Working Paper No. 20
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

# THE ARGUS-SOFTWARE

**Invited paper**

Submitted by Statistics Netherlands[1]

[1] Prepared by Anco Hundepool (ahnl@rnd.vb.cbs.nl).

# The ARGUS-software

Anco Hundepool
Statistics Netherlands
Methods and Informatics Department
P.O. Box 4000
2270 JM Voorburg, The Netherlands
E-mail: ahnl@rnd.vb.cbs.nl

**Abstract:** In this paper we will give an overview of the 5th framework CASC (Computational Aspects of Statistical Confidentiality) project and concentrate on the ARGUS software. This CASC-project can be seen as a follow up of the 4th Framework SDC-project. However, the main emphasis is more on building practical tools. The further development of the ARGUS-software will play a central role in this project. Besides this software development, several research topics have been included in the CASC-project. These research topics, both for the disclosure control of microdata as well as tabular data, aim at obtaining practical results that might be implemented in future version of ARGUS and find its way to the end-users.

**Keywords:** Statistical Disclosure Control, μ-ARGUS, τ-ARGUS, microdata, tabular data.

## 1. Introduction

Statistical Disclosure Control is a field in statistics that has attracted much attention in recent years. Decision-makers demand more and more detailed statistical information. And researchers have the capacity to perform complex statistical analysis on their powerful PCs and they desire detailed microdata. Therefore there is a growing pressure on the statistical offices to publish more and more detailed information. But the Statistical Institutes have to preserve the balance between their task as a data provider and their obligation to preserve the privacy of the respondents, who have trusted their individual information to them. Without respondents no statistical information. The CASC-project is an initiative to coordinate the research and development in Europe. It is partly subsidised by the 5th Framework program of the EU. As a follow-up of the SDC-project it aims at the combination of research and the development of practical tools, the ARGUS-software. We aim both at the SDC-problems for microdata as well as tabular data.

Statistical Disclosure Control is an increasingly important aspect of official statistics. The growing information need puts an increasing pressure on the National Statistical Institutes (NSI's) to publish more detailed statistical information. The NSI's are traditionally very well equipped to carry out large censuses and large scale surveys. These sources of information contain very detailed information about enterprises and individuals. The power of computer systems is no longer a barrier to the composition of very large and detailed tables - the traditional output of the NSI's.

New information systems, online databases and internet based systems of access make publishing these large tables a possibility, where previously the physical limit of the paper-publications would restrict the amount of detail that could be published. For the users of statistical information, (policy makers, researchers etc.) this is a very positive development. The NSI's will meet these requests for information using the new technology.

However, there is another side of the coin. When the NSI's are collecting the information needed to compose these large statistical databases, they have, for obvious reasons, guaranteed the confidentiality of the information provided by the respondents. Whether the information is collected via a voluntary survey or through a compulsory survey/census, it is vital for the NSI's to safeguard the respondent's confidentiality. As well as being a legal obligation this is also vital in maintaining the confidence of respondents. If the respondents have the feeling that their sensitive information is no longer safe in the hands of the NSI's then response rates will fall, and the value of the outputs will drop.

The 5th framework CASC project has made a major step forward in the development of practical tools for SDC. The main software developments in CASC are μ-ARGUS, the software

package for the disclosure control of microdata and τ-ARGUS for tabular data. Moreover, the CASC-project has also resulted in a long and impressive list of research papers. Already some of this research has been built into ARGUS while many others results will be implemented in future releases.

## 2. The CASC-team

The CASC project team brings together leading partners from 5 European countries. Both the NSI's as well as several universities participate in this project

| **The participating CASC-institutes** | |
|---|---|
| 1. Statistics Netherlands | 8. University La Laguna |
| 2. Istituto Nationale di Statistica | 9. Institut d'Estadistica de Catalunya |
| 3. University of Plymouth | 10. Institut National de Estadística |
| 4. Office for National Statistics | 11. TU Ilmenau |
| 5. University of Southampton | 12. Institut d'Investigació en Intelligència Artificial-CSIC |
| 6. The Victoria University of Manchester | 13. Universitat Rovira i Virgili |
| 7. Statistisches Bundesamt | 14. Universitat Politècnica de Catalunya |

As the CASC team is rather large we have formed a steering committee representing the five countries. The role of this steering committee is to coordinate all the work and also to bring the work of the five countries together.

### CASC Steering Committee

| **Institute** | **Country** | **Responsibility** |
|---|---|---|
| Statistics Netherlands | Netherlands | Overall management<br>Software development |
| Istituto Nationale di Statistica | Italy | Testing |
| Office for National Statistics | UK | |
| Statistisches Bundesamt | Germany | Tabular data |
| Universitat Rovira i Virgili | Spain | Microdata |

## 3. ARGUS Software development

### 3.1. Software concepts

As the CASC-project aims at practical solutions for disclosure control, we have given the development of the ARGUS software a central role in the project. The ARGUS software will play the binding factor between the different parts of the project. Research topics have only been included if they aim at results that either can be implemented in (future) version of ARGUS or aim at testing the methodology used in the CASC-project.

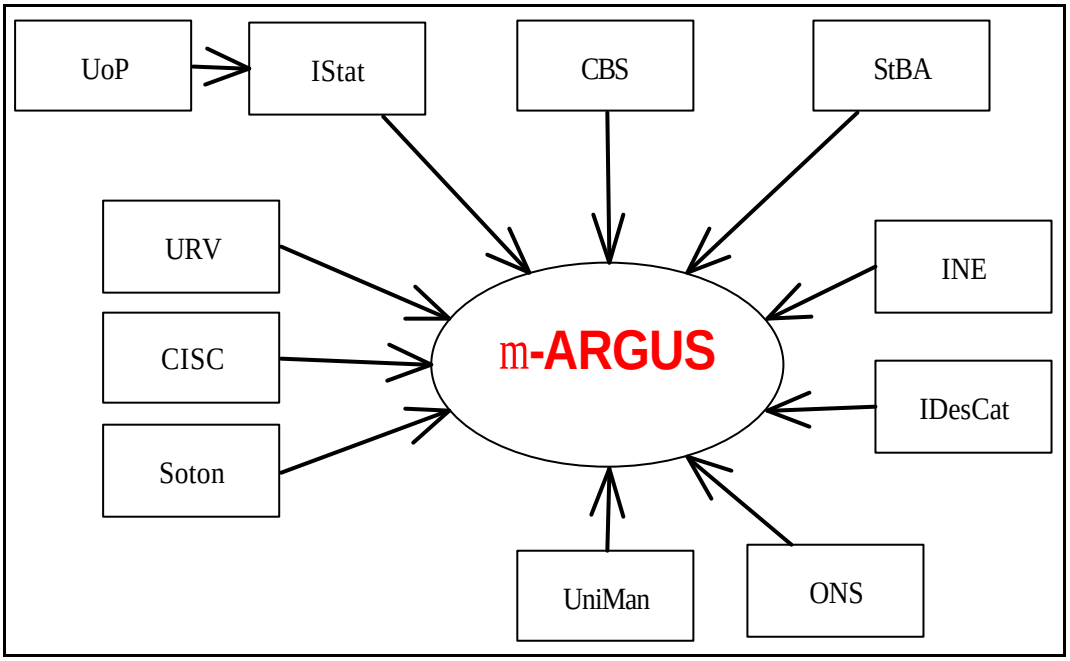
The starting point for the CASC-project were the ARGUS-twins resulting from the SDC-project. However as these twins had been developed with Borland C++, we have decided to convert the software to a more modern, up-to-date version of C++, i.e. Visual C++. However for the user-interfaces we use Visual Basic as a programming tool. This is an easier platform for the

development of user-interfaces, still meeting the needs of ARGUS. For the more crucial routines taking care of the heavy calculations we use Visual C++, which will lead to more efficient code. Some methods in ARGUS lead to complex computation problems, which justifies the choice for C++.

The routines build in Visual C++ will be compiled into an OCX-component, which can easily be used in the Visual Basic user-interface program. This guarantees a more flexible software concept and gives better options for the inclusion of additional routines for disclosure control even third party solutions. A first example is the link between ARGUS and the German GHQUAR/GHMITER software. However the aims with ARGUS in the CASC project are that ARGUS should be expanded into a control centre that will offer the user to choice of SDC-solutions. This also makes the comparison between the different solutions within one framework much easier.

## 4. m-ARGUS

The µ-ARGUS project team:




### 4.1. introduction

µ-ARGUS is based on a view of safety/unsafety of microdata that is used at Statistics Netherlands. In fact the incentive to build a package like µ-ARGUS was to allow data protectors at Statistics Netherlands to apply the general rules for various types of microdata easily, and to relieve them from the chore and tedium that producing a safe file in practice can involve. Not only should it be easy to produce safe microdata, it should also be possible to generate a logfile that documents the modifications of a microdata file.

The aim of statistical disclosure control is to limit the risk that sensitive information of individual respondents can be disclosed from data that are released to third party users. In case of a microdata set, i.e. a set of records containing information on individual respondents, such a disclosure of sensitive information of an individual respondent can occur after this respondent has been re-identified. That is, after it has been deduced which record corresponds to this particular individual. So, the aim of disclosure control should help to hamper re-identification of individual respondents represented in data to be published.

An important concept in the theory of re-identification is a key. A key is a combination of (potentially) identifying variables. An identifying variable, or an identifier, is one that may help an

intruder re-identify an individual. Typically an identifying variable is one that describes a characteristic of a person that is observable, that is registered (identification numbers, etc.), or generally, that can be known to other persons. This, of course, is not very precise, and relies on one's personal judgement. But once a variable has been declared identifying, it is usually a fairly mechanical procedure to deal with it in μ-ARGUS

In a disclosure scenario, keys (as described above) are supposed to be used by an intruder to re-identify a respondent. Re-identification of a respondent can occur when this respondent is rare in the population with respect to a certain key value, i.e. a combination of values of identifying variables. Hence, rarity of respondents in the population with respect to certain key values should be avoided. When a respondent appears to be rare in the population with respect to a key value, then disclosure control measures should be taken to protect this respondent against re-identification.

Therefore the occurrence of combinations of scores that are rare in the population should be avoided. To define what is meant by rare the data protector has to choose a threshold value $D_k$, for each key value k, where the index k indicates that the threshold value may depend on the key k under consideration. A combination of scores, i.e. a key value, that occurs not more than $D_k$ times in the population is considered unsafe, a key value that occurs more than $D_k$ times in the population is considered safe. The unsafe combinations must be protected, while the safe ones may be published.

Re-identification of an individual can take place when several values of so-called identifying variables, such as 'Place of residence', 'Sex' and 'Occupation', are taken into consideration. The values of these identifying variables can be assumed to be known to relatives, friends, acquaintances and colleagues of a respondent. When several values of these identifying variables are combined a respondent may be re-identified. See also Willenborg and De Waal (1996)

## 4.2. Statistical disclosure control measures

To avoid re-identification several techniques are available in μ-ARGUS, like global recoding (grouping of categories), local suppression, PostRAndomisation Method (PRAM), adding noise and microaggretation,

### 4.2.1. Global recoding

In case of global recoding several categories of a variable are collapsed into a single one. The effect will be that the number of records with the same key will rise. And the risk of re-identification will diminish. On the one hand side it is a very powerful instrument in μ-ARGUS to make a safe datafile. Many unsafe keys will disappear, but on the other hand a lot of detail can disappear as well. The data protector should use this method carefully and also keep in mind that if he cannot solve the unsafe keys here, he will have to apply many local suppressions (i.e. impute missing values). Future users of a dataset might not like this perspective and prefer a more aggregated categorisation for a variable without all these missing values.

It is important to realise that global recoding is applied to the whole data set, not only to the unsafe part of the set. This is done to obtain a uniform categorisation of each variable.

### 4.2.2. Local suppression

When local suppression is applied one or more values in an unsafe combination are suppressed, i.e. replaced by a missing value. This removes the possibility to use this key any longer fro re-identification. As keys often consists of several variables there is a freedom to select one of them for local suppression. Also several unsafe keys can be found in one records. μ-ARGUS offers two methods to do this efficiently. One is based on the minimising of the reduction of the entropy (i.e. preserving as much as possible the information), but as an alternative the user can specify his own priorities.

### 4.2.3. Top and bottom coding

Global recoding is a technique that can be applied to general categorical variables, i.e. without any requirement of the type. In case of ordinal categorical variables one can apply a particular global recoding technique namely top coding (for the larger values) or bottom coding (for the

smaller values). When, for instance, top coding is applied to an ordinal variable, the top categories are lumped together to form a new category. Bottom coding is similar, except that it applies to the smallest values instead of the largest. Top and bottom coding for categorical variables can be seen as special case of global recoding.

Top and bottom coding can also be applied to continuous variables. What is important is that the values of such a variable can be linearly ordered. It is possible to calculate threshold values and lump all values larger than this value together (in case of top coding) or all smaller values (in case of bottom coding). Checking whether the top (or bottom) category is large enough is also feasible.

### 4.2.4. The Post RAndomisation Method (PRAM)

PRAM is a disclosure control technique that can be applied to categorical data. Basically, it is a form of deliberate misclassification, using a known probability mechanism. Applying PRAM means that for each record in a microdata file, the score on one or more categorical variables is changed. This is done, independently of the other records, using a predetermined probability mechanism. Hence the original file is perturbed, so it will be difficult for an intruder to identify records (with certainty) as corresponding to certain individuals in the population. Since the probability mechanism that is used when applying PRAM is known, characteristics of the (latent) true data can still be estimated from the perturbed data file. See De Wolf et al (1998).

### 4.2.5. Microaggregation

Microaggregation is a family of statistical disclosure control techniques for *quantitative* (numeric) microdata, which belong to the substitution/perturbation category. The rationale behind microaggregation is that confidentiality rules in use allow publication of microdata sets if records correspond to groups of $k$ or more individuals, where no individual dominates (*i.e.* contributes too much to) the group and $k$ is a threshold value. Strict application of such confidentiality rules leads to replacing individual values with values computed on small aggregates (microaggregates) prior to publication. This is the basic principle of microaggregation.

To obtain microaggregates in a microdata set with $n$ records, these are combined to form $g$ groups of size at least $k$. For each variable, the average value over each group is computed and is used to replace each of the original averaged values. Groups are formed using a criterion of maximal similarity. Once the procedure has been completed, the resulting (modified) records can be published.

The method for multivariate fixed-size microaggregation implemented in µ-Argus tries to form homogeneous groups of records by taking into account the distances between records themselves and between records and the average of all records in the data set; this method will be called MDAV (multivariate microaggregation based on Maximum Distance to Average Vector).

### 4.2.6. Risk models

To be able to distinguish safe from unsafe microdata, it is necessary that a disclosure risk model is specified. Disclosure models can differ greatly in their degrees of sophistication. The basic model in µ–ARGUS is a fairly simple such model, namely one based on a thresholding rule. The understanding is that a combination of values is safe only if the (estimated) frequency of its occurrence in the population (or in the file) is above a certain threshold value.

An individual risk of disclosure allows one to estimate a measure of the chance of identification of each record in the released file on the basis of the actual values observed on the public variables. In the last few years a number of proposals have been made. Benedetti and Franconi (1998) propose a methodology for individual risk estimation based on the sampling weight, which is the approach used in this version of µ-ARGUS.

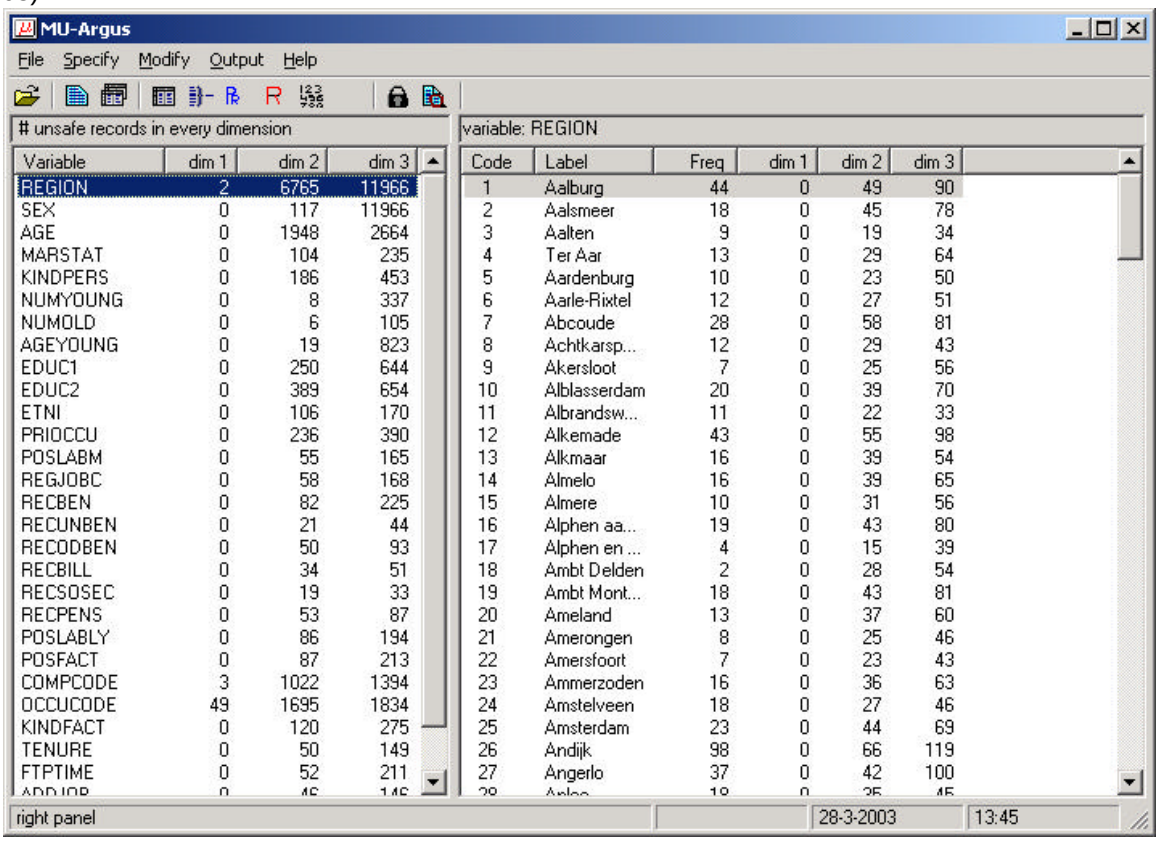
### 4.2.7. µ-ARGUS software

All these above mentioned methods have been implemented in the current versions of µ-ARGUS. We will continue to extend and improve µ-ARGUS, as our goal is to make all the SDC methodology easily available for the data-protectors. However it must be stressed that this software tools can only be applied by people with a basic understanding of the SDC-theory. They are not 'black-boxes', which will automatically produce a safe file.

The methods available in µ-ARGUS can be used to produce datafiles for different purposes. We make a basic distinction between datafiles that will be made available to established

researchers at universities et al. (possibly with a contract) and datafiles which will be made available to the general public. It goes without saying that in this case the much more strict rules have to be applied.

For more information on μ-ARGUS software we refer the μ-ARGUS-manual (Hundepool et al, 2003)



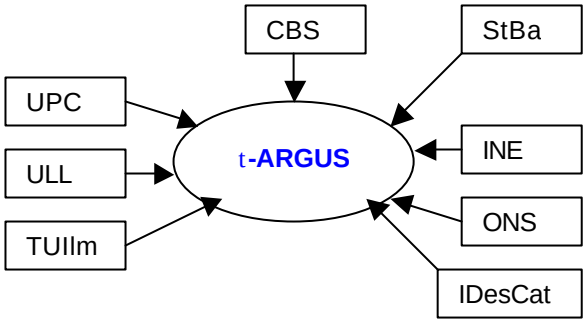

| Variable | dim 1 | dim 2 | dim 3 |
|---|---|---|---|
| REGION | 2 | 6765 | 11966 |
| SEX | 0 | 117 | 11966 |
| AGE | 0 | 1948 | 2664 |
| MARSTAT | 0 | 104 | 235 |
| KINDPERS | 0 | 186 | 453 |
| NUMYOUNG | 0 | 8 | 337 |
| NUMOLD | 0 | 6 | 105 |
| AGEYOUNG | 0 | 19 | 823 |
| EDUC1 | 0 | 250 | 644 |
| EDUC2 | 0 | 389 | 654 |
| ETNI | 0 | 106 | 170 |
| PRIOCCU | 0 | 236 | 390 |
| POSLABM | 0 | 55 | 165 |
| REGJOBC | 0 | 58 | 168 |
| RECBEN | 0 | 82 | 225 |
| RECUNBEN | 0 | 21 | 44 |
| RECODBEN | 0 | 50 | 93 |
| RECBILL | 0 | 34 | 51 |
| RECSOSEC | 0 | 19 | 33 |
| RECPENS | 0 | 53 | 87 |
| POSLABLY | 0 | 86 | 194 |
| POSFACT | 0 | 87 | 213 |
| COMPCODE | 3 | 1022 | 1394 |
| OCCUCODE | 49 | 1695 | 1834 |
| KINDFACT | 0 | 120 | 275 |
| TENURE | 0 | 50 | 149 |
| FTPTIME | 0 | 52 | 211 |

| Code | Label | Freq | dim 1 | dim 2 | dim 3 |
|---|---|---|---|---|---|
| 1 | Aalburg | 44 | 0 | 49 | 90 |
| 2 | Aalsmeer | 18 | 0 | 45 | 78 |
| 3 | Aalten | 9 | 0 | 19 | 34 |
| 4 | Ter Aar | 13 | 0 | 29 | 64 |
| 5 | Aardenburg | 10 | 0 | 23 | 50 |
| 6 | Aarle-Rixtel | 12 | 0 | 27 | 51 |
| 7 | Abcoude | 28 | 0 | 58 | 81 |
| 8 | Achtkarsp... | 12 | 0 | 29 | 43 |
| 9 | Akersloot | 7 | 0 | 25 | 56 |
| 10 | Alblasserdam | 20 | 0 | 39 | 70 |
| 11 | Albrandsw... | 11 | 0 | 22 | 33 |
| 12 | Alkemade | 43 | 0 | 55 | 98 |
| 13 | Alkmaar | 16 | 0 | 39 | 54 |
| 14 | Almelo | 16 | 0 | 39 | 65 |
| 15 | Almere | 10 | 0 | 31 | 56 |
| 16 | Alphen aa... | 19 | 0 | 43 | 80 |
| 17 | Alphen en ... | 4 | 0 | 15 | 39 |
| 18 | Ambt Delden | 2 | 0 | 28 | 54 |
| 19 | Ambt Mont... | 18 | 0 | 43 | 81 |
| 20 | Ameland | 13 | 0 | 37 | 60 |
| 21 | Amerongen | 8 | 0 | 25 | 46 |
| 22 | Amersfoort | 7 | 0 | 23 | 43 |
| 23 | Ammerzoden | 16 | 0 | 36 | 63 |
| 24 | Amstelveen | 18 | 0 | 27 | 46 |
| 25 | Amsterdam | 23 | 0 | 44 | 69 |
| 26 | Andijk | 98 | 0 | 66 | 119 |
| 27 | Angerlo | 37 | 0 | 42 | 100 |

Figure 1: μ-ARGUS main window
(overview of the unsafe combinations per variable)

## 5. t-ARGUS

The τ-ARGUS team




### 5.1. Introduction

Tables have traditionally been the major form of output of NSI's. There is also a longer tradition of studying the SDC-aspects of tabular data. Even in moderate sized tables there can be large disclosure risks. Take e.g. a cell in a table where there is only one contributor. The published cell

value is clearly the contribution of one respondent/enterprise. However the situation is more complex. Besides this the protection of the unsafe cells in a table is an even more complex task

## 5.2. Sensitive cells in magnitude tables

The well-known dominance rule is often used to find the sensitive cells in tables, i.e. the cells that cannot be published as they might reveal information on individual records. More particularly, this rule states that a cell of a table is unsafe for publication if a few (n) major contributors to a cell are responsible for a certain percentage (p) of the total of that cell. The idea behind this rule is that in that case at least the major contributors themselves can determine with great precision the contributions of the other contributors to that cell. The choice n=3 and p=70% is not uncommon, but τ-ARGUS will allow the users to specify their own choice.

As an alternative the prior-posterior rule has been proposed. The basic idea is that a contributor to a cell has better chances to estimate the competitors in a cell than an outsider and also that these kind of intrusions can occur rather often. The precision with which a competitor can estimate is a measure of the sensitivity of a cell. The worst case is that the second largest contributor will be able to estimate the larges contributor. If this precision is more than p% the cell is considered unsafe. An extension is that also the global knowledge about each cell is taken into account. In that case we assume that each intruder has a basic knowledge of the value of each contributor of q%.

Internationally and also in the Netherlands there is a shift from the dominance rule towards the prior-posterior rule. The reasons for this are the more intuitive approach, better numerical properties like the protection levels. Also waivers (contributors giving permission to publish their results) can be taken into account more easily. See Loeve (2001)

With these rules as a starting point it is easy to identify the sensitive cells, provided that the tabulation package has the facility not only to calculate the cell totals, but also to calculate the number of contributors and the n individual contributions of the major contributors. With this information τ-ARGUS can apply the sensitivity rules ands also perform the table redesign very easily. τ-ARGUS can produce the tables from the microdata files, also calculating the necessary additional information.

Traditionally τ-ARGUS could only read microdata files, but because of so many requests to be able to protect ready-made tables as well the next version of τ-ARGUS will have this facility. However with the restriction that the options for table redesign cannot be used any more.

A problem, however, arises when also the marginals of the table are published. It is no longer enough to just suppress the sensitive cells, as they can be easily recalculated using the marginals. Even if it is not possible to exactly recalculate the suppressed cell, it is possible to calculate an interval that contains the suppressed cell. This is possible if some constraints are known to hold for the cell values in a table. A common found constraint is that the cell values are all nonnegative.

If the size of such an interval is rather small, then the suppressed cell can be estimated rather precisely. This is not acceptable either. Therefore it is necessary to suppress additional information to achieve that the intervals are sufficiently large.

Several solutions are available to protect the information of the sensitive cells:

- Combining categories of the spanning variables (table redesign). Larger cells tend to protect the information about the individual contributors better.
- Suppression of additional (secondary) cells to prevent the recalculation of the sensitive (primary) cells.

The calculation of the optimal set (with respect to the loss of information) of secondary cells is a complex OR-problem. τ-ARGUS will be build around this solution and takes care of the whole process. A typical τ-ARGUS session will be one in which the users will first be presented with the table containing only the primary unsafe cells. The user can then choose how to protect these cells. This can be the combining of categories, equivalent to the global recoding of μ-ARGUS . The result will be an update of the table with presumably less unsafe cells (certainly not more). At

a certain stage the user requests the system to solve the remaining unsafe cells by finding secondary cells to protect the primary cells.

## 5.3. Sensitive cells in frequency count tables

In its simplest way sensitive cells in frequency count tables are defined as those cells that contain a frequency that is below a certain threshold value. This threshold value is to be provided by the data protector. This way of identifying unsafe cells in a table is the one that is implemented in the current version of τ-ARGUS It should be remarked, however, that this is not always the adequate way to protect a frequency count table. A greater risk in frequency tables is the so called group disclosure. If from a table is can be deduced that all contributor to a cell have a certain characteristic, this characteristic is revealed for contributors to a cell, (even many). This is also an undesirable situation. Current research at Statistics Netherlands aims at establishing better rules for these frequency tables.

## 5.4. Table redesign

In case in a table a great many sensitive cells appear to exist, it might be an indication that the spanning variables of the table are too detailed. In that case one could consider the possibility of combining certain rows and columns in the table. (This might not always be possible, from a publication policy point of view.) Otherwise the amount of secondary cell suppressions might just be too enormous. The situation is comparable to the case of microdata containing a lot of unsafe combinations. Rather than eliminating them with local suppressions one can remove them by using global recodings. The idea of table redesign is to combine rows, columns etc., by adding the cell contents of corresponding cells from the different rows, columns etc. It is a property of the dominance rule that a joint cell is more safe than any of the individual cells. So as a result of this operation the number of unsafe cells is reduced. One can try to eliminate all unsafe combinations in this way, but that might lead to an unacceptably high information loss. Instead, one could stop at some point, and eliminate the remaining unsafe combinations by using other techniques such as cell suppression.

## 5.5. Secondary cell suppression

Once the sensitive cells in a table - either of magnitude or a frequency count type - have been identified and there are not too many of these it might be a good idea to suppress these values. In case no constraints on the possible values in the cells of a table exist this is easy: one simply removes the cell values concerned and the problem is solved. In practice, however, this situation hardly ever occurs. Instead one has constraints on the values in the cells due to the presence of marginals and lower bounds for the cell values (typically 0). The problem then is to find additional cells that should be suppressed in order to protect the sensitive cells. The additional cells should be chosen in such a way that the interval of possible values for each sensitive cell value is sufficiently large. What is “sufficiently large” is to be specified by the data protector by specifying the protection intervals.

In general the secondary cell suppression problem turns out to be a hard problem, provided the aim is to retain as much information in the table as possible, which, of course, is a quite natural requirement. The optimisation problems that will then result are quite difficult to solve and require expert knowledge in the area of combinatorial optimisation.

## 5.6. Information loss in terms of cell weights

In case of secondary cell suppression it is possible that a data protector might want to differentiate between the candidate cells for secondary suppression. It is possible that he would like to preserve the content of certain cells as much as possible and is willing to sacrifice the values of certain other cells instead. A mechanism that can be used to make such a distinction between cells in a table is that of cell weights. In τ-ARGUS it is possible to associate different weights with the cells in a table. The higher the weight the more important the corresponding cell

value is considered and the less likely it will be suppressed. We shall interpret this by saying that the cells with the higher associated weights have a higher information content. The aim of secondary cell suppression can be summarised by saying that a safe table should be produced from an unsafe one, by minimising the information loss, expressed as the sum of the weights associated with the cells that have secondarily been suppressed.

τ-ARGUS offers several ways to compute these weights. The first option is to compute these weights as the sum of the contributions to a cell. Secondly this weight can be the frequency of the contributors to a cell, and finally each cell can be weighted as one, minimising the number of suppressed cells.

## 5.7. Solving the secondary cell suppression problem

Several approaches to solve this problem have been implemented in τ-ARGUS, each with its own characteristics and advantages and disadvantages

- The hypercube method
- The optimal solution
- The partial optimal solution
- The network solution

### 5.7.1. The hypercube method

The approach builds on the fact that a suppressed cell in a simple n-dimensional table without substructures cannot be disclosed exactly if that cell is contained in a pattern of suppressed, nonzero cells, forming the corner points of a hypercube.

The algorithm 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. To compute that bound, it is not necessary to implement the time consuming solution to the Linear Programming problem. 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. See Giessing (2002).

An implementation of this method by R. D. Repsilber of the Landesamt für Datenverarbeitung und Statistik in Nordrhein-Westfalen/Germany, offers a quick heuristic solution. The method has been implemented in τ-ARGUS The advantages are the speed of the solution even for very large tables and the fact that this method does not require a licence for commercial OR-software like the other solutions. A disadvantage might be that the solution will not be the optimal one, leading to over-suppression

### 5.7.2. The optimal solution

JJ Salazar (1998) has developed complex optimisation models to find the optimal solution for the secondary cell suppression. The models take into account the primary cells to be protected but also see to it that the cells cannot be recalculated to a given upper and lower protection level. These models have the flexibility to allow for different optimisation criteria, so it is possible to minimise the sum of the values of the cells to be suppressed, the sum of the frequencies of the individual cells of merely the number of cells to be suppressed.

The original Salazar models could only protect simple unstructured tables, but recently the models and the implementation have been extended for hierarchical and linked tables. Due to all the sub-totals present in these tables the intruder has many more options to recalculate suppression pattern and so the optimisation models have become much more complex.

It is to be expected that for very large tables the required computing time to find the optimal solution might be prohibitive in real life situations. But then alternatives are available in τ-ARGUS

The solution of these problems requires high performance OR-solvers which are only available commercially. In τ-ARGUS we have made provisions to solve the Salazar models with either Cplex or Xpress, two major solvers available.

### 5.7.3. The partial optimal solution (HITAS)

In real life situation most tables of NSI's tend to have one or more hierarchical spanning variable. As the original Salazar model could only handle non-hierarchical tables, an approximation has been build which breaks down the large hierarchical table into many unstructured sub-tables. This results in a whole tree of small sub-tables. Starting at the top this method then protects all these tables. As sometimes the suppression pattern influences a higher level of the tree a backtracking procedure will be carried out.

At the end of this procedure the whole table is protected. It proves to be a reasonable quick procedure, which has enabled us to protect very large table. See De Wolf (2002)

### 5.7.4. The network solution

Networks are often used in optimisation problems as an approximation of the full optimal solution. The advantages are that the solutions are obtained rather quickly, often at high quality. Therefore networks have been studied in the SDC area for a longer time. However the conclusions were that networks can only be used properly for 2-dimensional; tables. On the first sight this might be a serious drawback, but many very large tables produced by the NSI's are 2-dimensional, e.g. the foreign trade statistics.

So Jordi Castro (2003) has developed a network based solution, which is now available in τ-ARGUS. The first implementation only allows for non-hierarchical tables, but an extension for hierarchical tables is foreseen within the scope of the CASC-project.

## 5.8. The τ-ARGUS software

All the above mentioned solutions have been build in τ-ARGUS. The aim of τ-ARGUS is to make it into a control centre for tabular SDC. This will facilitate the users to apply the most appropriate method available for problem he faces. Like with μ-ARGUS τ-ARGUS is not a black box, which will just protect a table for you. τ-ARGUS is a control centre, which helps you to apple the appropriate SDC, measures and performs the complex computations involved.
For more information on τ-ARGUS software we refer the τ-ARGUS-manual (Hundepool et al, 2002)

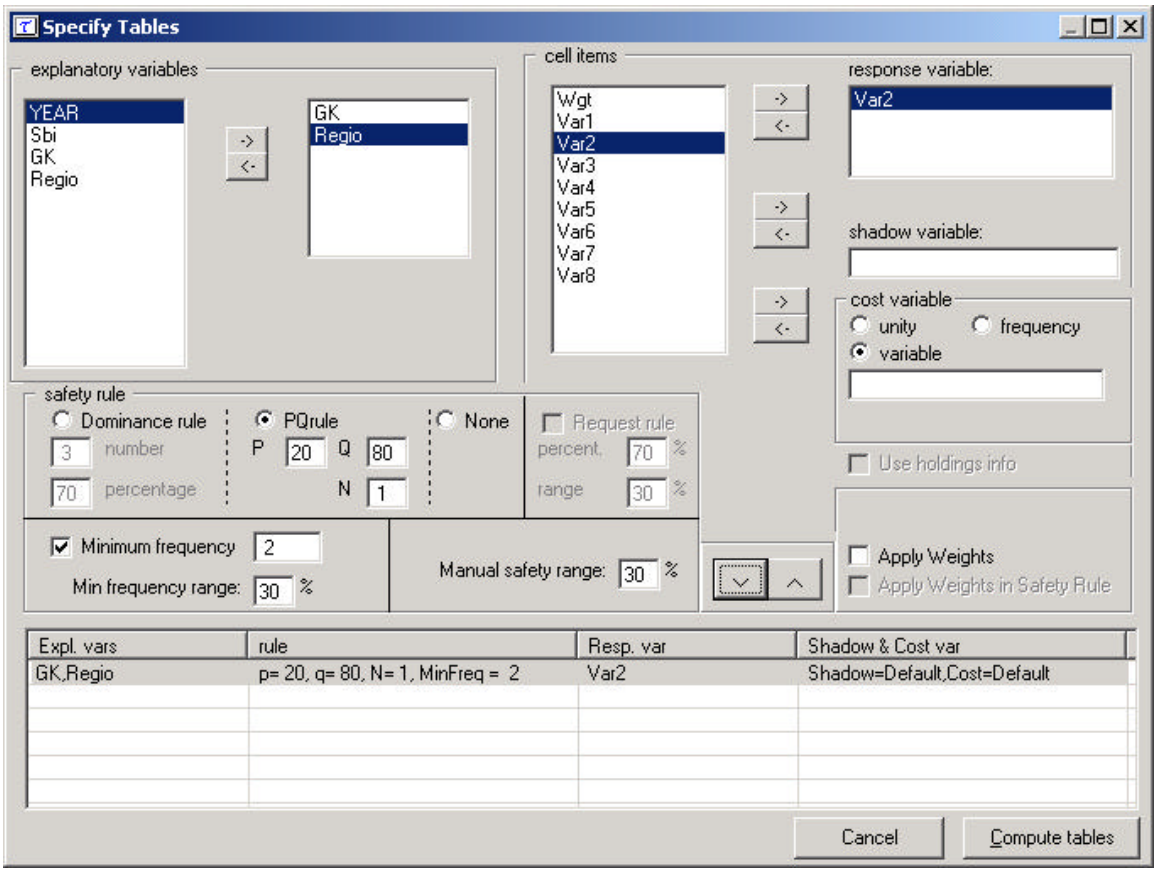



Figure 2: τ-ARGUS window to specify tables from a micro data file

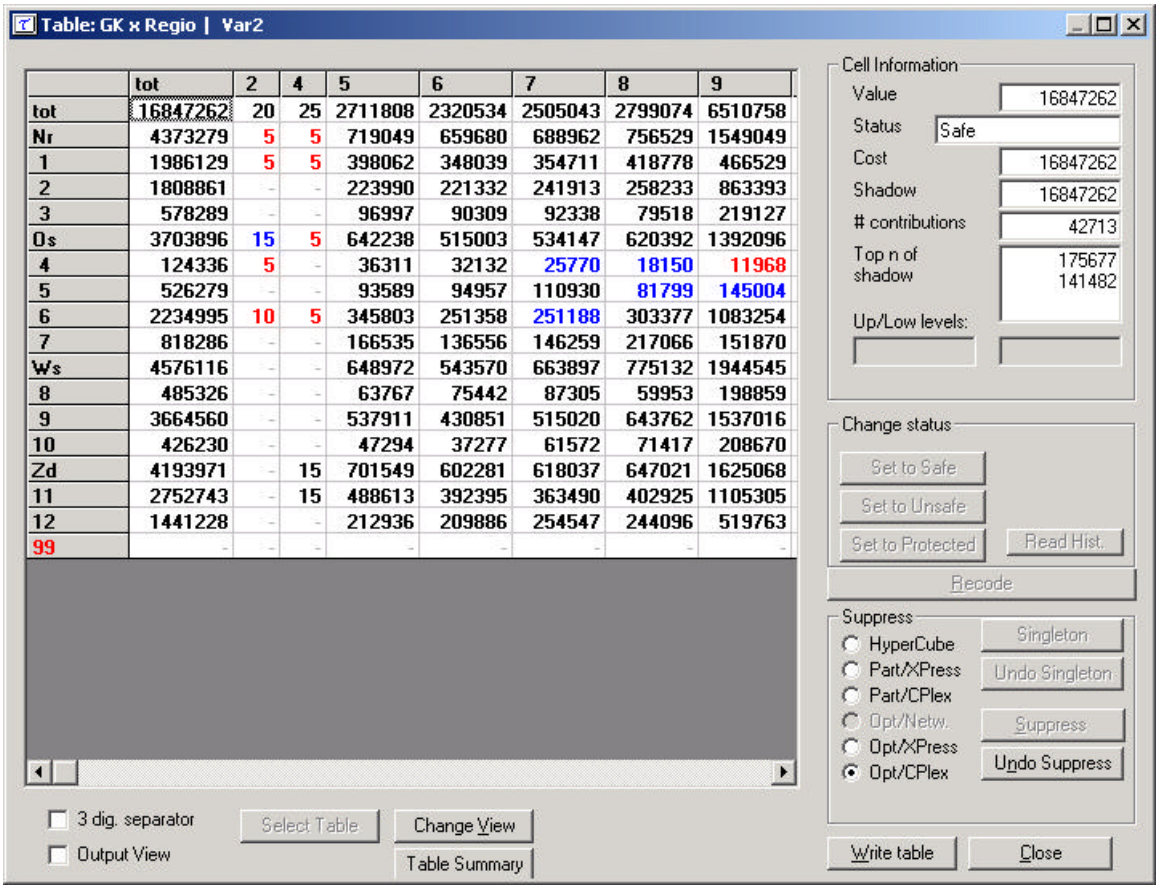



Figure 3: τ-ARGUS Table with unsafe cells

## 6. Conclusion

The major objective of this CASC-project is that the results will be used in real life situations in official statistics. The composition of the project team has been designed in such a way that the primary users, i.e. the NSI's, are active members. Seven statistical offices (5 national and two regional) participate in the project, either actively in the various stages of the development or as testers of the results. This reflects the needs and the interest of the NSI's for these kinds of tools.

Side effects of this project will be that the research community on Statistical Disclosure Control in Europe will work together. This joint effort will bring the state of the art to a higher level.

In order to disseminate the results of the CASC-project the project team will maintain a WEB-site. (http://neon.vb.cbs.nl/casc). Research papers resulting from this project as well as other material of interest for this field will find a place there. Also copies of μ-ARGUS and τ-ARGUS are available there.

## References


Benedetti, R. and Franconi, L. (1998), 'An estimation method for individual risk of disclosure based on sampling design'

Josep Domingo-Ferrer (2001), Josep Mateo-Sanz and Vicenc Torra, "Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk", *Paper presented at the NTTS/TTK Meeting Crete, Greece,* June 2001

J. Castro, "User's and programmer's manual of the network flows heuristics package for cell suppression in 2D tables" , research report DR 2003/07, Statistics and Operations Research Dept., Universitat Politècnica de Catalunya, 2003.

Fischetti, M. and J.J. Salazar-González (1998). *Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints.* Technical Paper, University of La Laguna, Tenerife.

Giessing, S. and Repsilber, D. (2002), 'Tools and Strategies to Protect Multiple Tables with the GHQUAR Cell Suppression Engine', in '*Inference Control in Statistical Databases'* Domingo-Ferrer (Editor), Springer Lecture Notes in Computer Science Vol. 2316.

Anco Hundepool, Aad van de Wetering, Ramya Ramaswamy, Luisa Franconi, Alessandra Capobianchi, Peter-Paul de Wolf, Josep Domingo, Vicenç Torra, Ruth Brand and Sarah Giessing (2003), μ-ARGUS user manual version 3.2, *Statistics Netherlands, Voorburg.*

Anco Hundepool, Aad van de Wetering, Peter-Paul de Wolf, Sarah Giessing,Matteo Fischetti, Juan-José Salazar and Alberto Caprara (2002), τ-ARGUS user manual 2.1, *Statistics Netherlands, Voorburg*

Anneke Loeve (2001), Notes on sensitivity measures and protection levels, Research paper 0129, *Statistics Netherlands, Voorburg*

Leon Willenborg and Ton de Waal (1996), Statistical Disclosure Control in Practice, Lecture Notes in Statistics Vol. 111, *Springer-Verlag, New York*

P.-P. de Wolf, J.M. Gouweleeuw, P. Kooiman and L.C.R.J. Willenborg (1998), Reflections on PRAM, Proceedings SDP98, Lisbon.

P.-P. de Wolf (2002). HiTaS: a heuristic approach to cell suppression in hierarchical tables. in '*Inference Control in Statistical Databases'* Domingo-Ferrer (Editor), Springer Lecture Notes in Computer Science Vol. 2316