to some extent. They also help to reduce the number of cases, where pre-published cells are selected as secondary suppressions. Strategies of this kind may thus help to reduce certain disclosure risks. But they did not turn out to be successful with respect to reducing information loss. In fact, they tend to increase the number of suppressions, as compared to strategy C (= no use of preference categories).

Table 2: Final number of secondary suppressions

Strategy	Number of secondary suppressions in tables 1 and 4...		
	...overall	...on the national level	...on the state level
B1	24 252	5 021	19 231
B2	25 169	5 338	19 831
C	23 965	4 961	19 004
A	26 802	1 076	25 726

4 Summary

The paper provided some background information on methodological concepts of GHM*ITER*, concerning data security and information loss. It identified some weak points in the data security concept of the program that lead to the problem that the program tends to suppress too many top level cells. In section 3.1 a strategy was proposed to deal with that disadvantage. This strategy however requires some post-processing to avoid a risk of inferential disclosure. Except for the post-processing step, the suggested strategy was tested empirically, and proved to work well in practice. Unfortunately, without the post processing step, the strategy results in a suppression pattern that has a certain risk of interval disclosure to it.

In sec. 3.2, strategies have been compared that relate to certain models of co-operation across agencies in the secondary cell suppression steps of the process of statistical (tabular) data production. The dilemma of a table production process involving several agencies is that for joint disclosure limitation either all agencies must wait (with their publications), until the last agency delivers their data, or that pre-published lower level cells have to be considered, with a damaging effect on the amount of data that can be published on the higher level. In that case there is the additional problem that GHM*ITER* may not be able to find a feasible suppression pattern without suppressing some of the pre-published cells, resulting in a risk of exact disclosure.

It turned out that, although the tested techniques proved to work well, our idea to make GHM*ITER* favour suppressions of a 'proper' historic suppression pattern does not seem to provide a way out of the dilemma. In the following outlook section, we therefore suggest further research and a modified approach.

5 Outlook and final remarks

The software τ-ARGUS for tabular data protection is currently further developed by Statistics Netherlands within the European project CASC. It is one of the great advantages of this package that it offers a variety of algorithms for secondary cell suppression. It will also allow to combine use of these algorithms. In section 3.1, this paper has proposed a strategy to control GHM*ITER* in such a way that suppression of high level cells is avoided to some extent. Unfortunately, there will be a risk of inferential disclosure in the resulting

suppression patterns. Some post-processing is thus necessary. We suggest to use the LP based Fischetti/Salazar [9] or HiTaS [3] algorithm for multiple tables for the post-processing. We would expect this method-mix to be useful in a situation where direct application of these methods to multiple large tables is not feasible due to enormous computational effort. Future practical experimentation is required to find out, if the speed-up of the method mix as compared to direct application of the LP based algorithm turns out to be of actual relevance.

In order to find a way out of the disclosure protection dilemma of a table production process involving several agencies, as explained in sections 3.2, and 5., further research should test if strategies similar to the one proposed in sec. 3.2 lead to better results, when applied with the LP based Fischetti/Salazar [9] or HiTaS [3] algorithms.

A modified approach to solve the problem using GHM*ITER* might involve a combination of the methods of section 3.2 with the methods suggested in sec. 3.1, namely to process higher level tables (say, NACE-0 instead of NACE-2) first.

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] 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)

[4] GE-Systems (2002): 'User Manual for CIF version 2.0', unpublished Manual, Eurostat, Luxembourg.

[5] Gießing, S. (1998), 'Looking for efficient automated secondary cell suppression systems: a software comparison', *Research in Official Statistics Journal 2/98*

[6] Gießing, S. (2001), 'New tools for cell suppression in τ-ARGUS: one piece of the CASC project work draft', paper presented at the Joint ECE/Eurostat Worksession on Statistical Confidentiality in Skopie (The former Yugoslav Republic of Macedonia), 14-16 March 2001

[7] Giessing, S. (2001), 'Nonperturbative Disclosure Control Methods for Tabular Data', In: *'Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies'* Doyle, Lane, Theeuwes, Zayatz (Eds), North-Holland

[8] 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)

[9] 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)

[10] Hundepool, A., 'The CASC project, In: *'Inference Control in Statistical Databases'* Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 2316)

[11] Hundepool, A., van de Wetering, A., de Wolf, P.P., Giessing, S., Fischetti, M., Salazar, J.J., Caprara, A. (2002), τ-ARGUS *users's manual, version 2.1*

[12] Repsilber, R. D. (1994), 'Preservation of Confidentiality in Aggregated Data', paper presented at the Second International Seminar on Statistical Confidentiality, Luxemburg, 1994

[13] Repsilber, D. (1999), 'Das Quaderverfahren' - in *Forum der Bundesstatistik, Band 31/1999: Methoden zur Sicherung der Statistischen Geheimhaltung*, (in German)

[14] Repsilber, D. (2000), 'Wahrung der Geheimhaltung sensibler Daten in mehrdimensionalen Tabellen mit dem Quaderverfahren' - in *Statistische Analysen und Studien Nordrhein-Westfalen, Landesamt für Datenverarbeitung und Statistik NRW, Ausgabe 3/2000* (in German)

[15] 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)

Acknowledgements

This work was partially supported by the EU project IST-2000-25069 Computational Aspects of Statistical Confidentiality.