NAZIOARTEKO ESTATISTIKA MINTEGIA
SEMINARIO INTERNACIONAL DE ESTADISTICA

Statistical Disclosure Control

Peter-Paul de Wolf
Rob Van de Laar

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA

59

SDC

Statistical Disclosure Control

Peter-Paul de Wolf

Coordinator Centre of Excellence on SDC (European Commission)

Rob Van de Laar

Member of Centre of Excellence on SDC

STATISTICS NETHERLANDS

AURKEZPENA

Urtero antolatzen dugun Nazioarteko Estatistika Mintegiaren **XXX.** urteurrenera heldu gara, kontutan harturik 1983. urtetik hona, estatistika alorrean mundu mailan ari diren ikertzaile aitzindari eta ospetsu asko gonbidatu izan ditugula.
Pasadan urtean, 2016an, "Big Data" gaur eguneko gaia eskeini genuen, Peter Struijs irakasle, Statistics Netherlands-etik etorria. Aurten ere, erakunde berberatik, Peter Paul de Wolf eta Rob Van de Laar adituak gonbidatu ditugu "Statistical Disclosure Control (SDC)" izeneko ikastaroa irakasteko, Vitoria-Gasteizko Letren Fakultatean, 2017ko azaroaren 20 eta 21an.

"Statistical Disclosure Control" (SDC), hau da, "Estatistika Zabalkundearen Kontrola" datuetan oinarritutako ikerketan erabilitako teknika bat da, pertsonak eta erakundeak identifikagarriak izan ez daitezen inkesten datu-analisien emaitzetan edo mikro-datuen argitalpenetan.
Baita "Estatistika konfidentzialtasunaren kontrola" izenburua erabil genezake, SDC teknika honen helburua hauxe delako: datuetan oinarritutako ikerketetan parte hartzen duten pertsonen eta inkestatuen konfidentzialtasuna babestea.

Gaur egun gure gizartean, erakundeok kudeatzen ditugun horrenbeste pertsonen eta erakundeen datuen babespena eta konfidentzialtasuna bermatzea ezinbestekoa da, hori dela eta, estatistika arloan ardura hau zabaldu nahi dugu; Eustat-eko bigarren oinarrizko printzipioa Konfidentzialtsuna baita.

Estatistika formakuntza honen zabalkundea ahalik eta pertsona eta erakunde gehienetara heldu ahal izateko, Eustat-eko web orrira jo dezakezue, **www.eustat.eus**, aurtengo Nazioarteko Estatistika Mintegiari buruzko informazioa eta baita 1983.urtetik hona, izan ditugun hizlarien txostenak ere eskuragarri izan ditzazuen.

Bukatu baino lehen, eskerrak eman nahi nizuke estatistikan erakutsitako interesagaitik eta bereziki aurtengo gaian: "Statistical Disclosure Control"

.

Vitoria-Gasteiz, azaroak 2017

JOSU IRADI ARRIETA
EUSTAT-eko Zuzendari Nagusia

PRESENTATION

We have arrived at the **30th** International Statistics Seminar. We have been organising these yearly seminars since 1983, and have consequently had the opportunity to invite innovative and renowned researchers in statistics on a global level.

One of the principles of official statistics is that of statistical confidentiality. It appears in our statistical legislation as the duty of statistical secrecy, which protects the identifiable details of both individuals and legal entities. It is one of the major concerns of statistical entities and this year we have invited the experts Peter Paul de Wolf and Rob Van de Laar to deliver a training course entitled "Statistical Disclosure Control" in the Arts Faculty of the University of the Basque Country in Vitoria-Gasteiz on 20 and 21 November 2017.

"Statistical Disclosure Control" (SDC) is a technique used in data based research in order to guarantee that no individual or organisation can be identified from the results of a survey analysis or in microdata publications.
We can also refer to this technique as "Control of Statistical Confidentiality", as the aim of this technique is the following: to protect the confidentiality of data obtained from those being surveyed and the subjects of the research.

Today, in our society it is vital for entities that handle so much personal and company data to guarantee the protection and confidentiality of the aforementioned data. We therefore want to transmit this responsibility in the statistical field, indeed the second basic principle of Eustat is confidentiality.

In order for this information to reach as many interested individuals and institutions as possible you have at your disposal Eustat's website: **www.eustat.eus**, where you can access information about the 2017 International Statistics Seminar and all past seminars since 1983. The technical notebooks of each speaker are also available.

Before concluding, I would like to thank you for the interest you have shown in statistics and especially this year's topic: " Statistical Disclosure Control"

Vitoria-Gasteiz, November 2017

JOSU IRADI ARRIETA
Director General of EUSTAT

PRESENTACIÓN

Hemos llegado al **XXX** Seminario Internacional de Estadística que año tras año llevamos organizando, teniendo en cuenta que desde el año 1983, hemos tenido ocasión de invitar a investigadores pioneros y reconocidos en materia estadística a nivel mundial.

Uno de los principios de la estadística oficial es el de la confidencialidad estadística que en nuestra legislación estadística aparece como el deber de secreto estadístico, que ampara los datos identificables tanto de la personas físicas como de las jurídicas. Es una de las grandes preocupaciones de los organismos estadísticos y por ello, este año hemos invitado a los expertos Peter Paul de Wolf y Rob Van de Laar para impartir el curso de formación titulado "Statistical Disclosure Control" en la Facultad de Letras de Vitoria-Gasteiz los días 20 y 21 de noviembre de 2017.

"Statistical Disclosure Control" (SDC), es decir, "Control de la Divulgación Estadística", es una técnica utilizada en la investigación basada en datos, para garantizar que ninguna persona u organización sea identificable a partir de los resultados de un análisis de encuestas o en publicaciones de microdatos.
También podríamos denominar a esta técnica "Control de la Confidencialidad estadística", ya que el objetivo de esta técnica es el siguiente: proteger la confidencialidad de los datos obtenidos de los encuestados y los sujetos de la investigación.

Hoy en día en nuestra sociedad, las entidades que gestionamos tantos datos de personas y empresas, es imprescindible garantizar la protección y confidencialidad de dichos datos. Por ello, en el área estadística queremos transmitir esta responsabilidad, teniendo Eustat como segundo principio básico de su Entidad la confidencialidad.

Para que esta difusión llegue al mayor número posible de personas e instituciones interesadas, podéis dirigiros a la página web de Eustat: **www.eustat.eus**, para acceder tanto a la información del Seminario Internacional de Estadística de este año como de jornadas anteriores, desde el año 1983, teniéndo acceso a los cuadernos técnicos de cada ponente.

Antes de finalizar, quisiera agradecerte por el interés mostrado por la estadística y especialmente el tema de este año: "Statistical Disclosure Control".

Vitoria-Gasteiz, noviembre 2017

JOSU IRADI ARRIETA
Director General de EUSTAT

BIOGRAFI OHARRAK

Peter-Paul de Wolf-ek Matematika doktoretza du Delft Unibertsitate Teknikoan (Holanda).1996.urtean Statistics Netherlands-en hasi zen lan egiten, Estatistika-Metodo Departamentuan.Gaur egun, Europar Komisioko (Eurostat) SDC (Statistical Disclosure Control) arloko Bikaintasun-Zentroaren koordinatzailea da. Horrez gain, SDC-ri buruzko proiektu batzuen nagusia da, EB-ek finantziatuta. Era berean, SDC-ri buruzko ikerketa metodologikoaren koordinatzailea da Statistics Netherlands-en.

Rob Van de Laar Fisika Aplikatuan Doktorea da Eindhoven Unibertsitate Teknikoan (Herbehereak). 1996.urtean Statistics Netherlands-en hasi zen lan egiten Estatistika- Metodo Departamentuan eta 2015. urtetik aurrera bere lan nagusia SDC gaian datza. Gaur egun, SDC arloko Bikaintasun-Zentroaren kidea da eta baita ikastaro metodologikoen Koordinatzailea ere.

BIOGRAPHICAL SKETCH

Peter-Paul de Wolf studied Mathematics at Technical University Delft (Holland) and he graduated “cum laude”. In 1996 he started his career at Statistics Netherlands at the department for Statistical Methods.
Currently he’s Coordinator of the Centre of Excellence on SDC (Statistical Disclosure Control) of the European Commission (Eurostat), project leader of several EU-funded projects on SDC and coordinator of the methodological research on SDC at Statistics Netherlands.

Rob Van de Laar studied Applied Physics at Technical University Eindhoven (Netherlands) and he graduated “cum laude”. In 1996 he started his career at Statistics Netherlands at the department for Statistical Methods. Currently he’s member of the Centre of Excellence on SDC and Coordinator of the methodological courses at Statistics Netherlands.

NOTAS BIOGRÁFICAS

Peter-Paul de Wolf es Doctor en Matemáticas en la Universidad Técnica de Delft (Holanda). En 1996 empezó a trabajar en Statistics Netherlands en el Departamento de Métodos Estadísticos. Actualmente es el Coodinador del Centro de Excelencia del SDC (Statistical Disclosure Control) de la Comisión Europea (Eurostat), además de Jefe de varios proyectos sobre SDC financiados por la UE y Coordinador de Investigación metodológica en SDC en Statistics Netherlands.

Rob Van de Laar Doctor en Física Aplicada en la Universidad Técnica de Eindhoven (Países Bajos). En 1996 empezó a trabajar en Statistics Netherlands y desde el 2015 su principal tema de trabajo es el SDC. Hoy en día es miembro del Centro de Excelencia del SDC y Coordinador de cursos metodológicos en Statistics Netherlands.

Index

1. Statistical Disclosure Control

1.1 General description

When publishing statistical information, a balance must be achieved between the interests of the data suppliers and the interests of the users. On the one hand, users want as much information as possible, and as detailed as possible. On the other, the data suppliers (people and companies, as well as the registration holders) require that their privacy is guaranteed. *Private lives and public policies: confidentiality and accessibility of government statistics* (Duncan et al., 1993) is the very relevant title of an American book about this problem.

For instance, what a National Statistics Institute (NSI) may and may not publish follows from its statistical disclosure control policy. Here, statistical disclosure control (SDC) means preventing that content-related conclusions about recognisable units are made based on published or otherwise available data from the NSI. An extensive discussion of SDC can be found in Hundepool et al. (2012).

It must not be possible to make such conclusions based on statistical publications. For NSIs these often comprise tables, web articles, press releases, and scientific articles. However, also if microdata is made available for scientific analysis, this fundamental principle of statistics must remain in force.

Individual publication must comply with the policy and other rules on SDC that NSIs have agreed upon. However, not all publications satisfy these rules in and of themselves. On the contrary, frequently a publication will have to be 'protected'. Different methods are available to protect microdata, table data and analysis results. The theme of SDC can thus be broken down into a number of sections:

- Statistical disclosure control methods for microdata,
- Statistical disclosure control methods for quantitative tables,
- Statistical disclosure control methods for frequency tables,
- Statistical disclosure control methods for analysis results.

The conflicting interests of privacy protection and information retention play an ongoing role in SDC. When using the different methods for SDC, these two aspects must be taken into account. The SDC policy sets down a minimum level of protection. The real skill of the person protecting the data is to use different disclosure control methods in such a way that the minimum required level of protection is achieved and that the information loss is as small as possible. This will be different in every situation, as the concept of information loss can have different meanings for the different users of the data.

The methods mentioned in this treatise will each be explained separately. However, in practice, for each situation, multiple methods will often be used at the same time to create 'safe' publications.

1.2 Scope and relationship with statistical bureaus

The statistical disclosure control policy of an NSI will not be described here. That policy is often laid down in an internal handbook. However, various available methods will be described that can be used to apply that policy.

When applying SDC methods, both the level of protection and the information loss of the publications must be examined. Since the concept of 'information' is subjective and therefore can be defined differently by each user (even in a single publication), it is not possible to prescribe a specific method for each specific situation. The methods will therefore be described along with their advantages and disadvantages, along with their effects on the level of protection and information loss. A staff member who is in charge of the statistical disclosure control of a publication (in whatever form), will subsequently have to choose the most suitable method for the publication in question.

1.3 Place in the statistical process

Statistical disclosure control traditionally takes place at the end of the statistical process: statistical disclosure control is applied immediately before publication (in whatever form). Ideally, account should be taken during the entire statistical process of the fact that, ultimately, the publication will have to satisfy the statistical disclosure control policy. However, measures can also be taken at the start of the statistical process, such as formulating the cover letter for participation in a survey ('informed consent').

The concept of SDC therefore plays a role during the entire statistical process. However, the specific methods as described in this document are only used at the end of the statistical process, immediately before publication.

1.4 Definitions

Term	Definition
μ-ARGUS	Software for the statistical disclosure control of microdata files
τ-ARGUS	Software for the statistical disclosure control of tables
Disclosure	The obtaining of information from statistical data about a specific recognisable person, household, company or institution
Identifying variable	Variable of which the value can contribute to the identification of a specific person, household, company or institution
Primary unsafe cell	Cell in a table that does not satisfy the disclosure control rules

Secondary unsafe cell	Cell in a table that does satisfy the disclosure control rules, but which must be suppressed to protect the primary unsafe cells
Structural zero cell	A cell for which it is generally known that, logically, this cell *cannot* have a contribution

An extensive English-language glossary for statistical disclosure control can be found at http://neon.vb.cbs.nl/casc/Glossary.htm.

1.5 Literature

Duncan, G.T., Jabine, T.B. and V.A. de Wolf (Eds.) (1993), *Private lives and public policies: confidentiality and accessibility of government statistics.* The National Academies Press, ISBN 0309086515.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K. and De Wolf, P.P. (2012), *Statistical Disclosure Control (Wiley Series in Survey Methodology)*. John Wiley & Sons, ISBN 9781118348215.

2. Statistical Disclosure Control of Microdata

2.1 General description and reading guide

2.1.1 General description

The methods described here are used to create protected microdata files. The extent to which the methods are used (or how strictly they are applied) depends partly on the type of file that is going to be released. This is described in detail in Hundepool et al. (2012), where a distinction is made between Public Use Files, Scientific Use Files and Secure Use Files.

The methods described in this section are easy to apply with the μ-ARGUS software package. This package was developed in a European context.

2.1.2 Reading guide

As the first step in the statistical disclosure control of microdata, it will have to be determined whether disclosure is possible: is there any information about individual respondents in the microdata that may not be disclosed? This 'sensitive' information usually concerns respondents that can be recognised as unique or rare cases in the microdata. Such respondents must be protected.

Several of the disclosure control methods that we will discuss here can be applied to categorical variables: global recoding (section 2.3) and PRAM (section 2.7). Top (and bottom) coding is mainly intended for continuous variables; see section 2.5. Local suppression (section 2.4) can be used for both categorical and continuous variables. There is also the possibility of adding noise to raising weights; see section 2.6.

Which method or combination of methods will ultimately be used in a specific situation cannot be determined in advance. The department responsible for the construction of the microdata file is also responsible for adequate statistical protection. When selecting the method or methods to be used, two competing aspects must be taken into account:

- disclosure risk
- information loss.

In general, it can be said that reducing the disclosure risk will lead to increased information loss. The converse is also true: the smaller the information loss, the larger the disclosure risk.

2.2 Scope and relationship with other sections

The methods that will be described in this subsection are applied directly on the microdata itself. As a result, different levels of protection can arise, which correspond to those of the Public Use Files or the Scientific Use Files.

Users of unprotected files (Secure Use Files) and of Scientific Use Files can generate output that does not necessarily satisfy the disclosure control policy of the data providing NSI. In these situations, other methods will have to be used to protect the output. For such methods, see section 'Statistical Disclosure Control of Analysis Results'.

2.3 Global recoding

2.3.1 Brief description

In the statistical disclosure control of Public Use Files and Scientific Use Files, often mainly the variables that can potentially be used to identify a respondent are examined. These types of variables are called *identifying* variables. Identifying variables are generally categorical variables. Combinations of categories of identifying variables tend to lead to unique or rare people. Consider, for example, 'Mayor in Amsterdam' (unique) or 'Female neurosurgeon older than 55 years of age from Staphorst[1]' (rare). Commonly used rules for Scientific Use Files (see Hundepool et al., 2012) state that such combinations must occur sufficiently often in the target population.

By combining categories of identifying variables, rare combinations can be made less rare.

2.3.2 Applicability

In the disclosure control of microdata files that are released by an NSI, certain combinations of identifying variables must occur sufficiently often in the population. In particular, if an identifying variable is present in a very detailed form in the file, global recoding can in many cases be used to sufficiently protect the file, while the information loss remains limited.

For some researchers, however, global recoding will remove too much detail, as a result of which they will no longer be able to perform their analyses. It is therefore the task of the NSI staff member who is charged with the statistical disclosure control of the file to assess whether global recoding is a suitable protection method for the case in question.

However global recoding does not have to be limited to identifying variables. Non-identifying variables can also be globally recoded, as long as they are categorical variables.

[1] Staphorst is a small village in the Bible Belt of the Netherlands.

In such an application, subject to the possible identification of a respondent, only less detailed (and therefore probably generally known) information would be disclosed.

2.3.3 Detailed description

Global recoding involves the adaptation of the code list of an (identifying) variable. If the variable is hierarchical (for example, region), an obvious aspect of the recoding is to delete some detailed levels. For example, in a recoding of the City/Town variable, all cities could be replaced by their associated counties/provinces.

After the code list of a variable is adapted, for *each* record, the score on that variable is adapted to the new code list. This is therefore done not only for the unsafe records, but also for the safe records.

2.3.4 Example

Figure 1 shows several records from a fictitious microdata file. The records are numbered for easy reference.

	Occupation	Region	Gender	Education	…
1	Mayor	Amsterdam	Man	High	…
2	Fisherman	Urk	Man	Low	…
3	Teacher	Amsterdam	Woman	High	…
4	Plumber	Papendrecht	Man	Medium	
	…	…	…	…	…

Figure 1: Several records from a fictitious microdata file

The mayor from Amsterdam is, obviously, unique. The variable ‘Region’ is now globally recoded by replacing the city names by the associated province. This generates the records as shown in Figure 2.

	Occupation	Region	Gender	Education	…
1	Mayor	Noord-Holland	Man	High	…
2	Fisherman	Flevoland	Man	Low	…
3	Teacher	Noord-Holland	Woman	High	…
4	Plumber	Zuid-Holland	Man	Medium	
	…	…	…	…	…

Figure 2: Records from Figure 1 after global recoding of ‘Region’

Now, in record 1, the mayor is no longer unique. Because the recoding is applied *globally*, the Region variable in the safe records 2 to 4 is also adapted.

2.4 Local suppression

2.4.1 Brief description

Another SDC method to deal with combinations of identifying variables that occur insufficiently often in the target population, is local suppression. In local suppression, the score on at least one of the variables in a combination that occurs insufficiently often in the target population is suppressed (or it is assigned the score of 'Unknown'). As a result, the combination of the remaining variables describes a potentially larger group in the target population.

2.4.2 Applicability

Local suppression is often used as the final disclosure control method. At this point, most of the protection has already been provided by other methods, and local suppression is used to protect the last unsafe records.

Local suppression leads to missing values in the file. The way in which these missing values are selected, however, is certainly not random: the goal is to protect records that belong to small, identifiable groups. The effect of these missing values on the analyses to be conducted is different from the effect of missing values as a result of non-response.

For that matter, local suppression does not have to be limited to identifying variables. Non-identifying variables can also be locally suppressed. In the event that a respondent is identified, this ensures that no sensitive information would be disclosed.

2.4.3 Detailed description

In local suppression, the value of an identifying or other variable is set to 'Unknown'. By suppressing the score for at least one variable from an insufficiently often occurring combination, a lower dimensional combination is, in fact, created. The result of this is that the combination will potentially describe a larger group of respondents in the target population.

Local suppression is only applied to unsafe records. It is possible that multiple unsafe combinations of identifying variables occur in a single record. By suppressing the right variable or variables in an intelligent manner, multiple unsafe combinations can sometimes be protected simultaneously.

If a microdata file has multiple records of people from the same household, this must be taken into account in local suppression. Such records may contain so-called household variables. These are variables for which each member of the household has the same score, for example, household income, household size and city/town. If an unsafe combination with a household variable occurs for at least one person from a household and this household variable is locally suppressed, then this variable must be suppressed for all the

people in that household. In that case, values may therefore also be suppressed in safe records.

When selecting the variable that is going to be suppressed from a rare combination of scores on identifying variables, this choice is, in principle, free. However, two options are possible in μ-ARGUS.

First, the user can indicate, by assigning weights to variables, the extent to which the suppression of the score on that variable is desired (or not). μ-ARGUS then chooses to suppress those variables for which the sum of the weights is as small as possible. Consequently, it is possible, for example, to refrain (to a certain extent) from locally suppressing those variables that have already been adapted through other protection methods.

In the second option, μ-ARGUS uses a type of entropy argument to select the variable or variables to be suppressed. Each variable is then assigned the following weight:

$$w_X = -\sum_{i=1}^{K_X} \frac{f_X(i)}{n} \log \frac{f_X(i)}{n}, \qquad (1)$$

where K_X is the number of categories of variable X, n the number of records in the microdata file and $f_X(i)$ the number of records with score i on variable X. As a result, variables with larger numbers of categories are suppressed less frequently than variables with only a few categories.

2.4.4 Example

Figure 3 shows some records from a fictitious microdata file. The records are numbered for easy reference.

	Occupation	Region	Gender	Education	…
1	Mayor	Amsterdam	Man	High	…
2	Fisherman	Urk	Man	Low	…
3	Teacher	Amsterdam	Woman	High	…
4	Plumber	Papendrecht	Man	Medium	
	…	…	…	…	…

Figure 3: Some records from a fictitious microdata file

The mayor from Amsterdam is, obviously, unique. The variable ‘Region’ is now locally suppressed by replacing the city name by the score ‘Unknown’ in the unsafe records. This generates the records as shown in Figure 4.

	Occupation	Region	Gender	Education	…
1	Mayor	Unknown	Man	High	…
2	Fisherman	Urk	Man	Low	…

3	Teacher	Amsterdam	Woman	High	…
4	Plumber	Papendrecht	Man	Medium	
	…	…	…	…	…

Figure 4: Records from Figure 3 after local suppression of 'Region'

Now, in record 1, the mayor is no longer unique. Since the suppression is applied *locally*, the Region variable in the safe records 2 to 4 is not suppressed.

2.5 Top-coding

2.5.1 Brief description

When protecting microdata, most attention is paid to the treatment of the identifying variables. They play an important role in the disclosure control. The numerical variables are often the variables that are of interest to the data user (and also a possible discloser), such as income, etc. The actual income for instance of an average Dutch person does not identify this person to a significant extent, but that is not the case for people with an extremely high income. Suddenly, the variable of income has become an identifying variable, and therefore the need for extra protection must be assessed.

Top-coding is a suitable method in this situation. It is a simple method, in which values above a certain threshold are replaced by the same standard value. This can be an indication such as ('many') or ('> threshold'). However, the mean of all records with a value above that threshold can also be used. The advantage of this last choice is that the mean for the top-coded variable remains the same for all records.

In addition, bottom-coding can be used in an equivalent manner.

It is clear that top-coding is only useful for numerical variables. For qualitative variables, global recoding (see section 2.3) can be used to obtain a sort of top-coding.

2.5.2 Applicability

This method can be used as additional protection in those situations where some extremes of numerical variables must be considered as identifying.

2.6 Adding noise to weights

2.6.1 Brief description

If the file contains raising weights (to correct for the sample and/or non-response), the data protector must consider whether, using the information about the sample design, certain

information could be retrieved from those raising weights that could lead to disclosure. A well-known example is that the region is often used as a stratification variable in sampling. If, in the protection with global recoding (see section 2.3), the region information is limited or possibly completely removed, a consideration should be made as to whether information can still be derived about the region from the value of the weighting variable. If, for example, the city is replaced by the province, it is still possible that the raising weight could reveal that it concerns a large city. And therefore it might become clear which (suppressed) city information this relates to.

This type of disclosure can be avoided by adding sufficient noise to the raising weight. By adding random noise, the raising weight will generally still be useful in analyses.

2.6.2 Applicability

In such cases as indicated above, knowledge about the sample design could divulge information which could contribute to the disclosure of data. This method can help to prevent information about individual respondents from being disclosed from raising weights.

2.6.3 Detailed description

An implementation of this method is available in μ-ARGUS. A percentage p can be indicated, so that μ-ARGUS will replace the weight w_i by a random value from the interval

$$\left[\frac{(100-p)}{100}w_i, \frac{(100+p)}{100}w_i\right]. \qquad (2)$$

2.7 PRAM

2.7.1 Brief description

The Post Randomisation Method (PRAM) is a method for the statistical disclosure control of categorical variables. PRAM can be considered as an intentional misclassification, for which the misclassification probabilities are recorded by the data protector. PRAM is also related to the Randomised Response (RR) technique. However, RR is performed when the questions are being asked, while PRAM is applied after the answer has been provided/collected.

When PRAM is used, for each record in a microdata file, the score on one or more categorical variables is changed (or not changed) based on a certain probability. This is done independently on all the records. The probability mechanism that determines the transition of the scores is recorded in advance in a so-called Markov matrix.

Because PRAM is a stochastic method, the disclosure risk is directly affected: if a discloser believes that he or she recognises a record, there is a certain probability that this record does *not* correspond to the person that the discloser is thinking of. After all, several scores on identifying variables are changed with a certain probability.

The fact that the probability mechanism used is known when PRAM is applied means that it is possible, using the protected microdata and the Markov matrix, to construct unbiased estimators for certain statistical attributes of the original data. In addition, techniques from the misclassification and the Randomised Response can also be used.

For a detailed description of PRAM, we refer to Gouweleeuw et al. (1998a and 1998b).

2.7.2 Applicability

Most NSI policies for the statistical disclosure control of Scientific Use Files state that the identification of individual people must be prevented (or, in any case, it must be made more difficult). To identify an individual person, a discloser will have to use identifying variables, such as gender, marital status, age and educational level. Naturally, this only works if the discloser is certain that the variables in the file provided are actually the true scores. By applying PRAM to identifying variables, this certainty is eliminated: there is now a positive probability that the score is no longer the original score.

In the statistical disclosure control of a Scientific Use File, it is generally not possible to include very detailed regional variables. This is particularly the case if other detailed identifying variables are present in the file. In this situation, the traditional statistical disclosure control methods, such as recoding, top-coding and local suppression, would produce a file that is virtually unusable for analyses in which the regional detail is important. PRAM would then be a possible alternative: the detailed level is maintained, but the actual score on an identifying variable can no longer be assumed with certainty.

A user of a file that is protected using PRAM, however, must have sufficient statistical knowledge to be able to correct his or her desired analysis method for the changes made to the records. How these methods must be adapted is known for several analysis methods. See, for example, Gouweleeuw et al. (1998a and 1998b), Van den Hout (1999), Van den Hout and van der Heijden (2002) and Ronning et al. (2004).

Files that are protected using PRAM are thus mainly intended for theoretically or otherwise experienced statisticians. In addition, microdata files on which PRAM is applied can also be used as ‘test files’; for example, to test scripts or to determine research trends. The ultimate definitive analysis would then have to be performed on the original (unprotected) file by means of remote execution or an onsite session.

2.7.3 Detailed description

For a detailed theoretical description of the method, please refer to Gouweleeuw et al. (1998a and 1998b).

The Markov matrix with transition probabilities plays an important role in the application of PRAM. The transition probabilities determine the level of protection and affect the information loss. It is therefore important to properly select these probabilities. Each user will experience information loss in a different way. It is therefore preferable to keep the users' wishes in mind when determining the transition probabilities. De Wolf (2006) provides different measures for information loss.

Because PRAM is a stochastic disclosure control method, the standard formulas for disclosure risk are not directly applicable. However, alternative formulas are provided in, for example, De Wolf (2006).

It should be clear that the selection of the transition probabilities is not an easy task. There is no universal way to take the right decision in every situation. The following questions play a role in determining the transition probabilities:

- To which variables will PRAM be applied?
- Will PRAM be applied independently on all variables or on a subset thereof?
- Are there impossible combinations that must be prevented by setting the associated transition probabilities to zero?
- What effect does it have on the information loss?
- What effect does it have on the disclosure risk?

For each case, the answers to these questions will determine the selection of the specific transition probabilities. There is therefore no universal method available to determine the ideal transition probabilities.

For an empirical study into the consequences of different possibilities for the transition probabilities on both the disclosure risk and the information loss, please refer to De Wolf (2006).

When selecting a matrix of transition probabilities, a number of typical structures are possible. For example, a band matrix with bandwidth *b* can be useful for ordinal variables such as Age. In that case, an age can be replaced with a certain probability by an age within plus or minus *b* years. Completely filled matrices are mainly useful for nominal variables with a limited number of categories, such as the variable Marital status. See Figure 5 for a few examples.

$$\begin{pmatrix} 0.70 & 0.10 & 0.10 & 0.10 \\ 0.02 & 0.94 & 0.02 & 0.02 \\ 0.09 & 0.09 & 0.73 & 0.09 \\ 0.11 & 0.11 & 0.11 & 0.67 \end{pmatrix} \qquad \begin{pmatrix} 0.80 & 0.20 & 0 & 0 \\ 0.10 & 0.80 & 0.10 & 0 \\ 0 & 0.20 & 0.60 & 0.20 \\ 0 & 0 & 0.10 & 0.90 \end{pmatrix}$$

(a) (b)

Figure 5: Examples of matrices with transition probabilities: (a) Completely filled matrix, (b) Band matrix with bandwidth 1 (one)

For other variables, a block matrix is a more obvious solution. For example, for a variable such as Region (at city/town level), we can consider a block matrix in which the blocks correspond to the Provinces. In this way, cities can only be replaced by other cities from the same province. See Figure 6 for an example of a block matrix with transition probabilities.

$$\begin{pmatrix} 0.90 & 0.10 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.20 & 0.80 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.70 & 0.20 & 0.10 & 0 & 0 & 0 \\ 0 & 0 & 0.10 & 0.80 & 0.10 & 0 & 0 & 0 \\ 0 & 0 & 0.15 & 0.15 & 0.70 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.75 & 0.15 & 0.10 \\ 0 & 0 & 0 & 0 & 0 & 0.09 & 0.82 & 0.09 \\ 0 & 0 & 0 & 0 & 0 & 0.05 & 0.05 & 0.90 \end{pmatrix}$$

Figure 6: Example of a block matrix with three blocks with transition probabilities.

2.7.4 Example

At present, μ-ARGUS only offers a limited facility to perform statistical disclosure control using PRAM. In that package, it is possible to apply PRAM per variable, for which the Markov matrix may either be a band matrix or a completely filled matrix. The bandwidth of a band matrix is adjustable, the same as the diagonal probabilities (the probabilities that certain categories do *not* change). The R-package sdcMicro has the possibility to use a Markov matrix of general form.

Because PRAM is a stochastic disclosure control method (only the transition *probabilities* are recorded), a protected file may look different after every application of PRAM: such a protected file is, in any case, the outcome of a probability experiment. Analyses can therefore only be corrected *in expectation* for the fact that they are used on a file that is protected using PRAM. This means that, for example, the expectation for corrected estimated parameters will be the same as the parameter estimates based on the original file.

To obtain an impression of possible adaptations of analyses, consider the simple case of PRAM applied to the variable Gender (two categories), in which we want to estimate the frequency table for the number of men and the number of women. We notate the variable Gender by ξ, where $\xi = 1$ = Male and $\xi = 2$ = Female. We notate the associated frequency table by $\mathbf{T}_\xi$. Suppose that the original file contains 100 men and 100 women, so $\mathbf{T}_\xi = (100, 100)^t$. PRAM is applied to the variable Gender using the following matrix with transition probabilities:

$$\mathbf{P} = \begin{pmatrix} 0.90 & 0.10 \\ 0.20 & 0.80 \end{pmatrix}. \qquad (3)$$

Written out: the probability that the gender Male will be changed to Female is 10%, the probability that the gender Female will be changed to Male is 20%. The variable ξ is notated as X after the application of PRAM. The frequency table of Gender based on the protected file is then written $\mathbf{T}_X$. It can easily be derived that

$$E(\mathbf{T}_X | \xi) = \mathbf{P}^t \mathbf{T}_\xi, \qquad (4)$$

where the expectation is conditional on the original file. In the example, this means that it is expected that 110 men and 90 women will occur in the protected file. An unbiased estimator for the original frequency table follows directly from equation (4), i.e.

$$\hat{\mathbf{T}}_\xi = \left(\mathbf{P}^t\right)^{-1} \mathbf{T}_X. \qquad (5)$$

The original frequency table will only be reproduced *in expectation* by this corrected estimator. In other words,

$$E(\hat{\mathbf{T}}_\xi | \xi) = \mathbf{T}_\xi. \qquad (6)$$

Suppose that the protected file contains 112 men and 88 women (Please note: this is an example, because this can differ for each realisation of the probability experiment), then the unbiased estimation (rounded to whole numbers) would be represented by

$$\hat{\mathbf{T}}_\xi = \left(\mathbf{P}^t\right)^{-1} \mathbf{T}_X = \begin{pmatrix} 0.90 & 0.20 \\ 0.10 & 0.80 \end{pmatrix}^{-1} \begin{pmatrix} 112 \\ 88 \end{pmatrix} = \begin{pmatrix} 103 \\ 97 \end{pmatrix}. \qquad (7)$$

Note that this corrected estimation of the frequency table is much closer to the original frequency than the uncorrected estimation (the direct count from the protected file), but that the exact original values were not obtained.

2.8 Literature

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. and De Wolf, P.P. (1998a), *Post Randomisation for Statistical Disclosure Control: Theory and Implementation*, Journal of Official Statistics, vol. 14, 4, pp. 463 – 478.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. and De Wolf, P.P. (1998b), *The post randomisation method for protecting microdata*, Qüestiió, Quaderns d'Estadística i Investigació Operativa, vol. 22, 1, pp. 145 – 156.

Van den Hout, A. (1999), *The analysis of data perturbed by pram*, Delft University Press, Delft.

Van den Hout, A. and Van der Heijden, P.G.M. (2002), *Randomized response, statistical disclosure control and misclassification: a review*, International Statistical Review 70(2), pp. 269 – 288.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K. and De Wolf, P.P. (2012), *Statistical Disclosure Control (Wiley Series in Survey Methodology)*. John Wiley & Sons, ISBN 9781118348215.

Hundepool, A., v. d. Wetering, A., Ramaswamy, R., Franconi, L., Polettini, S., Capobianchi, A., de Wolf, P.P., Domingo, J., Torra, V., Brand, R. and Giessing, S. (2007), *μ-ARGUS user manual 4.1*, Voorburg.

Ronning, G., Rosemann, M. and Strotmann, H. (2004), *Estimation of the probit model using anonymized micro data*, Paper prepared for the 'European Conference on Quality and Methodology in Official Statistics (Q2004)', Mainz, 24–26 May 2004.

De Wolf, P.P. (2006), *Risk, Utility and PRAM*, in 'Privacy in Statistical Databases 2006', Domingo-Ferrer, J. and Franconi, L. (Eds.), LNCS 4302, Springer-Verlag, Berlin Heidelberg, pp. 189 – 204.

3. Statistical Disclosure Control of Quantitative Tables

3.1 General description and reading guide

3.1.1 General description

The statistical disclosure control of quantitative tables encompasses the production of quantitative tables that satisfy the policy for statistical disclosure control and, as such, can be published. Quantitative tables are tables in which the cell values are composed by summation of a continuous variable over all the contributors to a cell. This is in contrast to frequency tables in which only the *number* of contributors per cell is given. Other rules apply to frequency tables, and other protection methods may be more suitable than those for quantitative tables. Disclosure control methods for frequency tables are discussed in the next section 'Statistical Disclosure Control of Frequency Tables'.

If exactly one or two contributors produce a cell total, it is clear that this cell cannot be published. In the case of a single contributor, individual information is released directly, and in the case of two contributors, one contributor can exactly calculate the other contribution by subtracting his or her own contribution from the cell total.

However, undesirable situations can arise also if there are more than two contributors in a cell. In principle, in the statistical disclosure control of quantitative tables, we must prevent (or at least make it more difficult) that any individual contribution can be estimated too accurately. This may occur, for example, in the case that a very large contributor is present in a single cell along with several relatively small contributors. In this case, the second-largest contributor can calculate that the largest contribution does not contribute more than the cell total minus the second-largest contribution to the cell. A relatively good estimate of the contribution of the largest contributor can be obtained as a result, usually in conflict with the disclosure control rules of the NSI.

The presence of empty cells also requires extra attention. In some cases, an empty cell will be a so-called *structural zero cell*. This means that it is generally known that, logically, it is *impossible* for this cell to have a contribution. Such cells can therefore not be used in the disclosure control: whatever you do, everyone knows that they must be empty cells.

At the same time, reliable information can sometimes be disclosed using *non-structural zero cells*. If there are contributors in such a cell, there is actually a sort of group disclosure: it is immediately clear that all the contributors have provided a contribution of zero (assuming that the contributions are non-negative). If there are no contributors in the cell, but it is not impossible for a contributor to be in this cell, this in itself also reveals direct information.

The methods described in this section can be easily applied using the software package τ-ARGUS. This package was developed in a European context.

3.1.2 Reading guide

As the initial step in determining the correct statistical disclosure control for a quantitative table, it will first have to be determined whether disclosure is possible. In the first instance, the basis for this is 'common sense': is there information present in the table that may not be disclosed about individual respondents? For quantitative tables, such information is generally a respondent's individual contribution to the total of a specific cell in the table.

In addition, an objective method is needed to determine which cells in the table contain respondents that potentially run a risk of their individual contribution being disclosed. The p% rule (see section 3.3) is intended to identify such primary unsafe cells. This method can only be used for quantitative tables and not for frequency tables.

After the unsafe cells have been identified, the table will generally have to be protected. There are three general methods available for this purpose: restructuring the table (see section 3.4), suppressing cells (see section 3.5) and rounding (see section 3.6). The relatively new method Controlled Tabular Adjustment (CTA) is not described in this text. A detailed description can be found in Hundepool et al. (2012).

Which method or combination of methods is ultimately used in a specific situation cannot be determined in advance. This depends to a significant extent on the intended users. For example, in Eurostat regulations, it is not always possible to restructure the table, and cell suppression will often have to be chosen. When selecting the method or methods to be used, two competing aspects must be taken into account:

- disclosure risk;
- information loss.

In general, it can be said that reducing the disclosure risk will lead to increased information loss. The converse is also true: the smaller the information loss, the larger the disclosure risk.

3.2 Scope and relationship with other sections

This section discusses methods that can be used for the statistical disclosure control of quantitative tables. This chapter does *not* discuss any methods that can only be used for frequency tables. For such methods, see the section 'Statistical Disclosure Control of Frequency Tables'.

A few of the methods described in this chapter can, in principle, be used for quantitative tables as well as for frequency tables. Such methods will only be mentioned briefly in the section 'Statistical Disclosure Control of Frequency Tables'.

The methods in this chapter can be divided into two variants: methods for determining the primary unsafe cells of a quantitative table and methods for making tables with unsafe cells suitable for publication.

3.3 *P* % rule

3.3.1 Brief description

The goal of statistical disclosure control is to prevent the disclosure of information about individual contributors to a table, or at least to make this more difficult. To achieve this, the cells where there is a risk of possible disclosure will first have to be identified. An objective measure is needed for this, one that indicates how well an individual contribution to a cell can be estimated based on the published table. The *p* % rule provides for this. This is also the method to use to indicate to what extent the disclosure control rule has been violated and how large the measures to be taken must be.

3.3.2 Applicability

Before a statistical disclosure control method can be applied to a quantitative table, it must first be indicated where potential problems occur in that table. The *p* % rule indicates how well a contributor in a cell would be able to estimate another contributor in that same cell. This serves to determine the so called primary unsafe cells, and it also gives an indication how much protection must be provided to satisfy the policy for the publication of quantitative tables.

With this method, account can also be taken of possible authorisations/waivers: contributors who have indicated that they do not object to publications from which their contribution can be derived. Such contributors are then simply excluded in the application of the *p* % rule.

The *p* % rule may only be used:

- in the case of quantitative tables;
- with non-negative contributors;
- for which the largest contributors are identifiable for the discloser;
- on non-empty cells with a positive cell total.

3.3.3 Detailed description

Let T_A be the cell value of cell A in the table in question. Denote the largest contributor without a waiver by X_s and the largest of the remaining contributors by X_r. Then cell T_A is unsafe if:

$$\frac{(T_A - X_r) - X_s}{X_s} < \frac{p}{100}. \qquad (8)$$

That is, in the situation of no waivers, a cell is unsafe if the second-largest contributor can estimate the largest contributor with an accuracy exceeding *p* %.

It is simple to see that this is the worst case scenario: if the second-largest contributor *cannot* estimate the largest contributor more accurately than *p* %, then no other contributor can estimate an arbitrary other contributor more accurately than *p* %, and therefore the cell is safe. In other words: the most accurate estimation can be made by the second-largest contributor, when this party estimates the largest contribution.

The value of the difference between the left side and the right side of the inequality in formula (8) also indicates how much protection an unsafe cell needs. For more detail, please refer to Loeve (2001).

With software τ-ARGUS, it is easy to apply the *p* % rule. This method is one of the standard built-in rules that can be used to identify the primary unsafe cells. Moreover, τ-ARGUS automatically calculates how much protection an unsafe cell needs and uses that in the further protection of the table concerned. To make it possible for τ-ARGUS to identify the primary unsafe cells using the *p* % rule, however, it is necessary that the input for τ-ARGUS consists of the microdata from which the table concerned is composed. To apply the *p* % rule, information is needed, in any case, about the individual contributors. For more information about the use of τ-ARGUS, please refer to the associated manual.

The value selected for *p* is determined by for instance the NSI's policy. That policy could for example give an interval within which *p* should be selected (e.g., $5 \leq p \leq 15$). The exact value for *p* is determined by the producer of the table and may never be revealed to external parties, because this could help them in the calculation of the suppressed cells.

A large value for *p* results in strict disclosure control, because, when estimating an arbitrary contribution in that case, not even a relatively 'large' error is allowed. A small value for *p* results in less strict disclosure control, because a cell is only unsafe in this situation if a contribution can be estimated very accurately.

3.3.4 Example

In this fictitious example, we look at a cell in a table with turnover according to SBI (the Dutch Standard Industrial Classification) and Region. Suppose that the cell with 'SBI = 32' and 'Region = Noord Brabant' consists of four contributors with the values 324, 4, 2 and 10. Suppose that we want to use the *p* % rule where $p = 5$, then we must first sort the contributors: $X_1 = 324$, $X_2 = 10$, $X_3 = 4$ and $X_4 = 2$. The cell total T_A is then 340. If we calculate the quotient from formula (8), we obtain the value 0.0185. This is clearly smaller than 5 %, and therefore the cell is unsafe.

3.4 Table restructuring

3.4.1 Brief description

Section 3.3 describes when a cell in a table must be considered unsafe. In general, cells with a limited number of contributors or a cell with one or two large contributors are the obvious candidates to be characterised as unsafe. All unsafe cells must be protected. Before performing suppression on a large scale, restructuring the table can also be considered. By combining rows and/or columns, cells are pooled and the content per cell is increased. The result of this is that fewer cells are identified as unsafe by the *p* % rule, as described in section 3.3.

3.4.2 Applicability

This method will generally lead to fewer unsafe cells in the table. Combining cells creates cells that are safer than the individual cells before they were combined.

There are no methodological conditions for using this method. However, externally imposed obligations sometimes specify what detail level of a table must be published. This may be a Eurostat obligation, but the policy of the NSI can also mean that a certain detailed level of a table must be published. In these cases, the method can be applied from a technical perspective, but its use is prevented by (external) policy decisions.

Furthermore, an assessment must be made between the information loss resulting from the larger number of crosses (suppressed cells) that are needed to protect the table, and the information loss resulting from combining columns/rows, for which fewer suppressions are needed.

3.4.3 Detailed description

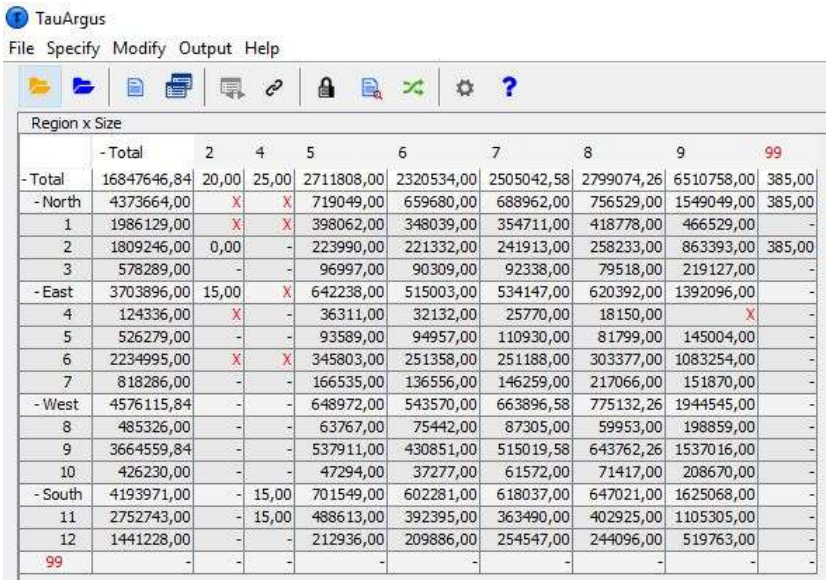
The software package τ-ARGUS has provisions for recoding rows and/or columns in tables. In this regard, a distinction is made between two situations:

- In the case of a hierarchical spanning variable, the recoding implies that certain splits are omitted at the lowest level.
- In the case of an unstructured spanning variable, users are free to combine the columns or rows of a table as they choose.

3.4.4 Example

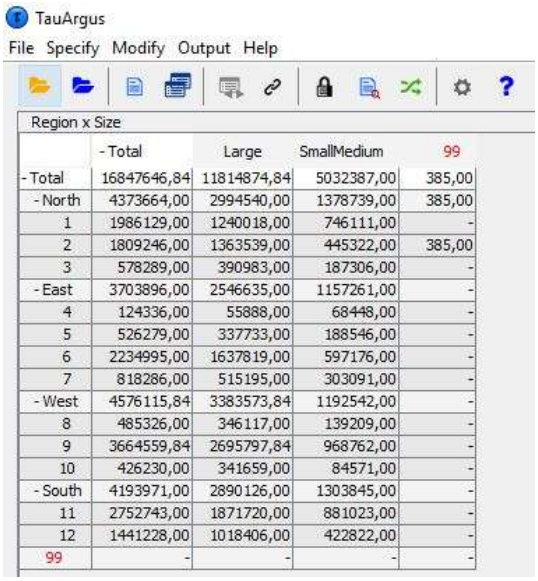
Figure 7 presents a fictitious table of the turnover according to Region (rows, hierarchical) and SizeClass (columns). The unsafe cells according to e.g. a *p*%-rule, are denoted with a red '×'. Figure 8 provides two possible restructuring possibilities for this table. In the first variant the variable SizeClass is recoded such that the categories 2 to 6 are combined into

the category MediumSmall, and that the categories 7, 8 and 9 are combined into the category Large. Note that, in this way, all the primary unsafe cells are combined to create safe cells. For the other possibility as a result of the recoding of the variable Region the smallest detail level has been removed. This restructuring does not resolve all the problems: the primary unsafe cells for regions North and East are still present in the table.



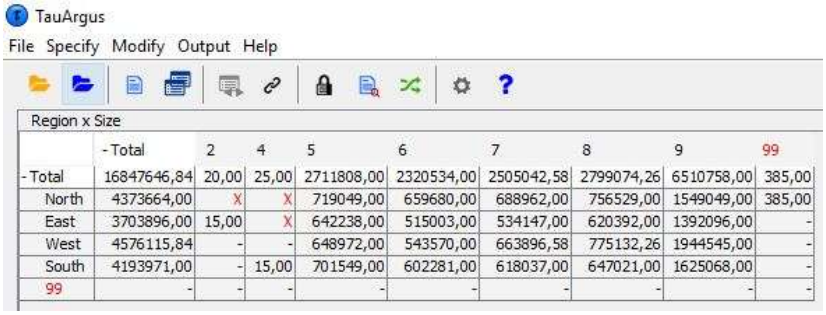
	- Total	2	4	5	6	7	8	9	99
- Total	16847646,84	20,00	25,00	2711808,00	2320534,00	2505042,58	2799074,26	6510758,00	385,00
- North	4373664,00	X	X	719049,00	659680,00	688962,00	756529,00	1549049,00	385,00
1	1986129,00	X	X	398062,00	348039,00	354711,00	418778,00	466529,00	-
2	1809246,00	0,00	-	223990,00	221332,00	241913,00	258233,00	863393,00	385,00
3	578289,00	-	-	96997,00	90309,00	92338,00	79518,00	219127,00	-
- East	3703896,00	15,00	X	642238,00	515003,00	534147,00	620392,00	1392096,00	-
4	124336,00	X	-	36311,00	32132,00	25770,00	18150,00	X	-
5	526279,00	-	-	93589,00	94957,00	110930,00	81799,00	145004,00	-
6	2234995,00	X	X	345803,00	251358,00	251188,00	303377,00	1083254,00	-
7	818286,00	-	-	166535,00	136556,00	146259,00	217066,00	151870,00	-
- West	4576115,84	-	-	648972,00	543570,00	663896,58	775132,26	1944545,00	-
8	485326,00	-	-	63767,00	75442,00	87305,00	59953,00	198859,00	-
9	3664559,84	-	-	537911,00	430851,00	515019,58	643762,26	1537016,00	-
10	426230,00	-	-	47294,00	37277,00	61572,00	71417,00	208670,00	-
- South	4193971,00	-	15,00	701549,00	602281,00	618037,00	647021,00	1625068,00	-
11	2752743,00	-	15,00	488613,00	392395,00	363490,00	402925,00	1105305,00	-
12	1441228,00	-	-	212936,00	209886,00	254547,00	244096,00	519763,00	-
99	-	-	-	-	-	-	-	-	-

Figure 7: Quantitative table for turnover according to region and size class



	- Total	Large	SmallMedium	99
- Total	16847646,84	11814874,84	5032387,00	385,00
- North	4373664,00	2994540,00	1378739,00	385,00
1	1986129,00	1240018,00	746111,00	-
2	1809246,00	1363539,00	445322,00	385,00
3	578289,00	390983,00	187306,00	-
- East	3703896,00	2546635,00	1157261,00	-
4	124336,00	55888,00	68448,00	-
5	526279,00	337733,00	188546,00	-
6	2234995,00	1637819,00	597176,00	-
7	818286,00	515195,00	303091,00	-
- West	4576115,84	3383573,84	1192542,00	-
8	485326,00	346117,00	139209,00	-
9	3664559,84	2695797,84	968762,00	-
10	426230,00	341659,00	84571,00	-
- South	4193971,00	2890126,00	1303845,00	-
11	2752743,00	1871720,00	881023,00	-
12	1441228,00	1018406,00	422822,00	-
99	-	-	-	-

(a) Recoding of SizeClass (all primary unsafe cells have been protected)



	- Total	2	4	5	6	7	8	9	99
- Total	16847646,84	20,00	25,00	2711808,00	2320534,00	2505042,58	2799074,26	6510758,00	385,00
North	4373664,00	X	X	719049,00	659680,00	688962,00	756529,00	1549049,00	385,00
East	3703896,00	15,00	X	642238,00	515003,00	534147,00	620392,00	1392096,00	-
West	4576115,84	-	-	648972,00	543570,00	663896,58	775132,26	1944545,00	-
South	4193971,00	-	15,00	701549,00	602281,00	618037,00	647021,00	1625068,00	-
99	-	-	-	-	-	-	-	-	-

(b) Recoding of Region (not all primary unsafe cells have been protected)

Figure 8: Two possible restructuring possibilities used on the table from Figure 7

3.5 Cell suppression

3.5.1 Brief description

A frequently used method to protect primary unsafe cells is to suppress (not publish) certain cells. The cell value is then simply replaced by an '×'.

In a quantitative table when the marginals are also provided, however, it is often not sufficient to suppress only the primary unsafe cells. If a suppressed cell is the only suppressed cell in a row, the suppressed value can, after all, simply be calculated by subtracting the other cell values in that row from the corresponding row-total.

To sufficiently protect primary unsafe cells, it is therefore also necessary to suppress other cells which, in themselves, are safe. This is called *secondary suppression*. It is not easy to perform this in such a way that the primary unsafe cells are protected sufficiently while also ensuring that not too much information is removed from the table. Furthermore,

account must also be taken of the fact that structural zero cells cannot be used as secondary suppressions: everyone knows that, by definition, these cells are empty.

To prevent a situation where suppressed, primary unsafe cells can be calculated exactly, secondary suppressions are therefore necessary. However, also a 'too accurate' estimation for a suppressed cell is not desirable. Indeed, what is the difference between the following statements: 'This suppressed cell actually has a value of 10000' and 'This suppressed cell actually has a value of between 9998 and 10002'. Given a suppression pattern, it is always[2] possible to calculate an interval in which a suppressed cell must lie. The method of 'Cell Suppression' must then also produce a suppression pattern for which the intervals that can be calculated are sufficiently large. The size of these intervals is determined by the rule that is used to determine the primary unsafe cells.

Fischetti and Salazar (2000) have developed a method to solve the above problem in an optimal manner. Their method is, in theory, applicable to arbitrary, additive tables with non-negative contributors. In practice, however, their solution involves too much computing time if the tables become too large, either in size or complexity. This is why a number of suboptimal methods have been developed to find suitable suppression patterns for larger and/or more complex tables.

For example, the 'modular approach' (a.k.a. HiTaS) splits a hierarchical table into a large number of non-hierarchical subtables and applies the optimal method to each individual subtable. By correctly combining the results, a suboptimal solution can be obtained for the entire table, with a significantly shorter computing time. See De Wolf (2002).

The 'hypercube approach' can also protect large tables by protecting the subtables in a certain iterative way. The protection of each subtable also takes place suboptimally. Consequently, the approach is relatively fast, but, in general, more cells are suppressed than strictly necessary to obtain a protected table. See Giessing and Repsilber (2002).

3.5.2 Applicability

This method can be used to adequately protect quantitative tables with cells that do not satisfy the requirements of the statistical disclosure control policy. In particular, if the table cannot be restructured further or at all, the cell suppression method can be used effectively.

The contributions to the table to be protected must be non-negative, the table must be additive, and the marginals must also be provided.

In the modular approach as implemented in τ-ARGUS, the table must be four-dimensional at a maximum. Each dimension may be hierarchical. Linked tables can be protected by copying the suppressions from one table to the other, and then protecting the tables. This

[2] In the case that the table is composed of non-negative contributors and the marginals are also given.

should then possibly be performed in an iterative manner. For a certain class of linked tables, it is possible to solve the linked tables problem automatically.

In the hypercube approach as implemented in τ-ARGUS, the table may be seven-dimensional at a maximum. The table may be hierarchical in every dimension. Linked tables are also possible in principle.

It should be mentioned that for both approaches, from a performance perspective, the recommendation is to avoid using long, unstructured (non-hierarchical) code lists.

3.5.3 Detailed description

The software package τ-ARGUS has a provision to apply cell suppression to quantitative tables. If the original microdata is used as input, τ-ARGUS will determine the primary unsafe cells with the associated safety intervals (see also section 3.3).

After this, τ-ARGUS will have to determine a suppression pattern that guarantees the necessary safety intervals. There are various options for this. We will discuss the two approaches that are the most interesting for NSIs.

3.5.3.1 Modular approach

For a detailed description and an elaborated example of the modular approach, see De Wolf (2002).

Generally, the modular approach can be described as follows:

1. Split the hierarchical table into all logical non-hierarchical subtables.
2. Group the subtables in classes in such a way that all tables in a single class can be protected independently of each other. For a suitable classification, see De Wolf (2002).
3. Protect all tables in class *K*.
4. If no secondary suppressions are placed in the marginals of the subtables of class *K*, continue with class $K + 1$, including any secondary suppressions in the inside of a table as primary suppressions for class $K + 1$.
5. If secondary suppressions do have to be placed in a marginal of at least one subtable, go back to class $K - 1$, including only the secondary suppressions in the marginals as primary suppressions.
6. Repeat steps 4 and/or 5 until all subtables have been protected at the lowest (most detailed) hierarchical level.

All non-hierarchical subtables will be protected using the mixed integer approach from Fischetti and Salazar (2000). In this approach, the required safety intervals are guaranteed, while a certain cost function is minimised. This cost function can be selected in different

ways, as a result of which various forms of information loss can be minimised. This minimisation takes place *locally*, so that the ultimate solution for the entire (hierarchical) table does not necessarily also have to be optimal.

In selecting the cost function in τ-ARGUS, several options can be selected, including:

- A variable from the dataset (such as the quantitative value on which tabulation takes place);
- A constant (so that the number of suppressions is minimised);
- The number of contributors per cell (so that the total number of suppressed contributions is minimised).

In the disclosure control of a subtable, also the so-called singletons problem must be taken into account: cells with only one contribution. If such cells are in a suppression pattern, the contributors involved can reverse part or all of the suppression pattern. After all, they know what their own contribution is and can therefore fill in that suppressed value, as a result of which it may also be possible to calculate other suppressed cells. In the current implementation of the mixed integer approach in τ-ARGUS, it is not possible to keep each conceivable combination of a singleton with another suppressed cell under control while searching for a suppression pattern. However, it is possible to take account of the combinations within a single row, column or layer[3] in the table. The combinations which must be taken into account consist of exactly two primary unsafe cells in a single row, column or layer, of which at least one cell is a singleton. A 'virtual' cell is constructed that consists of the combined two primary unsafe cells. It is prevented that that 'virtual' cell can be recalculated, by adding that 'virtual' cell to the table-structure.

3.5.3.2 Hypercube approach

For a more detailed description of the hypercube approach, see Giessing and Repsilber (2002).

In this approach too, a hierarchical table is split into non-hierarchical subtables. The non-hierarchical subtables are then protected in a certain order, where the subtables at the highest level are dealt with first.

For each subtable, all possible hypercubes are constructed for each primary unsafe cell in which that primary unsafe cell is one of the corner points. For each hypercube, the interval is calculated around the primary unsafe cell if all other corner points of the hypercube are also suppressed. If that interval is large enough (depending on the protection rule used), the associated hypercube is designated as 'feasible'. The information loss is then calculated for

[3] A row in a three dimensional table consists of the cells with coordinates (*r, k, l*) where *k* and *l* are fixed. A column then consists of the cells with the coordinates (*r, k, l*) where *r* and *l* are fixed. A layer then consists of the cells with coordinates (*r, k, l*) where *r* and *k* are fixed.

each feasible hypercube. Finally, the admissible hypercube with the smallest information loss is selected to protect the primary unsafe cell concerned.

No linear programming problem needs to be solved in order to calculate the safety intervals resulting from a hypercube. This significantly accelerates the procedure. The hypercube approach is therefore, in general, faster than the modular approach, for which a mixed integer programming problem needs to be solved.

After all subtables are protected in this way, the entire procedure is repeated. Secondary suppressed cells from a certain subtable that also occur in other subtables are considered as primary unsafe cells in those other subtables, and dealt with as such. This process is repeated until no more changes take place.

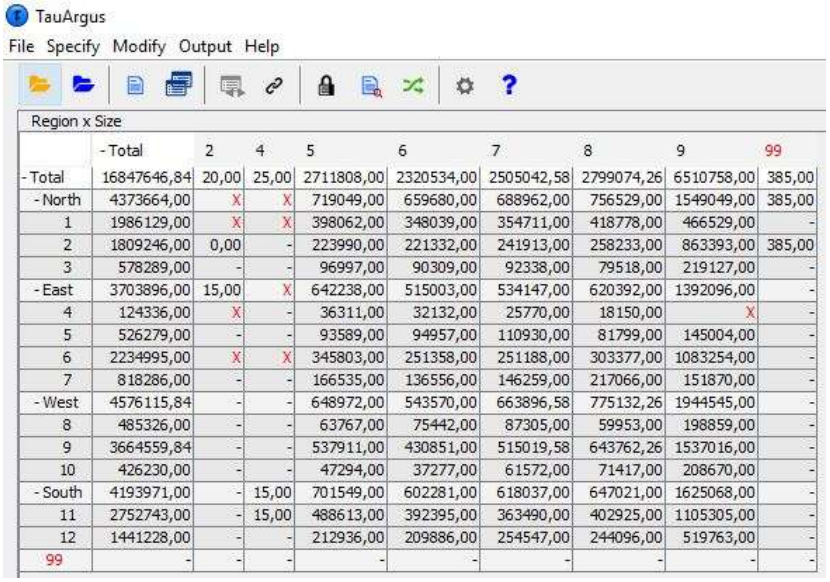
Note that the use of hypercubes to protect primary unsafe cells is a sufficient but not necessary condition for a safe suppression pattern. In other words, in some cases, the combination of the different hypercubes will not lead to an optimal suppression pattern, but it will always produce a safe suppression pattern. Consequently, this approach tends to suppress more cells than necessary for a safe suppression pattern.

This approach also takes account of the so-called singletons. A cell with only one contributor would indeed allow all suppressed corner points of a hypercube to be calculated. Therefore the extra requirement in the case of singletons is that this type of cell must be a corner point of at least two different hypercubes.

3.5.4 Example

Using τ-ARGUS, it is easy to apply cell suppression to a quantitative table. Both the modular approach and the hypercube approach are implemented in τ-ARGUS. It is also possible to select multiple information loss measures for the cost function that must be minimised. For the use of τ-ARGUS, please refer to the associated manual (Hundepool et al., 2003).

Figure 9 presents an example of a table in which only the primary unsafe cells are suppressed.

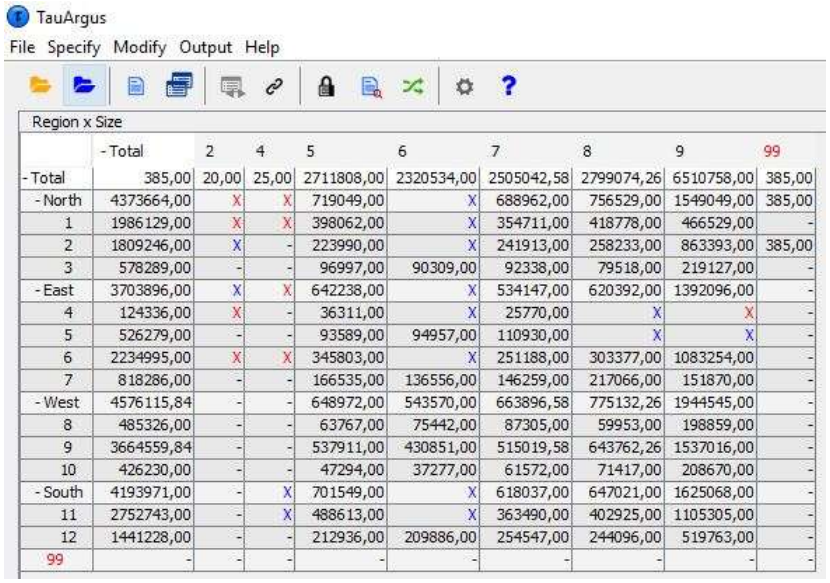


	- Total	2	4	5	6	7	8	9	99
- Total	16847646,84	20,00	25,00	2711808,00	2320534,00	2505042,58	2799074,26	6510758,00	385,00
- North	4373664,00	X	X	719049,00	659680,00	688962,00	756529,00	1549049,00	385,00
1	1986129,00	X	X	398062,00	348039,00	354711,00	418778,00	466529,00	-
2	1809246,00	0,00	-	223990,00	221332,00	241913,00	258233,00	863393,00	385,00
3	578289,00	-	-	96997,00	90309,00	92338,00	79518,00	219127,00	-
- East	3703896,00	15,00	X	642238,00	515003,00	534147,00	620392,00	1392096,00	-
4	124336,00	X	-	36311,00	32132,00	25770,00	18150,00	X	-
5	526279,00	-	-	93589,00	94957,00	110930,00	81799,00	145004,00	-
6	2234995,00	X	X	345803,00	251358,00	251188,00	303377,00	1083254,00	-
7	818286,00	-	-	166535,00	136556,00	146259,00	217066,00	151870,00	-
- West	4576115,84	-	-	648972,00	543570,00	663896,58	775132,26	1944545,00	-
8	485326,00	-	-	63767,00	75442,00	87305,00	59953,00	198859,00	-
9	3664559,84	-	-	537911,00	430851,00	515019,58	643762,26	1537016,00	-
10	426230,00	-	-	47294,00	37277,00	61572,00	71417,00	208670,00	-
- South	4193971,00	-	15,00	701549,00	602281,00	618037,00	647021,00	1625068,00	-
11	2752743,00	-	15,00	488613,00	392395,00	363490,00	402925,00	1105305,00	-
12	1441228,00	-	-	212936,00	209886,00	254547,00	244096,00	519763,00	-
99	-	-	-	-	-	-	-	-	-

Figure 9: Quantitative table for turnover according to region and size class

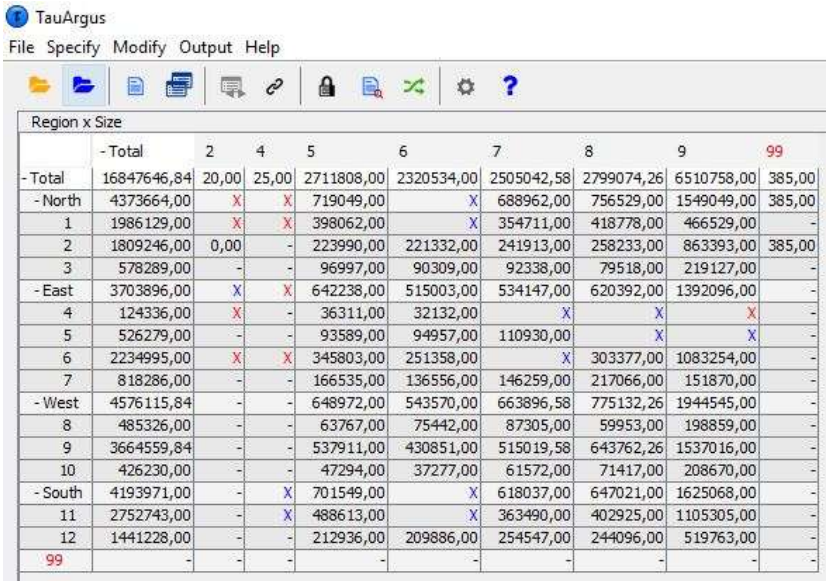
It is clear that this is not sufficient: both the cell (East, 4) and the cell (4, 9) can be directly calculated: (East, 4) = 3 703 896 – 15 – 642 238 – 515 003 – 534 147 – 620 392 – 1 392 096 = 5 and (4, 9) = 1 392 096 – 145 004 – 1 083 254 – 151 870 = 11 968.

Figure 10 shows the suppression pattern that was determined with τ-ARGUS using the hypercube approach. Figure 11 shows the same based on the modular approach. The primary suppressions are denoted with a red '×', the secondary suppressions with a blue '×'. Of course, in a publication, it should be impossible to make a distinction between primary and secondary suppressions.



	- Total	2	4	5	6	7	8	9	99
- Total	385,00	20,00	25,00	2711808,00	2320534,00	2505042,58	2799074,26	6510758,00	385,00
- North	4373664,00	X	X	719049,00	X	688962,00	756529,00	1549049,00	385,00
1	1986129,00	X	X	398062,00	X	354711,00	418778,00	466529,00	-
2	1809246,00	X	-	223990,00	X	241913,00	258233,00	863393,00	385,00
3	578289,00	-	-	96997,00	90309,00	92338,00	79518,00	219127,00	-
- East	3703896,00	X	X	642238,00	X	534147,00	620392,00	1392096,00	-
4	124336,00	X	-	36311,00	X	25770,00	X	X	-
5	526279,00	-	-	93589,00	94957,00	110930,00	X	X	-
6	2234995,00	X	X	345803,00	X	251188,00	303377,00	1083254,00	-
7	818286,00	-	-	166535,00	136556,00	146259,00	217066,00	151870,00	-
- West	4576115,84	-	-	648972,00	543570,00	663896,58	775132,26	1944545,00	-
8	485326,00	-	-	63767,00	75442,00	87305,00	59953,00	198859,00	-
9	3664559,84	-	-	537911,00	430851,00	515019,58	643762,26	1537016,00	-
10	426230,00	-	-	47294,00	37277,00	61572,00	71417,00	208670,00	-
- South	4193971,00	-	X	701549,00	X	618037,00	647021,00	1625068,00	-
11	2752743,00	-	X	488613,00	X	363490,00	402925,00	1105305,00	-
12	1441228,00	-	-	212936,00	209886,00	254547,00	244096,00	519763,00	-
99	-	-	-	-	-	-	-	-	-

Figure 10: Suppression pattern for the table from Figure 9, using the hypercube approach



	- Total	2	4	5	6	7	8	9	99
- Total	16847646,84	20,00	25,00	2711808,00	2320534,00	2505042,58	2799074,26	6510758,00	385,00
- North	4373664,00	X	X	719049,00	X	688962,00	756529,00	1549049,00	385,00
1	1986129,00	X	X	398062,00	X	354711,00	418778,00	466529,00	-
2	1809246,00	0,00	-	223990,00	221332,00	241913,00	258233,00	863393,00	385,00
3	578289,00	-	-	96997,00	90309,00	92338,00	79518,00	219127,00	-
- East	3703896,00	X	X	642238,00	515003,00	534147,00	620392,00	1392096,00	-
4	124336,00	X	-	36311,00	32132,00	X	X	X	-
5	526279,00	-	-	93589,00	94957,00	110930,00	X	X	-
6	2234995,00	X	X	345803,00	251358,00	X	303377,00	1083254,00	-
7	818286,00	-	-	166535,00	136556,00	146259,00	217066,00	151870,00	-
- West	4576115,84	-	-	648972,00	543570,00	663896,58	775132,26	1944545,00	-
8	485326,00	-	-	63767,00	75442,00	87305,00	59953,00	198859,00	-
9	3664559,84	-	-	537911,00	430851,00	515019,58	643762,26	1537016,00	-
10	426230,00	-	-	47294,00	37277,00	61572,00	71417,00	208670,00	-
- South	4193971,00	-	X	701549,00	X	618037,00	647021,00	1625068,00	-
11	2752743,00	-	X	488613,00	X	363490,00	402925,00	1105305,00	-
12	1441228,00	-	-	212936,00	209886,00	254547,00	244096,00	519763,00	-
99	-	-	-	-	-	-	-	-	-

Figure 11: Suppression pattern for the table from Figure 9, using the modular approach

3.5.5 Quality indicators

If a table is protected by means of cell suppression, it is possible to calculate the realised safety interval for each suppressed cell. Given the suppression pattern and the structure of the table, two LP-problems must be solved for each suppressed cell (minimising and maximising the value for the suppressed cell).

If τ-ARGUS is used for the protection of a quantitative table, at the end of the session, a report is generated that contains the steps taken and the associated results. It is also possible during the session to obtain information about the protected or unprotected table (for example: the number of primary unsafe cells, the number of secondary suppressions, information loss).

3.6 Additive rounding

3.6.1 Brief description

When rounding cell values in a quantitative table, the exact cell values are only known within a certain interval. A table with primary unsafe cells can also be protected in this way. The extent to which rounding is performed will, of course, have an impact on the size of the intervals. If each cell is rounded independently, the additivity of the table will not necessarily be maintained.

Of course, there is a simple way to guarantee the additivity: by rounding the cells in the interior of the table independently of each other and then recalculating the marginals. As a result, however, the marginals can deviate significantly from the rounded original values.

In additive rounding, the table is rounded such that the additivity is maintained and that the rounded table deviates from the original as little as possible. Furthermore, it is possible to perform additive rounding in such a way that safety intervals specified in advance can also be guaranteed. Whether this can be achieved, however, depends on the size of the selected rounding base in relation to the safety intervals.

3.6.2 Applicability

Additive rounding can be used for the statistical disclosure control of both quantitative tables and frequency tables. Often, a presentation argument will also play a role: a large number of significant figures suggests a high degree of precision that is not always justified because of sampling errors and measurement errors. Rounding the table values reduces this false precision to a certain extent.

3.6.3 Detailed description

In additive rounding, the cell values in a table are rounded to multiples of a rounding base b, keeping the totals and subtotals in the table equal to the sum of the corresponding parts.

Oftentimes, additive rounding is performed in a 'zero restricted' manner. In other words, cell values that are already a multiple of the rounding base are not changed, while the other cell values are rounded to one of the adjacent multiples of that rounding base. The rounded values are selected such that the sum of the absolute deviations of the cell values in the rounded table with respect to the cell values in the original table is minimised, under the restriction that the rounded table remains additive. As a result, it is possible that cell values are not rounded to the closest multiple of the rounding base.

In certain conditions, it is not possible to construct a rounded table under the scenario described above. In that case, the restriction that rounding is performed to one of the adjacent multiples of the rounding base is weakened by allowing a cell value to also be rounded to non-adjacent multiples of the rounding base. This weakening can be limited slightly by determining a maximum for the number of steps that may exist between the rounded value and the original value.

In the case of 'zero restricted' additive rounding using rounding base $b > 0$ of the non-negative number $z = ub + r$, where $0 \leq r < b$, rounding is performed on the number a, such that

$$a \in \{ub, (u + 1_{(0,b)}(r))b\} \qquad (9)$$

where $1_{(0,b)}(r)$ is equal to 1 if $r \in (0, b)$ and equal to 0 if $r = 0$.

This means that, in the case that $r = 0$, a is always rounded to ub and, in the case that $r \in (0, b)$, a is always rounded to ub or to $(u + 1)b$.

If, however, the restriction is weakened by a maximum of $K > 0$ steps further than the adjacent multiples of the rounding base, then rounding is performed on the number a, such that

$$a \in \left\{ \left(0 \vee (u + j) \right) b \mid j = -K, \ldots, \left(K + 1_{(0,b)}(r) \right) \right\} \qquad (10)$$

where $x \vee y = \max(x, y)$.

Multiple additive rounded versions may exist for a given table. These are all *feasible* tables. The table closest to the original table can subsequently be selected from the feasible tables. In τ-ARGUS, the distance that is minimised is represented by

$$\sum_{i=1}^{N} \left| z_i - a_i \right| \qquad (11)$$

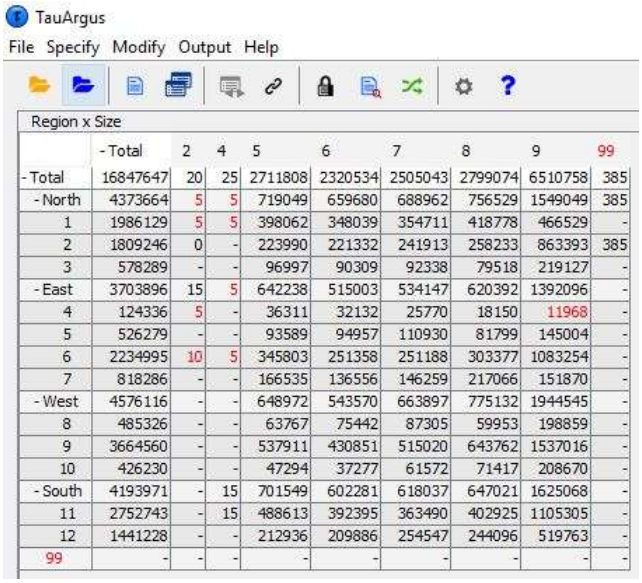
where N is the number of cells in the table (including all totals and subtotals), z_i the cell values in the original table and a_i the corresponding rounded cell values.

Finding the optimal solution is a problem that requires intensive computation (NP-complete). For large tables, this can result in unacceptably long calculation times. Partitioning is built into τ-ARGUS for this reason: a large table can be split into a number of subtables that are rounded individually. After these subtables are rounded, they are combined, calculating (if necessary) the totals and subtotals in question from the rounded parts.

3.6.4 Example

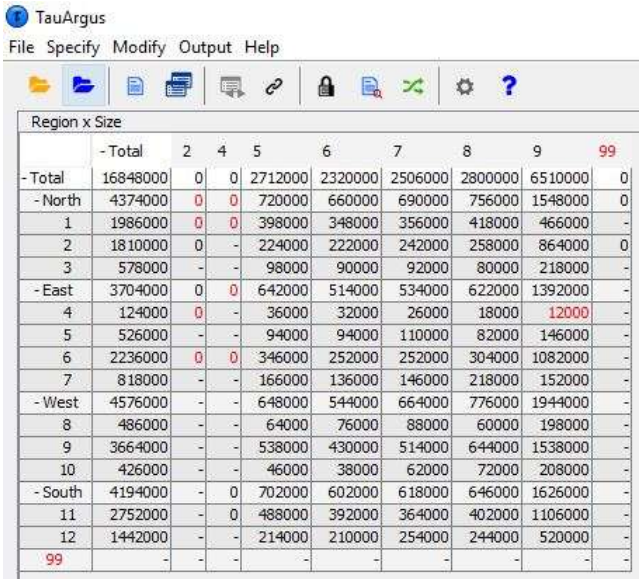
τ-ARGUS can be used to easily perform additive rounding on quantitative tables, while the desired protection margins are guaranteed.

Figure 12 presents an example of a table that contains a number of primary unsafe cells (marked red). Figure 13 contains the associated additively rounded table, with a rounding base of 2000. Of course, in a publication, the primary unsafe cells are not allowed to be recognisable.



	- Total	2	4	5	6	7	8	9	99
- Total	16847647	20	25	2711808	2320534	2505043	2799074	6510758	385
- North	4373664	5	5	719049	659680	688962	756529	1549049	385
1	1986129	5	5	398062	348039	354711	418778	466529	-
2	1809246	0	-	223990	221332	241913	258233	863393	385
3	578289	-	-	96997	90309	92338	79518	219127	-
- East	3703896	15	5	642238	515003	534147	620392	1392096	-
4	124336	5	-	36311	32132	25770	18150	11968	-
5	526279	-	-	93589	94957	110930	81799	145004	-
6	2234995	10	5	345803	251358	251188	303377	1083254	-
7	818286	-	-	166535	136556	146259	217066	151870	-
- West	4576116	-	-	648972	543570	663897	775132	1944545	-
8	485326	-	-	63767	75442	87305	59953	198859	-
9	3664560	-	-	537911	430851	515020	643762	1537016	-
10	426230	-	-	47294	37277	61572	71417	208670	-
- South	4193971	-	15	701549	602281	618037	647021	1625068	-
11	2752743	-	15	488613	392395	363490	402925	1105305	-
12	1441228	-	-	212936	209886	254547	244096	519763	-
99	-	-	-	-	-	-	-	-	-

Figure 12: Quantitative table for turnover according to Region and Size class



	- Total	2	4	5	6	7	8	9	99
- Total	16848000	0	0	2712000	2320000	2506000	2800000	6510000	0
- North	4374000	0	0	720000	660000	690000	756000	1548000	0
1	1986000	0	0	398000	348000	356000	418000	466000	-
2	1810000	0	-	224000	222000	242000	258000	864000	0
3	578000	-	-	98000	90000	92000	80000	218000	-
- East	3704000	0	0	642000	514000	534000	622000	1392000	-
4	124000	0	-	36000	32000	26000	18000	12000	-
5	526000	-	-	94000	94000	110000	82000	146000	-
6	2236000	0	0	346000	252000	252000	304000	1082000	-
7	818000	-	-	166000	136000	146000	218000	152000	-
- West	4576000	-	-	648000	544000	664000	776000	1944000	-
8	486000	-	-	64000	76000	88000	60000	198000	-
9	3664000	-	-	538000	430000	514000	644000	1538000	-
10	426000	-	-	46000	38000	62000	72000	208000	-
- South	4194000	-	0	702000	602000	618000	646000	1626000	-
11	2752000	-	0	488000	392000	364000	402000	1106000	-
12	1442000	-	-	214000	210000	254000	244000	520000	-
99	-	-	-	-	-	-	-	-	-

Figure 13: Table from Figure 12, protectively additively rounded with a rounding base of 2000

3.7 Literature

Fischetti, M. and Salazar Gonzales, J.J. (2000), *Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints*, Journal of the American Statistical Association, vol. 95, pp. 916 – 928.

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 (Ed.), LNCS 2316, Springer-Verlag Berlin Heidelberg, pp. 181 – 192.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K. and De Wolf, P.P. (2012), *Statistical Disclosure Control (Wiley Series in Survey Methodology)*. John Wiley & Sons, ISBN 9781118348215.

Hundepool, A., Van de Wetering, A., Ramaswamy, R., Wolf, P.P. de, Giessing, S., Fischetti, M., Salazar, J.J., Castro, J. and Lowthian, P. (2007), *τ-ARGUS user manual 3.2*, Voorburg.

Loeve, A. (2001), Notes on sensitivity measures and protection levels, BPA 01892-01-S-TMO.

De Wolf, P.P. (2002), *HiTaS: A Heuristic Approach to Cell Suppression in Hierarchical Tables*, In: 'Inference Control in Statistical Databases' Domingo-Ferrer (Ed.), LNCS 2316, Springer-Verlag Berlin Heidelberg, pp. 74 – 82.

4. Statistical Disclosure Control of Frequency Tables

4.1 General description and reading guide

4.1.1 General description

The statistical disclosure control of frequency tables encompasses the production of frequency tables that satisfy the policy on statistical disclosure control and can be published as such. Frequency tables are tables in which the number of contributors per cell is given. This is in contrast to quantitative tables in which the cell values are created by summation of a continuous variable over all the contributors to a cell. Other rules apply to quantitative tables, and other protection methods may be more suitable than those for frequency tables. Disclosure control methods for quantitative tables are discussed in the previous section 'Statistical Disclosure Control of Quantitative Tables'.

Most statistics Acts require the protection of recognisable data about statistical units. A violation of the statistical confidentiality ('disclosure') boils down to the combination of two facts: the recognition of a unit and the disclosure of further detailed information about that unit.

For frequency tables, this can be formulated as follows. The user must first recognise a contributor or group of contributors in the table. This is followed by a statement about these contributor(s) due to the frequency distribution over the cells. The statement that the table makes possible about this group must provide more information about the members of the group than just the group size. In this sense, knowledge that is needed to recognise the members of the group is not considered information about the members of the group.

The statutory requirement is satisfied if the table does not provide any information about an individual statistical unit as such. The statistical professional standards and an NSI's own interest in the continuity of having enough response in their surveys and/or registers however, leads in certain cases to the requirement that the table does not provide information about groups of statistical units (people or households, etc.). In particular, that is the case if the table contains variables that could provide harmful or potentially damaging information about these groups. Such data will be referred to hereinafter as 'sensitive data'.

The methods described in this section can be easily applied using the software package τ-ARGUS. This package was developed in a European context.

4.1.2 Reading guide

As the first step in determining the correct statistical disclosure control for a frequency table, it will have to be determined whether disclosure is possible. In the first instance, the basis for this is ‘common sense’: is there information present in the table that may not be disclosed about individual respondents? For frequency tables, such information can be hidden in the spanning variables. Part of the spanning variables can be considered as identifying variables and the rest as sensitive. With respect to sensitivity, an additional distinction could be made with regard to the degree of sensitivity.

Once unsafe situations have been identified, the table will generally have to be protected. There are three general methods available for this purpose: restructuring the table (see section 4.4), suppression (see section 4.5) and rounding (see section 4.6).

Which method or combination of methods is ultimately used in a specific situation cannot be determined in advance. This depends significantly on the intended users. For example, in Eurostat regulations it is not always possible to restructure the table, and cell suppression or rounding will often have to be selected. When selecting the method or methods to be used, two competing aspects must be taken into account:

- disclosure risk;
- information loss.

In general, it can be said that reducing the disclosure risk will lead to increased information loss. The converse is also true: the smaller the information loss, the larger the disclosure risk.

4.2 Scope and relationship with other sections

This section discusses methods that are used for the statistical disclosure control of frequency tables. This chapter does *not* discuss any methods that are only used for quantitative tables. For these methods, see the section ‘Statistical Disclosure Control of Quantitative Tables’.

A few of the methods described in this chapter can, in principle, be used for both quantitative tables and frequency tables. Such methods will be mentioned more extensively in the section ‘Statistical Disclosure Control of Quantitative Tables’.

The methods in this chapter can be divided into two variants: methods for determining the primary unsafe cells of a frequency table and methods for making tables with unsafe cells suitable for publication.

4.3 Temporary standardisation of a frequency table

4.3.1 Brief description

The goal of statistical disclosure control is to prevent the disclosure of information about individual contributors to a table, or at least to make this more difficult. To achieve this, the cells will first have to be identified where there is risk of a possible disclosure. In frequency tables, at least two aspects play a role in this: recognisable groups and sensitive variables. Stated briefly, an unsafe situation occurs in a frequency table either if a cell that corresponds to a recognisable group of respondents contains too few respondents, or if the distribution of the respondents from a recognisable group is too concentrated in one or two categories.

To do this, it is useful to look at the frequency table in a standard format. In some cases, the frequency table will already be available in this format. In others, it will be necessarily to temporarily convert it. Once the frequency table has been protected, it can be restructured in its original format. Another possibility would be to convert the table only as a thought experiment to apply the rules to find unsafe cells.

4.3.2 Applicability

Before a frequency table can be statistically protected, it will first have to be indicated where possible problems arise in that table. For this purpose, it is convenient to temporarily look at the frequency table in a standard way, so that a clear distinction is visible between identifying and sensitive variables.

4.3.3 Detailed description

To detect unsafe situations in frequency tables, it is necessary to divide the spanning variables into identifying variables and sensitive variables. The qualification for each variable can, in principle, be determined by the department concerned. To promote coordination between the different frequency tables to be published, it is a good idea to keep track of this centrally.

Next, the frequency table can be restructured temporarily so that the identifying variables are included in the left-hand column and the categories of the sensitive variables are present in the other columns. This must also involve ‘hidden’ variables that define the population or subpopulation that is the subject of the frequency table.

4.3.4 Example

Suppose that Table 1 is a frequency table of the number of people who, in a certain year, have died from a non-natural death, as stated in a publication[4].

Type of non-natural death	Gender	Age						
		Total	<15	15-<20	20-<40	40-<60	60-<80	>=80
Suicide	Total	1530	8	43	418	674	298	89
	Man	1027	8	34	297	453	181	54
	Woman	503	-	9	121	**221**	117	35
Murder and manslaughter	Total	141	11	13	74	34	8	1
	Man	96	9	5	47	32	2	1
	Woman	45	2	8	27	2	6	-
Traffic accident	Total	880	47	120	380	67	179	87
	Man	636	23	87	315	52	98	61
	Woman	244	24	33	65	15	81	26
Workplace accident	Total	81	-	3	30	42	6	-
	Man	79	-	3	28	42	6	-
	Woman	2	-	-	2	-	-	-
Personal accident	Total	2013	64	6	120	60	481	1282
	Man	834	32	2	100	56	223	421
	Woman	1179	32	4	20	4	258	**861**
Other/unknown	Total	110	2	6	24	8	37	33
	Man	63	1	4	18	7	20	13
	Woman	47	1	2	6	1	17	20
Total	Total	4755	132	191	1046	885	1009	1492
	Man	2735	73	135	805	642	530	550
	Woman	2020	59	56	241	243	479	942

Table 1: *Number of people who died from a non-natural death in year J*

The standardised form of this frequency table for statistical disclosure control is created by placing the identifying variables in the left-hand column and the sensitive variables in the other columns. In Table 1, the identifying variables are ‘Gender’ and ‘Age’. The sensitive variable is the variable ‘Type of non-natural death’. The standardised version of Table 1 is presented in Table 2. The bold figures in Table 2 denote cells where the counts are ‘too concentrated’, e.g., more than 90% (221 out of 243) of the women between 40 and 60 that died a non-natural death, committed suicide. This is not as easy to deduce from Table 1.

[4] The figures in the table are fictitious.

		Type of non-natural death						
Gender	Age	Suicide	Murder and manslaughter	Traffic accident	Workplace accident	Personal accident	Other / Unknown	Total
Man	<15	8	9	23	-	32	1	73
	15-<20	34	5	87	3	2	4	135
	20-<40	297	47	315	28	100	18	805
	40-<60	453	32	52	42	56	7	642
	60-<80	181	2	98	6	223	20	530
	>=80	54	1	61	-	421	13	550
	Total	1027	96	636	79	834	63	2735
Woman	<15	-	2	24	-	32	1	59
	15-<20	9	8	33	-	4	2	56
	20-<40	121	27	65	2	20	6	241
	40-<60	**221**	2	15	-	4	1	243
	60-<80	117	6	81	-	258	17	479
	>=80	35	-	26	-	**861**	20	942
	Total	503	43	220	2	1147	46	1961
Total	<15	8	11	47	-	64	2	132
	15-<20	43	13	120	3	6	6	191
	20-<40	418	74	380	30	120	24	1046
	40-<60	674	34	67	42	60	8	885
	60-<80	298	8	179	6	481	37	1009
	>=80	89	1	87	-	1282	33	1492
	Total	1530	139	856	81	1981	109	4696

Table 2: *Standardised version of Table 1*

4.4 Table restructuring

4.4.1 Brief description

Section 4.3 describes how unsafe situations can be discovered in frequency tables by temporarily looking at the table in standardised form. If unsafe cells are subsequently found, the table will have to be protected before it can be published. An initial option to make a frequency table with unsafe cells suitable for publication is to restructure the table. By combining categories, the content per cell is increased. This affects the distribution of the sensitive spanning variable(s) among the various categories. This method can also be used to increase the content per recognisable group.

4.4.2 Applicability

This method will generally lead to fewer unsafe cells occurring in the table. By combining rows and/or columns, cells are combined and the content per cell is increased. The distribution of the sensitive spanning variable(s) among the various categories is also affected as a result.

There are no methodological conditions for using this method. However, externally imposed delivery obligations sometimes specify what detail level of a table must be published. This may be a Eurostat obligation, but Statistics Netherlands policy can also imply that a certain detail level of a table must be published. In these cases, the method can

be used from a technical perspective, but this is prevented by external or other policy decisions.

4.4.3 Detailed description

The standardised version of the frequency table is used to determine whether an unsafe situation is present. The restructuring can take place in two ways:

a. Restructuring the original table

b. Restructuring the standardised version of the table.

If option a is selected, the table will have to be examined again in the standardised form after restructuring to determine whether the disclosure control rules have been satisfied. The standardised form will also have to be used to determine which unsafe cells must be dealt with. In the case of option b, it is immediately clear which cells must be dealt with, but the restructuring will still have to be converted to the original table.

4.4.4 Example

Table 2 shows that the distribution of the respondents among the various categories of the sensitive variable does not satisfy the disclosure control rules in two rows. These rules require that a cell may not contain a concentration of nearly all the respondents. The cell (Woman, 80+, Personal accident) contains 91% of the total group of women aged 80+ who died a non-natural death, and the cell (Woman, 40-60, Suicide) also contains 91% of the total group of women between 40 and 60 years of age who died a non-natural death.

Based on the original table, the choice could be made to not split the causes of death ‘Suicide’ and ‘Personal accident’ in terms of gender.

Based on the standardised form, the choice could be made to condense the ages categories to ‘<15’, ‘15-<20’, ‘20-<60’ and ‘>=60’. This creates a table where there is no longer a strongly concentrated distribution of recognisable groups in a single cell. See Table 3 for the associated table in its original form.

Type of non-natural death	Gender	Age Total	<15	15-<20	20-<60	>=60
Suicide	Total	1530	8	43	1092	387
	Man	1027	8	34	750	235
	Woman	503	-	9	342	152
Murder and manslaughter	Total	141	11	13	108	9
	Man	96	9	5	79	3
	Woman	45	2	8	29	6
Traffic accident	Total	880	47	120	447	266
	Man	636	23	87	367	159
	Woman	244	24	33	80	107
Workplace accident	Total	81	-	3	72	6
	Man	79	-	3	70	6
	Woman	2	-	-	2	-
Personal accident	Total	2013	64	6	180	1763
	Man	834	32	2	156	644
	Woman	1179	32	4	24	1119
Other/unknown	Total	110	2	6	32	70
	Man	63	1	4	25	33
	Woman	47	1	2	7	37
Total	Total	4755	132	191	1931	2501
	Man	2735	73	135	1447	1080
	Woman	2020	59	56	484	1421

Table 3: Protected version of Table 1

4.5 Suppression

4.5.1 Brief description

A frequently used method to protect primary unsafe cells is to suppress (not publish) certain cells. The cell value is then simply replaced by an '×'.

In a frequency table where the totals and subtotals are also provided, however, it is often not sufficient to suppress only the primary unsafe cells. After all, if a suppressed cell is the only suppressed cell in a row, the suppressed value is simple to calculate by subtracting the other cell values in that row from the associated marginal. The same possibilities are available as with suppression in quantitative tables (see section 3.5).

4.5.2 Applicability

Unsafe situations in frequency tables can be divided into two cases:

a. The recognisable group is too small;

b. The distribution of the recognisable group among the sensitive variable(s) is too concentrated in a single sensitive cell.

To determine a suitable suppression pattern, it is necessary to know how one can comply with the disclosure control rules imposed. In many algorithms, so-called safety intervals are used for this purpose. These are the minimum intervals for primary suppressed cells that should arise from the suppression pattern. At present, contrary to the case of

quantitative tables, no method is available to calculate the minimum intervals for primary unsafe cells in frequency tables. The method as described in Fischetti and Salazar (2000) is therefore not directly applicable as yet.

4.5.3 Detailed description

If a row total in the standardised form of the table is too small (the recognisable group is too small), this cell will have to be suppressed. Of course, multiple cells will have to be suppressed to prevent the row total from being calculated. In general, this will mean that the total row will have to be suppressed, including a second possibly 'safe' row.

A second situation that may arise is a sufficiently large row total which, however, is too concentrated in a single sensitive category of the variable. In that case, the row total is, in principle, suitable for publishing. The cell associated with the category of the sensitive variable in which the respondents are concentrated can then be viewed as the primary cell to be suppressed. In a table with totals and subtotals, one must also look for secondary cells to be suppressed. In many algorithms, safety intervals are used for this purpose. These are the minimum intervals for primary suppressed cells that should follow from the suppression pattern. At present, contrary to the case for quantitative tables, no method is available to calculate minimum intervals for primary unsafe cells in frequency tables. The method as described in Fischetti and Salazar (2000) is therefore also not directly applicable as yet.

An additional problem is formed by what is called 'meaningful aggregates' in the disclosure control rules. If multiple cells in a row are suppressed, the total of these suppressed cells is actually published. If the suppressed cells form a meaningful aggregate, then the respondents may also not be too concentrated in that combined cell. Account should therefore be taken of this when determining secondary suppressions. It is not yet clear if the Fischetti and Salazar model (2000) is general enough to take this into account.

4.5.4 Example

Table 4 shows a suppression pattern in which it is assumed that the aggregate 'Suicide' + 'Personal accident' is not a 'meaningful aggregate'. Both problematic cells are suppressed by placing '×'-s in the cells.

Type of non-natural death	Gender	Age						
		Total	<15	15-<20	20-<40	40-<60	60-<80	>=80
Suicide	Total	1530	8	43	418	674	298	89
	Man	1027	8	34	297	×	181	×
	Woman	503	-	9	121	×	117	×
Murder and manslaughter	Total	141	11	13	74	34	8	1
	Man	96	9	5	47	32	2	1
	Woman	45	2	8	27	2	6	-
Traffic accident	Total	880	47	120	380	67	179	87
	Man	636	23	87	315	52	98	61
	Woman	244	24	33	65	15	81	26
Workplace accident	Total	81	-	3	30	42	6	-
	Man	79	-	3	28	42	6	-
	Woman	2	-	-	2	-	-	-
Personal accident	Total	2013	64	6	120	60	481	1282
	Man	834	32	2	100	×	223	×
	Woman	1179	32	4	20	×	258	×
Other/unknown	Total	110	2	6	24	8	37	33
	Man	63	1	4	18	7	20	13
	Woman	47	1	2	6	1	17	20
Total	Total	4755	132	191	1046	885	1009	1492
	Man	2735	73	135	805	642	530	550
	Woman	2020	59	56	241	243	479	942

Table 4: Suppression pattern for the protection of Table 1

4.6 Additive rounding

4.6.1 Brief description

In frequency tables, rounding is a rather natural method of disclosure control. First of all, the exact cell values are only known in a certain interval when rounding is used. The extent to which rounding is performed will, of course, have an impact on the size of the intervals. Second, an unrounded frequency table creates the impression of great precision: the counting has been performed down to the individual units. In the case of estimated frequencies, this is false precision. Rounding also cuts down on this false precision.

If each cell is rounded independently, the additivity of the table will not necessarily be maintained. Of course, there is a simple way to guarantee the additivity: by rounding the cells in the interior of the table independently of one other and then recalculating the marginals. As a result, however, the marginals can deviate significantly from the rounded or unrounded original values.

In additive rounding, the table is rounded such that the additivity is maintained and that the rounded table deviates from the original as little as possible. The size of the rounding base determines the extent to which the frequency table is protected: the larger the rounding base, the greater the protection will generally be. At present, no method is available to automatically determine the correct rounding base.

4.6.2 Applicability

Additive rounding can be used for the statistical disclosure control of both quantitative tables and frequency tables. Often, a presentation argument will also play a role: a large number of significant figures suggests a high degree of precision that is not always justified because of sampling errors and measurement errors. Rounding the table values reduces this false precision to a certain extent.

4.6.3 Detailed description

See section 3.6.

4.6.4 Example

Table 5 shows a rounded version of Table 1, where additive rounding is performed using rounding base 50.

Type non-natural death	Gender	Age Total	<15	15-<20	20-<40	40-<60	60-<80	>=80
Suicide	Total	1550	0	50	400	700	300	100
	Man	1050	0	50	300	450	200	50
	Woman	500	-	0	100	250	100	50
Murder and manslaughter	Total	150	0	0	100	50	0	0
	Man	100	0	0	50	50	0	0
	Woman	50	0	0	50	0	0	-
Traffic accident	Total	850	50	150	350	50	200	50
	Man	600	0	100	300	50	100	50
	Woman	250	50	50	50	0	100	0
Workplace accident	Total	100	-	0	50	50	0	-
	Man	100	-	0	50	50	0	-
	Woman	0	-	-	0	-	-	-
Personal accident	Total	2000	50	0	150	50	450	1300
	Man	850	50	0	100	50	200	450
	Woman	1150	0	0	50	0	250	850
Other/unknown	Total	100	0	0	0	0	50	50
	Man	50	0	0	0	0	50	0
	Woman	50	0	0	0	0	0	50
Total	Total	4750	100	200	1050	900	1000	1500
	Man	2750	50	150	800	650	550	550
	Woman	2000	50	50	250	250	450	950

Table 5: Rounded version of Table 1, with rounding base 50. An '0' is a rounded 0, a '-' is an empty cell

4.7 Literature

Fischetti, M. and Salazar Gonzales, J.J. (2000), *Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints,* Journal of the American Statistical Association, vol. 95, pp. 916 – 928.

Hundepool, A., Van de Wetering, A., Ramaswamy, R., De Wolf, P.P., Giessing, S., Fischetti, M., Salazar, J.J., Castro, J. and Lowthian, P. (2007), *τ-ARGUS user manual 3.2*, Voorburg.

5. Statistical Disclosure Control of Analysis Results

5.1 General description and reading guide

5.1.1 General description

In addition to problems and methods for protecting microdata, quantitative tables and frequency tables described in the previous paragraphs, there is still a very broad, diverse group of statistical output. This concerns the results of various types of statistical analyses and model estimations. In principle, these results also run a risk of disclosing the data for individual respondents and must therefore be treated with care. The risk of disclosure is present particularly in the case of outliers. When determining whether these results are sufficiently safe, the underlying frequency tables are often examined. There is often a strong correlation between the analysis model and an underlying frequency table.

5.1.2 Reading guide

The problem of determining whether the results of statistical analyses are sufficiently safe arises mainly when checking the output of OnSite working and Remote Access. This is where many statistical analyses are performed on unprotected data (Secure Use Files), while the users would very much like to use and publish the results of their research outside of the safe environment provided by the NSI. Checking the output is a necessary part of this valued service. With respect to checking the output, it does not matter at all whether the output is obtained through OnSite or Remote Access. In both cases, the same analyses are performed on the same data files using the same tools (SPSS, SAS etc.).

Because this problem does not only occur at a particular individual NSI, but actually at every statistical bureau in Europe, it was decided to make this a subject of the Statistical Disclosure Control ESSnet project (2008-2009). The ESSnet was partly subsidised by Eurostat. One of the tasks in the ESSnet project was to draw up guidelines for checking output. For this subject, use is also made of these ‘Guidelines for Output Checking’, which can be found on the ESSnet website:
(http://neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf).

The subject is still under development. In another European co-funded project (Data Without Boundaries, see http://www.dwbproject.org) these guidelines were updated and amended with examples. See at the bottom of the page http://www.dwbproject.org/access/guides.html (Bond et al., 2013).

Due to the diversity of the problem, both in terms of the number of possible analysis methods and the number of different statistical packages, each with their own forms of output, it is not possible to develop ready-made software for this purpose.

5.2 Scope and relationship with other sections

This section discusses methods that can be used to determine whether the results of statistical analyses are sufficiently safe. To a large extent, use is made of the results of a European project group that drew up these guidelines. These guidelines also discuss tables. However, because these subjects have already been covered in the previous chapters, these subjects from the guidelines are less relevant here.

5.3 Disclosure control of analysis results

The methods for the protection of analysis results tie in with these European guidelines.

A number of considerations played a role when drawing up the guidelines for output checking. Of course, it is not possible to fully discuss all possible forms of output. The number of different methods available in SAS and SPSS is so large that it is impossible to assess all of these methods with respect to their possible disclosure risks. Just consider the size of the SPSS and SAS documentation.

Another aspect that plays an important role in the guidelines is their feasibility in practice. In assessing output, we must take account of two possible errors: first, incorrectly approving unsafe results and, second, incorrectly holding back safe results.

In the guidelines, two methods are provided for each subject. A ‘Rule of Thumb’, which primarily minimises the first error, and a ‘principles-based’ approach that tries to minimise both errors.

The idea behind this distinction is that much of the research output can easily be handled by the simple rule. If the output is not allowed due to the ‘Rule of Thumb’, and the researcher wants it approved anyway, extra work must be performed (also by the researcher) to demonstrate that the results are indeed safe.

Here is a list of the types of output that are currently discussed in the guidelines:

Descriptive statistics	Frequency tables
	Magnitude tables
	Maxima, minima and percentiles (incl. median)
	Mode
	Means, indices, ratios, indicators
	Concentration ratios
	Higher moments of distributions (incl. variance, covariance, kurtosis, skewness)
	Graphs: pictorial representations of actual data

Correlation and Regression Analysis	Linear regression coefficients
	Non-linear regression coefficients
	Estimation residuals
	Summary and test statistics from estimates (R^2, χ^2 etc.)
	Correlation coefficients
	Factor analysis
	Correspondence analysis

For the rest, please refer to the European Guidelines.

5.4 Literature

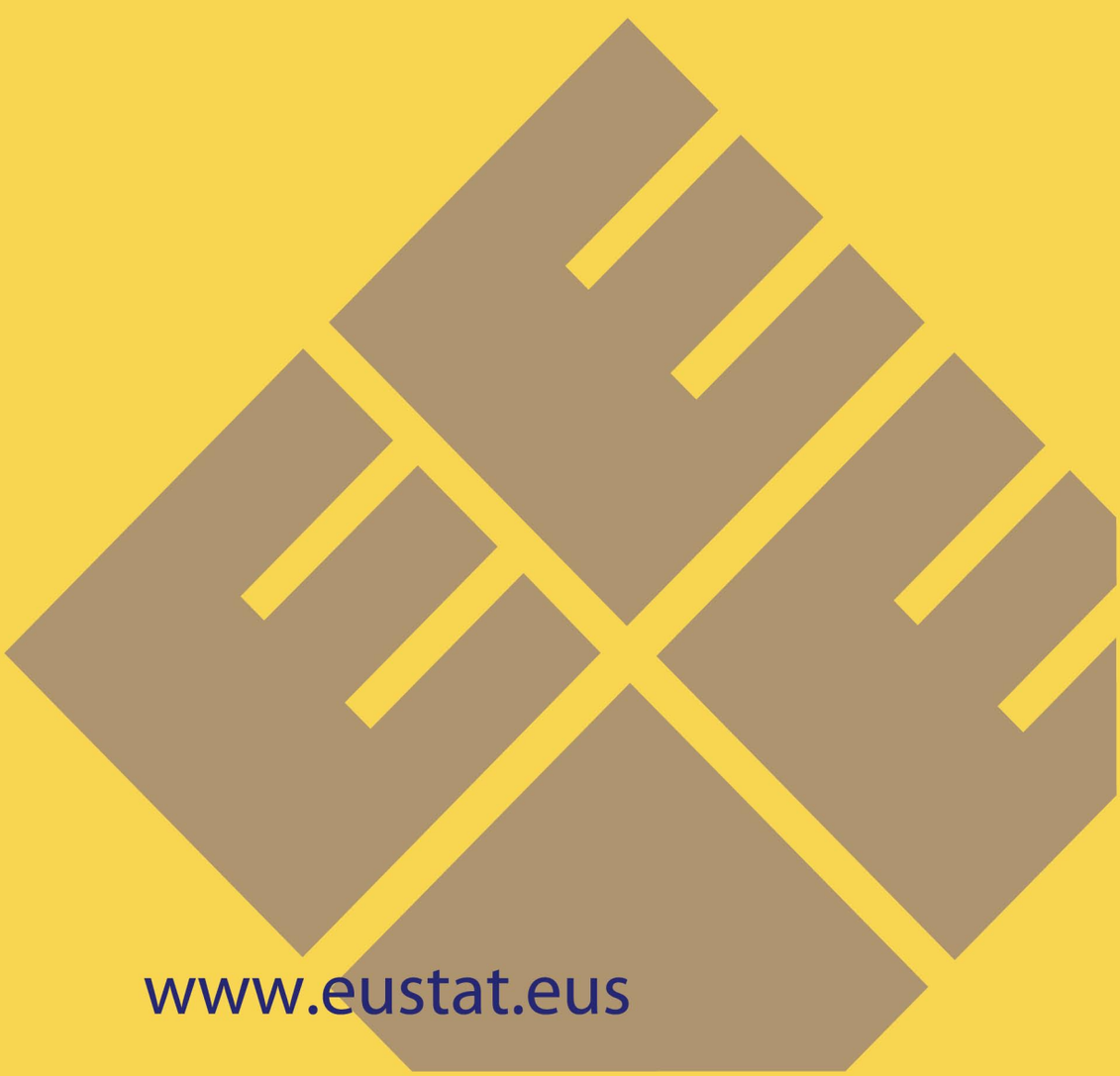
Ritchie, F., Welpton, R., Franconi, L., Lucarelli, M., Seri, G., Brandt, M., Guerke, C., Hundepool, A.J. and Mol, J. (2010), *Guidelines for the checking of output based on microdata research*, ESSnet-SDC-project. http://neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf

Bond, S., Brandt, M. and De Wolf, P.P. (2013), *Guidelines for Output Checking*, Data without Boundaries project. http://www.dwbprojet.org/export/sites/default/access/doc/dwb_standalone-document_output-checking-guidelines.pdf

Erakunde autonomiaduna
Organismo Autónomo del

EUSKO JAURLARITZA
GOBIERNO VASCO

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA



www.eustat.eus