Scenarios of attack: the data intruder's perspective on statistical disclosure risk

Article · January 1999

CITATIONS
56

READS
172

2 authors:

Mark James Elliot
The University of Manchester
96 PUBLICATIONS **1,110** CITATIONS
SEE PROFILE

Angela Dale
The University of Manchester
8 PUBLICATIONS **362** CITATIONS
SEE PROFILE

Netherlands Official Statistics

Volume 14, Spring 1999

Special issue

Statistical disclosure control

Voorburg

Prinses Beatrixlaan 428
P.O. Box 4000
2270 JM Voorburg (Netherlands)

Telephone : . .31 (070) 337 38 00
Fax : . .31 (070) 387 74 29
E-mail: lhka@cbs.nl
Internet: http://www.cbs.nl

Heerlen

Kloosterweg 1
P.O. Box 4481
6401 CZ Heerlen (Netherlands)

Telephone : . .31 (045) 570 60 00
Fax : . .31 (045) 572 74 40

Key figure A-125/1999

Subscription: Dfl. 42.00 per year
Price per copy: Dfl. 20.00

ISSN 0920-2048

Postage will be charged.

Contents

Editors
Jeroen Pannekoek
Leon Willenborg

Coordinating editor
Lieneke Hoeksma

Preface

Jeroen Pannekoek and Leon Willenborg

National statistical offices make information available to the public, traditionally in the form of aggregated (tabular) data but more recently also in the form of data sets containing individual records (microdata) that allow users to construct their own aggregates and do other analyses. It is in the interest of the users to make the information as detailed as possible but this interest conflicts with the obligation that statistical offices have to protect the confidentiality of the information provided by the responding units (establishments or individuals). Thus, statistical offices are confronted with the problem of ensuring that the risk of a breach of confidentiality (a disclosure) is acceptably low, while at the same time preserving as much as possible the information content of the data to be released [1)]. In response to this problem statistical offices try to assess the level of the risk of disclosure and where necessary take measures to diminish this risk. In addition to legal and organisational measures (contracts, modes of access, segments of users with their privileges and duties, etc), these measures include statistical disclosure control. These latter measures are intended to modify the data in such a way that it is more difficult to recognise individuals or that information about such an individual is less detailed or (slightly) perturbed.

This special issue of *Netherlands Official Statistics* is devoted to the problem of disseminating information while preserving confidentiality. It concentrates on the practices at statistical offices in dealing with this problem and the motivations for these practices. This involves topics such as evaluation of the disclosure risk, legal and organisational aspects, methods for reducing the disclosure risk and implementation of these methods in software, especially μ- and τ-ARGUS. The ARGUS software was developed by researchers from Statistics Netherlands, Eindhoven University of Technology (the Netherlands), the University of Padova (Italy) and the University of La Laguna in Tenerife (Spain), collaborating in the Statistical Disclosure Control (SDC) project, under sponsorship of the European Union through Esprit [2)].

Before we give a short summary of the contents of this issue we will sketch a background to statistical disclosure control methodology and practice. In particular we shall try to point out the crucial role of a so-called *disclosure scenario* in this area. A disclosure scenario is in fact a model for the actions and knowledge that a hypothetical intruder is assumed to apply when attacking a data set. Once such a model is assumed one can try to develop probabilities that each of the separate steps in a disclosure risk scenario are carried out successfully by an intruder. In this way one arrives at a formal disclosure risk model, in which it is possible to assign a probability to the event that an intruder correctly identifies an individual represented in the data. In practice, it is, however, by no means always necessary to develop a formal probability model. Instead one can try to work with informal, and simplified, disclosure risk models, which are based on qualitative information or in which certain parameters are postulated without a probabilistic reasoning to support them. Such models are not necessarily inferior in practice, only more suited for practical purposes. For instance, it is generally impossible to quantify the probability that an intruder will make an attempt at disclosure, or that the intruder has a database at his disposal to match the statistical data against.

An example of a simplified disclosure risk model is one that Statistics Netherlands uses to protect its microdata for researchers.. The idea is that an intruder should not be able to make a disclosure on the basis of spontaneous recognition, which is made operational in terms of certain low-dimensional combinations of scores on identifying variables that have to be checked. The requirement is that each such combination that exists in a microdata set appears frequently enough, i.e. more frequently than a certain threshold value which the data protector has to specify. This kind of approach is also the basis of μ-ARGUS.

Another important consequence of the availability of a disclosure scenario is that this provides a rational basis to counteract – or at least hamper – disclosure attempts of an intruder. It therefore guides the data protector in his goal to modify the original data so that a data set is produced that can be released. In fact, a formal disclosure risk model can be used to even quantify the reduction of disclosure risk (given the adopted disclosure scenario) as a result of the application of suitable data modification techniques [3)].

It should be borne in mind that a disclosure risk model is only one of the ingredients that a data protector should use to produce a set of sufficiently protected data, suitable for external release. Another ingredient is information loss. Intuitively the aim of protecting a set of data (tables or microdata), given criteria for the safety, is to modify the data in such a way that the resulting data are safe (according to the criteria adopted) and the information loss is minimised. There are various ways in which information loss can be quantified: one can use entropy measures, mean square errors or (more or less subjectively chosen) weights. The actual choice of a suitable information loss function is not always easy in practice: it requires that the data protector should be well informed about the actual use of the data. In case of user groups with different research interests there may be a conflict that is impossible to settle by releasing a single protected data set [4)]. But where interests do not conflict, the data protector should be able to arrive at an appropriate information loss function. In fact he may use it implicitly when protecting a data set. However, if at least part of this data modification process is carried out automatically (say, using one of the ARGUS packages), the data protector needs to specify such information loss explicitly. For instance, one can use τ-ARGUS to carry out automatic (secondary) cell suppression or rounding, and μ-ARGUS to find optimum combinations of global recodings and local suppressions. Solving these problems requires non-trivial (mixed integer) linear programming problems to be solved.

To make things more concrete, let us assume that an anonymised microdata file is released (i.e. direct identifiers like names and addresses have been discarded). An often used scenario for this case is a matching scenario: a re-identification of a record, and hence a disclosure of the identity of the responding unit, occurs because a record from the released data file can be matched with a record from an external database that contains a direct identifier. In this scenario the following sequence of events is supposed to take place:

– Someone (the intruder) attempts to identify an individual in the released data set.
– This intruder has an external database available with a number of identifying variables (or key-variables) that are also part of the released data file.
– At least one of the records of the external database can be matched with a record of the released data set, which implies that the combination of values of the key-variables is unique in both data sets.
– The match is a correct match. This is certain if either of the data sets covers the whole population but not so otherwise.

As will be apparent from the articles and the references in this issue, this simplified scenario can be extended and otherwise modified in a number of ways. To mention just one modification: the matching information (values of key-variables) need not be obtained from an external database, for certain individuals such as friends, colleagues, fellow members of professional and social organisations, it may be part of the personal knowledge of the intruder. The 'matching' that is carried out is not done on a computer but in the head of an intruder. It is clear that this sort of matching is

different from computer matching, being on the one hand more flexible (a human being can improvise better), but on the other more restricted (a computer is quicker in processing a large amount of data).

All the articles in this issue relate to this disclosure scenario, either by trying to assess qualitatively or quantitatively the level of the disclosure risk induced by some of the events described above, or by describing methods to diminish the disclosure risk, in terms of the various events that a disclosure scenario distinguishes. We now turn to the articles in this issue.

The article by **Elliot and Dale** concentrates on the probability of an attempt to identify an individual being made and describes different disclosure scenarios performed by different types of intruders with different motives and means for there disclosure attempts. Thus they give an inventory of the different kinds of 'attack' a statistical office has to be prepared for. In addition they give a critical evaluation of the key variables that should be the target of disclosure protection measures, thereby taking measurement error into account.

There are two contributions from **De Waal and Willenborg**. The first considers the problem of whether an intruder using linear programming techniques is able to recalculate exactly the value of cells in a super-table from a set of published linked tables, which are viewed as marginal tables of this super-table. The article shows how this problem relates to a similar and more standard problem for a single table with some suppressed cells. Their second contribution considers ways to quantify the information loss in a microdata set that has been modified by global recodings and local suppressions. The relevance of such measures for automatic global recoding and local suppression is also discussed.

The articles by **Kooiman, Nobel and Willenborg**, by **Zayatz, Massell and Steel**, and by **Thorogood** describe statistical disclosure limitation practices and research at Statistics Netherlands, the United States Census Bureau and Eurostat, respectively. As detailed in these three contributions, for the statistical offices concerned the protection of confidentiality is a legal requirement. To meet this requirement, these offices aim at both reducing the probability of an attempt and the probability of an identification given an attempt. An example of a measure to reduce the probability of an attempt is to restrict access to certain microdata files to named researchers who have signed a contract explicitly forbidding attempts to identify individual units. An even more restricted form of access is to not distribute microdata files to the users, but allow access only via on-site facilities, i.e. at the premises of the statistical office. In order to diminish the possibility of an identification measures are taken to reduce the amount of detail in the key-variables thereby reducing the possibility of a correct match with an external database. Examples of such procedures are adding noise and microaggregation (a form of rounding) for continuous variables and collapsing of categories and setting values to 'unknown' (suppression) for categorical variables. All such procedures can be performed by ARGUS. In practice, restricting access and reducing the amount of detail are complementary: data sets with very detailed information may be accessible only via on-site facilities whereas data sets with very little detail may be made available for the general public.

The article by **Schulte Nordholt** discusses in detail the disclosure limitation procedures for various outputs, a number of tables as well as a microdata file, of the Netherlands Annual Survey on Employment and Earnings. This application shows the kind of operational criteria that can be applied to consider a data set to be 'safe" for external release. For tables a dominance rule is applied of the form: a cell is unsafe for publication if the *n* major contributors to that cell are responsible for at least *p* percent of the total cell value. For the microdata file the number of population units for each key value, or combination of key values, must exceed a specified threshold. Disclosure limitation techniques were suppression of values and collapsing of categories for both the tables and the microdata and these techniques were performed by both ARGUS packages. The pros and cons of cell suppression versus collapsing of categories are discussed as well as some problems and strategies for dealing with 'linked" tables, i.e. a set of tables produced from the same microdata file and possibly with common marginals[5]. Disclosure limitation for the microdata file aims at reducing the amount of detail in the (categorical) key variables by global recoding and local suppression.

This special issue of *Netherlands Official Statistics* is devoted to current disclosure limitation practices as well as research into the assessment of the disclosure risk. We finish our introduction here and shall let the respective authors speak for themselves. But before we do, we would like to express our thanks to Mark Elliot, Angela Dale, Ton de Waal, Peter Kooiman, Joris Nobel, Laura Zayatz, Paul Massell, Phil Steel, David Thorogood and Eric Schulte Nordholt for their willingness to submit a contribution to this issue, and for the pleasant cooperation we, as editors, have experienced working with them.

Notes

1) Zero disclosure risk is often not a realistic requirement if the data are to be of any statistical value. Possible residual disclosure risks can sometimes be covered by appropriate legal and organisational measures (see below).

2) The consortium for the SDC project also included researchers from ISTAT and the University of Rome in Italy, the Universities of Leeds, Manchester and Southampton, and the ONS in the UK.

3) By suitable we mean that if the disclosure risk model adopted does not assume the presence of measurement errors then one cannot apply perturbative techniques such as adding noise, data swapping, PRAM, etc., which all introduce artificial measurement errors into the data. In this case one should use non-perturbative techniques such as global recoding or local suppression. The application of perturbative techniques requires a disclosure risk model that takes measurement error explicitly into account.

4) This situation arises when the safety criteria employed imply that two identifying variables cannot be given in great detail at the same time, so that a trade-off in detail is necessary. For instance it would be possible that region and profession cannot be both specified in great detail. The data protector is then forced to choose between a file with detailed region and a less detailed profession variable, or vice versa. A practical solution to such a dilemma – as manifested in e.g. the Dutch Labour Force Survey - is to produce a file with detailed region or detailed profession in alternate years. In fact this idea can be extended to arrive at topical files, which contain as much information as possible on a particular topic (e.g. minorities, etc.) to the detriment of detail in other variables. This idea is actually being taken up for the Labour Force Survey just mentioned.

5) This implies that the tables in this set cannot be protected individually, but should be protected collectively.

Scenarios of attack: the data intruder's perspective on statistical disclosure risk

Mark Elliot and Angela Dale [1)]

1. Introduction

The release of valuable microdata files from sources such as the census of population is constrained by fears of individual identification and hence disclosure of information. Much work has been done on analysing the risks of disclosure once an attempt has been made. However, the complexity of the social, psychological, and political factors that affect the probability of an attempt being made in the first place has meant that little work has been done on the overall likelihood of a disclosure attempt. Such an analysis is a crucial first step in a project to investigate the risk of statistical disclosure, because the psychological components of an attempt (e.g. goals and motives) will influence the type of attack made, the strategy used, and the probability of identification or disclosure given an attempt.

Most literature on statistical disclosure control starts with the assumption that an intruder has attempted to disclose information in a confidential data file. Analyses of the likelihood of that attempt being successful (e.g. Paass, 1988; Bethlehem et al., 1990; Lambert, 1993) provide a basis for proposals for disclosure avoidance techniques (e.g. Dalenius, 1986; Blien et al., 1990; Fuller 1993).

Marsh et al. (1991) argued for the release of samples of anonymised records (SARs) from the British 1991 Census on the basis of the estimated probability of a specified individual in the SAR being correctly matched to records in an external database, and of that match being verified. They classified the risk of a breach in confidentiality as the total probability of an individual being identified from the SAR. Formally this was defined as:

pr(identification) = pr(identification|attempt) pr(attempt)

They acknowledged the difficulty of assessing the value of pr(attempt) and, assuming that an attempt might be made, concentrated their efforts on a theoretical analysis of pr(identification|attempt), showing that it was negligible. However, they assumed the probability of an attempt to be low because the chances of identification, given an attempt, would be low. That is, the low probability of success would dissuade anybody from attempting it in the first place. However, the value of pr(identification|attempt) will be dependent on what method of identification is actually attempted - and this, itself, depends on the motivations of those making the attempt.

Marsh et al.'s analysis is also based on a disclosure attempt made by a large database holder wishing to enhance further that database. To draw general conclusions from this analysis regarding the risk to confidentiality presupposes: (i) that the matching procedure analysed is the optimum strategy for breaking confidentiality and (ii) that holders of large databases are the main threat to confidentiality. These two points are related and as Elliot (1996) demonstrates there are other possible uses for correctly matched target records besides the enhancement of databases. Some of these other uses indicate a different approach to attempts at identification and produce different values for pr(identification|attempt).

Marsh et. al. (1991) identify two main sources of identification attempts:

– holders of databases, such as credit reference agencies, which wish to update their information on individuals, by using census data.
– journalists or computer hackers who wish to identify people solely to discredit the census.

These two sources represent a dichotomy between attempts where the disclosure of information is the primary goal and attempts where the consequential effects (such as decline in public confidence) is the primary concern.

However, as Elliot(1996) argues this list should be extended to include political parties and other political organisations, government departments or law enforcement agencies, and individuals attempting to steal another's identity. [2)]

The dichotomy between types of attempt highlights a further important point. In order to establish what a potential data intruder would do, one needs to breakdown each attempt into its psychological and pragmatic components. Centrally, one needs to understand the motivations of the data intruder, without which it is impossible to clarify the nature of the attack.

Clearly then, a system of categorising attempts is needed. Section two of this paper describes such a scheme (which has been developed by Elliot (1996) to analyse possible attacks on the GB Sample of anonymised records SARs). Section 3 focuses on the particular issue of key variable definition, which can be viewed as the outcome of a scenario analysis.

2. Classification scheme for attacks

To analyse the probability of an attempt one needs to identify the pragmatic components of the attempt - inputs to, and processes of, the decision making process that leads to the attempt being made. The classification scheme presented here consists of 11 components:

– Motivation
– Means
– Opportunity
– Attack type
– Key/matching variables
– Target variables
– Effect of data divergence
– Likelihood of success
– Goals achievable by other means?
– Consequences of attempt
– Likelihood of attempt

This scheme is not definitive, some of the descriptive categories could be split into two or more parts, others could be joined .The decision when clarifying the scheme was to optimise the level of refinement, without the scheme becoming too unwieldy and difficult to grasp.

Motivation

The motivation for an attempt comprises two elements:

Rationale: A strategic level description of the main motivations of the data intruder (for example, to discredit the census).

Goals: A specification of the state of the world that the data intruder wishes to achieve; that is, an operational definition of what would be achieved by the disclosure attempt (for example, the release of a match into the public domain). It is important to note that it is this state of the world that the data intruder pursues, not disclosure of confidential data *per se*. The latter is simply instrumental in achieving the former.

Means

There are three elements in the means by which an attack would be made.
Available skills: Statistical skills are needed to identify and apply the optimum matching technique and to interpret the results. Computing skills are also needed, for example in use of appropriate analysis packages.
Available knowledge: Any factual knowledge that assists the data intruder in their attempt. As well as information stored in databases, available knowledge might include local knowledge (i.e. information about a locality such as the prevailing housing type) and direct knowledge (i.e. by visiting an address one can establish the housing type, by talking to an individual one can establish their occupation).
One particular type of knowledge is *response knowledge*; if an intruder knows that a target individual is in the target database then his/her chances of success are greatly enhanced. In this situation we say that the intruder has response knowledge of the target. This is particularly important for microdata produced from sample surveys.
Necessary computational power: Enough computing power to perform the analyses required to achieve the goals of the data intruder.

Opportunity

In order to have the opportunity to disclose information from a database, the intruder must have access to that database. Typically microdata sets are distributed only to those who have signed the necessary licenses, which are legally binding and which prohibit the licensee from attempting to identify an individual or household in the data files and from passing the data to an unlicensed individual.
Therefore, access to the target data sets could come through several routes:
– through an authorised data set user.
– in collusion with an authorised data set user, either voluntarily (because the colluder shares the data intruder's goals or has some other goal/reward, for example payment!) or involuntarily (if the colluder is threatened or blackmailed).
– a breach of a computing system containing a copy of the data set, either by the person/group making the attempt or by a hired hacker.
– a copy of the data set is stolen.

Because of the legal restrictions on the use of the data sets, these opportunities usually involve illegal activity. Therefore, in pursuit of their goals the data intruders must either accept the consequences of their illegal activity, or not release information into the public domain, or release information in a way that conceals the data intruder's identity.
It is virtually impossible to calculate accurately the dependent probability of such access to any given target microdata set. Clearly this will depend on the number of individuals having legitimate access. However, given that public use data sets typically have a large number of users, for many types of attempt it is probably prudent to assume that pr(opportunity)~1. That is, if any individual or organisation has decided to use a given data set for an attack, then they will be able to gain access to it.

Types of attack

An attack type is a general method for achieving a class of goals. That is, the optimum method of attack is determined almost exclusively by the goals of the data intruder. Five different types of attack are outlined below.

1. *Database cross match*. An outside database with several fields that are identical to or recodable to target record fields (key matching variables) is cross-matched with the target data set. The most likely goal for this type of attack would be the enhancement of an outside database.
2. *Match for a single specific individual.* The intention behind this type of attack is to enhance or verify information available about a target individual. Matching information could (although need not) be from an outside database - as a hypothetical example, the Inland Revenue may want to search the target data set for income related information of an individual suspected of tax evasion.
3. *Match for an arbitrary individual.* This type of attack concerns attempts where the data intruder is not interested in information gathering but in the (political) consequences of claiming that identification has been achieved and information could be disclosed. The data intruder is not interested in the actual identity of the individual who has been identified. A journalist in search of a good story might choose this route.
4. *Specific group of individuals.* This type of attack is an alternative to 3 for certain scenarios and would have the same goal. A set of individuals is selected either because they are distinctive (e.g. they may come from a minority ethnic group) or because a great deal of matching information is available on them (e.g. they belong to an occupation with a register of all members).
5. *Reversed matching.* Strictly, this is not a separate type of attack but a variation on any of the other types. Rather than starting with an individual or set of individuals in the outside world and attempting to identify them in the target data set, the data intruder starts with the target data set and locates one or more individuals with distinctive characteristics and attempts to find them in the world.

Key/matching variables

For all disclosure attempts key variables form an essential step to achieving identification. Key variables are those which are available to the data intruder and which are also available in the target data set and can therefore allow individuals to be matched.
Ideally, from the intruders viewpoint the coding method of a key variable would be the same on both target and identification data sets. However, it is often the case (particularly for the more discriminating variables) that coding of a variable will vary. This makes the matching of these variables more complex. Sometimes, as Elliot and Dale(1998) demonstrate, it is possible to construct additional variables which bridge between definitions on the target and identification data sets, possibly using several variables from each. In section 3 the paper examines this issue of *key variable quality* in more detail.

Target variables

Target variables are those whose values the data intruder wishes to disclose. Target variables are mostly relevant for attempts at disclosure where the motivation of the data intruder is to gain information. In scenarios where the data intruder is motivated by the secondary consequences of an attack, identification may be sufficient and therefore the information content of target variables is of little importance. Given this, target variables have two central properties *usefulness* and *sensitivity.*
Usefulness: For a variable to be a target it must contain information which improves upon or verifies data already available to the data intruder.

Sensitivity: The personal sensitivity of a target variable concerns the importance of the information disclosed for the individual concerned. General sensitivity can be seen as the prevailing perception of the personal sensitivity of the disclosed information. For example, the number of cars one possesses might be regarded as of low general sensitivity, but for an individual who is trying to keep his or her wealth hidden from the Inland Revenue it may be of high sensitivity.

Effect of data divergence

All databases contain errors and inaccuracies. These may be errors in information supplied by respondents, errors in recording by interviewers or transcribing by coders. Sometimes items of information are missing and, in the case of the British census, missing or inconsistent values are imputed using 'hot deck' methods (Mills and Teague, 1991). Data available from censuses and surveys are likely to be several years old by the time they are available for analysis outside the Office for National Statistics (ONS). This means that individual and household characteristics will have changed since the date of data collection and historical data may not be available.
Collectively, these sources of 'noise' in the data are termed *data divergence*. The term refers to two situation types:
– *data-data divergence* - differences between data sets, and
– *data-world divergence* - differences between data sets and the world.
Generally, both types will reduce the likely success rate of matching attempts. However, where two databases diverge from the world in the same way, referred to as *parallel divergence*, then the probability of matching is unaffected. This is the case, for example, when a respondent has lied consistently or when two data sets both have out of date data, but nevertheless have identical values.

Likelihood of success

Likelihood of success is *not* the same as the likelihood of achieving identification given an attack. Rather, it refers to the likelihood of the data intruder achieving his goal - which, in some scenarios may not be identification *per se* (for example, in scenario 3 the hypothetical journalist may get his 'good story' without a fully verified match).

Goals achievable by other means?

This is a crucial factor in determining the likelihood of an attempt being made. Can the data intruder achieve his goals by other means, which are easier to execute, legal and/or have a equal or better likelihood of success?

Consequences of attempt

Each scenario must also consider the likely consequences of an attempt. These will be dependent on the goals of the data intruder and the success or failure of the attempt. Broadly, consequences can be divided into two groups (i) whether or not confidentiality is broken, and (ii) the effect on public confidence.

(i) Whether or not confidentiality has been broken
We assume confidentiality to have been broken if at least one verified match has taken place, thereby identifying the target data set record of an individual. Confidentiality may not be broken by an attempt, either because a match cannot be achieved with the specified set of variables, or a match cannot be verified. Alternatively, the data intruder may not intend to break confidentiality *per se,* but to demonstrate that it could be done (as when an individual who is identified has colluded in the exercise).

(iia) Knowledge of the attempt to breach confidentiality is released into the public domain.
Whether or not an attempt is successful, information might be released into the public domain regarding the attempt. Release of such information in itself could be dangerous in terms of the effect on public confidence.
If the attempt is known to be successful then this effect would be exaggerated. If the attempt is unsuccessful or demonstrably unverified then the effect is potentially double-edged. The mere knowledge that an attempt has been made might have a damaging effect on the public perception of disclosure risk, by bringing the issue to the attention of the public. Also the fact that an attempt has been made indicates that someone believes a break in confidentiality is possible, which may have an adverse effect on perceived risk.
Against this, an unsuccessful attempt could be presented as an indication of the security of the confidentiality guarantee. This would be dependent on the public relations or 'fire-fighting" policy of the relevant national statistical Institute.

(iib) Knowledge that a breach of confidentiality is possible is released into the public domain.
This applies to scenarios where disclosure is demonstrated without a breach in confidentiality taking place (i.e. where the identified individual(s) have colluded in the disclosure exercise - scenario 4, below). Again the effect on public confidence would be dependent on the PR or 'fire-fighting" policy of the census department.

(iic) Details of matched individuals have been released into the public domain.
This is the most damaging consequence in terms of public confidence and co-operation with the census. When details of matched individuals pass into the public domain, the press, particularly the tabloid press, may run a 'personal story" which will tend to magnify (in the public mind) the importance of the confidentiality break.
This will be so even if the match is unverified.
Additional damage may be done if the information disclosed is sensitive or personally embarrassing to the matched individual. In the case of census records this is unlikely.

Likelihood of attempt

The likelihood of an attempt is a summary of the pragmatic estimate of the likelihood of an attempt being made given the other factors defining an attempt.
Again, because of the pragmatic/common sense nature of many of the inputs into this categorisation it is not possible to put a numerical probability to this likelihood.

Outputs of scenario analysis

Analysing a scenario of attack gives two outputs: the first is an estimate of the likelihood of particular attack being made against a particular database. This provides information about the value of considering the risk calculating: pr(identification|attempt) for that type of attack. The other output is the set of key variables that can then be used to estimate the risks associated with an attack given an attempt.

3. Key variables

The term *key variable* is used widely throughout the statistical disclosure literature. In general the term refers to a variable whose value is, for a given *target individual,* (a) known to a *data intruder* (an individual or organisation wishing to disclose information about the target individual) and (b) present in a *target database* to which the data intruder has access. It thus provides the basis for identification of a target individual.
Although the term has been widely used, it has not been operationally defined even for an *ad hoc* database. This paper also seeks to contribute to a more general discussion on the nature of key variables that is relevant cross nationally and able to inform the implementation of ARGUS or similar SDC software in different national contexts. Section 2 provides a first attempt at a taxonomy of key variables.

Key variables in the literature

The criteria employed by researchers in the selection of key variables have generally been *ad hoc*.
Greenberg and Voshell (1990) define key variables as '…those variables which, taken together, may contribute to the linking of a record to its respondent". In their own experiment Greenberg and Voshell select various combinations of 15 key variables; however they do not explain the rationale behind their choice of variables. The variables they chose are based on the 1980 US census and designed to mimic the Survey of Income and Program Participation (SIPP). They simply state that the 15 variables are '…possible key variables from SIPP". They are tenure, household type, race, ethnicity, children, marital status, payment, employment/unemployment/ veteran status, disability, household class, household income, social security, public assistance, and other income. [3] Bethlehem et al. (1990) use age, sex, marital status and household composition without any particular justification, however their purpose is only illustrative.
Muller et al. (1992) make the choice of key variables on the pragmatic basis of availability of an 'identification file" (this was the *Gelehrtyenkalender*: a directory of scientists) with which to conduct their experiments; thus the choice of key variables is exactly delimited by the overlap between the two files in their experiments. Muller et al.'s key variables are: region, sex, year of birth, half year of birth, industry, occupation, occupational status, specification of the professional activity, subject of the university degree.
Carter et al. (1991) use only four key variables: age/date of birth, sex, mother tongue, and household size in addition to geography. They state that the variables chosen are '… reasonably representative in terms of the degree of uniqueness likely to be met in practice.'
Marsh et al. (1991) select eight key variables for their experiments on population uniqueness using data from Tuscany, Italy: province, tenure, relationship to head of household, sex, marital status, occupational group, position at work and age. The selection is based on an empirical assumption: 'We selected as key variables the sort of socio-demographic information which large commercial databases in Britain may hold on individuals'
Paass (1988) conducts a more complex analysis based around six data intrusion scenarios, with the number of keys varying from seven to 68 (for a household file). As with Marsh et al. the exact choice is based on a pragmatic assumption of what the hypothetical data intruder might already know.
To summarise, the existing literature uses key variables in an *ad hoc* fashion based on pragmatic assumption or methodological considerations such as the availability of outside data. In the following section we examine key variables more closely in order to develop a more general framework.

A taxonomy for key variables

Variables that serve as keys may be classified in two ways. The first, the *key type,* refers to where the information may be found (in a public list, through personal knowledge etc.) and how available it is. The second, the *key quality* indicates how useful a key is in matching an individual uniquely in a target database.

Types of keys
Key variables may take a variety of forms. Broadly one can group them as:
1. *Publicly available information*
2. *Personal/ informal knowledge*
3. *Organisational database (usually commercial or government)*

1. Publicly available information
This is information that is publicly available as databases or directories. Examples are: Electoral Register (giving full name and address for adults in the household), telephone directories (billing addressee's name and household address), trade directories and registers of professional associations (full name, registered address, qualifications, date of registration). These are information formats that are common cross nationally and would be useable as key variable sources by a data intruder using target data from any country.

2. Personal/ informal knowledge
This includes information gained as a result of knowledge of a locality, for example local housing information. This type of information is particularly pertinent to released data of fine geographical detail. It also includes knowledge of other individuals through personal contact, either at work or through leisure activities. Brief social contact is likely to yield information on sex, ethnic group, and also possibly occupation, an estimate of age and marital status. Longer term friendships could easily yield detailed information on type of housing, household composition, age, sex and occupation of family members.

3 Organisational database (usually commercial or government)
This refers to databases held by commercial companies, for example lifestyle databases held by marketing firms, and also databases held by government agencies, used, for example, to administer benefits. All these databases will typically hold basic information on address, phone number, sex, age or date of birth. Additional information will vary with the type of database and is reported more fully below.

Meta-knowledge about keys

Meta-knowledge about keys is not information about actual values of key variables, but knowledge about the variables themselves. An obvious example is social knowledge about baselines. For example one knows *from common sense epidemiology* that the following combination of key variables are rare:
- black female bank managers
- 16 year-old widows
- Bangladeshis living in remote rural areas.

If one knows of such a person and finds a corresponding record in the target database then this common sense social knowledge might be used as a substitute for verification. [4] This is termed the special uniques problem (see Elliot et al. 1998 for a full discussion).

Key quality

Another dimension of key variables is the quality of the information that they contain. There are several elements of this.
Differentiation: Some keys are highly differentiating in that they are usually recorded at a high level of detail. Within the British census data, occupation, age and geography are the most obvious examples.
Skew: For some variables certain responses are far more likely than others. The most extreme example from the British census is the exclusive use of an inside toilet, where 'no' is a rare response and therefore a 'no' response provides a high differentiation of the respondent from the rest of the population. Conversely, 'yes' yields almost no information at all.
Measurement and coding error: It is well established that response errors and data coding errors or inconsistencies are more likely on certain variables than others. For example, in the British census *number of rooms* has a very high response error rate and as such is of limited use as a key (Heady et al., 1996). Target and identification data sets often diverge on variables where complex coding is required (such as occupation). Different data sets use different coding methods and there are always some responses for which the correct choice of code is ambiguous.
– *Stability*: The value of key variables may change over time. Whilst this can be due to measurement or coding error, this section is concerned with 'real' change or lack of it.

– *absolute stability* - variables which show no change over time, for example sex and date of birth.
– *one directional incrementation* - variables which change in one direction only. The most obvious example is age, which, for a given individual, can only increase incrementally. Other examples include number and level of educational qualifications.
– *categorical shifts - v*ariables on which the individual or household changes category, often related to life events such as moving house or changing jobs. Thus variables such as housing tenure or occupation are subject to categorical shifts.
– *miscellaneous predictable variation* - variables which change in a way that is more or less predictable, for example; number of children in household will often increase when the children are young and then decrease, once the children start to reach 16. This sort of stability is only probabilistic though, for example a family could suddenly take on foster children.

From this a two way classification of keys can be produced:

Quality	Availability	
	High	Low
High	Prime Keys	Critical keys
Low	Background keys	Inefficient Keys

Prime keys are those which we would expect to be present in every matching attempt.
Background keys are those which may be available but which are of limited value, due to high response error or instability.
Critical keys are those which are less likely to be available to a data intruder, but if available are useful.
Inefficient keys are those which a data intruder is unlikely to have access to and are of limited value if they do. For example, even if an intruder had reliable information on number of rooms in a household, this information is often misreported by survey respondents and therefore it is inefficient as a key.

- To summarise a high quality key variable is one which:
- differentiates the population into many categories.
- exhibits skew on those variables
- is not subject to large data divergence (response errors, coding variations etc.)
- is stable over time (either absolutely or through changing predictably)

4. Conclusions

Statistical disclosure control often uses a 'data-centric' rather than 'intruder-centric' analyses, this causes decisions about key variable selection to be *ad hoc* and therefore the assessments of risk to be ungrounded. Conducting an analysis of the scenarios of attack that might apply to a particular target data set provides (i) a rational estimate of a likelihood of a particular attack occurring and (ii) a set of key variables which are grounded in what is possible for the intruder. This paper has described a scheme for conducting such analysis.

References

Bethlehem, J.G., W.J. Keller and J. Pannekoek. 1990. Disclosure Control of Microdata. In: *Journal of American Statistical Association*. 85(409) pp. 38-45.

Blien, U., H. Wirth and M. Muller. 1990. *Identification Risk for Microdata Stemming from Official Statistics*. Discussion paper presented at the International Symposium on Statistical Disclosure Avoidance, The Hague, Netherlands.

Carter, R., J.-R. Boudreau and M. Briggs. 1991. *Analysis of the Risk of Disclosure for Census Microdata*. Statistics Canada Working Paper.

Dalenius, T. 1986. Finding a Needle In a Haystack. In: *Journal of Official Statistics* Vol. 2 No. 3 pp. 329-336

Elliot, M. J. 1996. *Attacks on confidentiality using the samples of anonymized records*. Paper presented to third international seminar on statistical confidentiality, Bled, Slovenia, October 1996.

Elliot, M. J. and A. Dale. 1998. *Disclosure Risk for Microdata*. Report to the European Union ESP/ 204 62/DG III.

Elliot, M. J., C.J. Skinner and A. Dale. 1998. *Special Uniques, Random Uniques and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk*. Proceedings of 1st International Conference on Statistical Data Protection. Lisbon, March, 1998.

Fuller, W.A. 1993. Masking procedures for Microdata Disclosure Limitation. In: *Journal of Official Statistics* Vol.2 No. 3 pp. 383-406

Greenberg, B. and L. Voshell. 1990. *The Geographic Component of Disclosure Risk for Microdata*, American Bureau of the Census SRD Research Report; Census/SRD/RR-90/13.

Lambert, D. 1993. Measures of Disclosure Risk and Harm. In: *Journal of Official Statistics* 9(2), pp. 313-332.

Marsh, C., A. Dale and C. Skinner. 1994. Safe Data versus safe Settings: Access to Microdata from the British Census. In: *International Statistical Review,* 62, 1, pp. 35-53.

Marsh, C., C. Skinner, S. Arber, B. Penhale, S. Openshaw, J. Hobcraft , D. Lievesley and N. Walford. 1991. The case for Samples of Anonymized Records from the 1991 Census. In: *Journal of the Royal Statistical Society* 154(2) pp. 305-340.

Mills, I. and A. Teague. Editing and imputing in the 1991 Census, In: *Population Trends*, 63, pp. 30-37.

Paass, G. 1988. Disclosure risk and Disclosure Avoidance for Microdata. In: *Journal of Business and Economic Statistics* 6(4) pp. 487-500

Notes

1) Both authors work at the University of Manchester's Centre for Census and Survey Research. The study on the scenarios of attack was partr of a larger European wide project headed by Leon Willenborg of Statistics Netherlands and funded by the European Union - ESPRIT Grant number 20462.
2) This is not to presuppose that such organisations are likely to make an attack, but is simply to ensure that all empirically possible types of attack are considered.

Exact disclosure in a super-table

Ton de Waal and Leon Willenborg

1. Introduction

The data published by a statistical office is often presented in tabular form. Several techniques exist to avoid disclosure of tabular data: cell suppression, rounding of values and data perturbation, see for example Willenborg and De Waal (1996). In this article we discuss another technique. Instead of publishing the entire table, i.e. the super-table, we publish a set of lower-dimensional tables, sometimes referred to as a set of linked tables, each of which presents only a subset of the variables from the entire table. Information about the cell values of the super-table can be obtained by linear programming techniques.

This approach is used in case of the more familiar cell suppression problem (cf. Cox, 1980). In fact the situation considered in the present paper is an extreme case of this: the entire super-table is suppressed and only some marginal tables of this table are released.

The aim of the present article is to investigate the possibility of reconstructing one or more values in the super-table exactly by combining information from the lower-dimensional tables. This requires two comments.

1. This problem is more restrictive than the one traditionally considered in cell suppression, namely whether any sensitive (and suppressed) cell can be recalculated within a certain predefined accuracy. If so, the corresponding cell is considered insufficiently protected. It is not too difficult to state the type of problem considered in the present paper similarly, but the current formulation of the problem offers a nice stepping stone to this more general information.
2. Exact calculation of the value of a cell in the super-table is not necessarily a disclosure in the ordinary sense of this concept. This is only the case if the cell concerned in the super-table is considered to be sensitive. This is what we shall implicitly assume. It is not an unreasonable assumption, because if no or only a few cells of the super-table were supposed to be sensitive, this super-table rather than some of its marginal tables would have been published.

It should also be remarked that the purpose of the present paper is directed at increasing the awareness of the problems associated with the publication of sets of linked tables, rather than giving a complete and scientific treatment of the problem of how to protect a set of such tables adequately. It should be noted that this is also reflected in the title of the paper: we address a disclosure problem rather than investigating its solution. For the interesting question is of course: which linked tables can be safely published instead of the corresponding super-table. And this question is not considered here.

Section 2 formulates the problem mathematically and introduces the notation we use. The problem statement is illustrated in Section 3 by some simple examples. Section 4 deals with a theoretical consideration of the problem and presents a possible way of solving it involving linear programming.

In Section 5 some extensions such as cell suppression and functional dependencies between variables are discussed. The article concludes with Section 6 in which a number of remarks are made and some open questions are posed which are or can be subjects for further research. It should be stated that we do not aim at giving a complete solution of the problem we are considering in this article. Our intention is merely to sketch the problem and to indicate a possible way of solving it. For more information on statistical disclosure control in general we refer to Willenborg and De Waal (1996).

2. Problem definition and notation

Suppose we have an N-dimensional super-table. The table presents outcomes of N different variables where variable j is supposed to be divided into m_j different categories $(j=1,...,N)$. Let $x_{i1}...{}_{in}$ $(i_j = 1,...,m_j)$ denote the value of cell $(i_1,...,i_N)$ in this table. This value may indicate the number of times the combination of properties $i_1,...,i_N$ occurs in the original data set on which the N-dimensional super-table is based (when we consider a frequency count table), or it may represent the magnitude (e.g. the profits) of this combination. For frequency count tables we have to demand that the cell values $x_{i1}...i_N$ $(i_j = 1,...,m_j)$ are non-negative; for magnitude tables this will depend on the meaning of the presented data.

Instead of releasing the entire N-dimensional table, we want to release a number of lower-dimensional tables in order to avoid disclosure of certain cell values of the original N-dimensional table. To this end, sets of variables are selected. The aggregation for variables of such a set then defines a subtable of the original table in the following way. Suppose we want to construct a table with the aggregations for the first k variables. The value of the cell $(i_1,..,i_k)$ in this k-dimensional table becomes

$$\sum_{j_{k+1}=1}^{m_{k+1}} \cdots \sum_{j_N=1}^{m_N} x_{i_1 \ldots i_k j_{k+1} \ldots j_N} \quad ,$$

where $i_l = 1,...,m_l$ and $l=1,...,k$. What we do is in fact nothing more than adding cell values over the variables k+1 until N. Aggregations for other variables can be obtained in a similar way. In this way we can construct a number of lower-dimensional tables from the original table.

Example 1: Consider a three-dimensional frequency count table in which each variable can assume two different values. Let the cell values be

$x_{111} = 8$, $x_{211} = 6$, $x_{112} = 4$, $x_{212} = 4$, $x_{121} = 5$, $x_{221} = 2$, $x_{122} = 2$, $x_{222} = 1$.

We want to determine the aggregation of this table for the first two spanning variables. The value of cell (1,1) in this table becomes $x_{111} + x_{112} = 8 + 4 = 12$ (since we have to add over the third variable). The other cell values can be obtained in a similar way. The two-dimensional aggregation becomes

Table 1

Variable 1	Variable 2	
	1	2
1	12	7
2	10	3

To avoid trivialities we shall assume that each variable appears in at least one table, that each variable has at least two categories, and that none of the released subtables is the super-table itself. The problem we shall discuss in this article is now whether it is possible to reconstruct exactly one or more cell values $x_{i_1}...i_n$ from the original table by combining information obtained from the lower-dimensional tables.

3. Examples

In this section we shall study some simple examples to illustrate our problem statement. All the data in the tables we present are fictitious.
Example 2: Consider a table with three variables, which can each assume two different values. So the table consists of eight cells. Information from this table is represented by Table 1 and the following two two-dimensional tables.

Table 2

Variable 1	Variable 3	
	1	2
1	13	6
2	8	5

Table 3

Variable 2	Variable 3	
	1	2
1	14	8
2	7	3

The question is now whether we can exactly determine at least one of the cell values of the original three-dimensional table. From the first table we can derive the following equations.

cell (1,1): $x_{111} + x_{112} = 12$,
cell (1,2): $x_{112} + x_{112} = 7$,
cell (2,1): $x_{211} + x_{212} = 10$,
cell (2,2): $x_{221} + x_{222} = 3$.

Similar equations can be derived from the second and third tables. The result is a system of 12 equations with 8 unknowns.
The number of equations can be reduced in the following way. Consider the equations that can be derived from cell (1,1) and (2,1) in the first table, viz.

$$x_{111} + x_{112} = 12, \quad (2.1)$$
$$x_{211} + x_{212} = 10, \quad (2.2)$$

and the equations that result from cell (1,1) and (1,2) in the third table, i.e.

$$x_{111} + x_{211} = 12, \quad (2.3)$$
$$x_{112} + x_{212} = 12,. \quad (2.4)$$

So we have four equations in four unknowns. The following relation can immediately be verified: (2.1)+(2.2)=(2.3)+(2.4). Therefore, one of the equations is redundant and can be omitted. It turns out that in this way five of the twelve equations are redundant and can be omitted without affecting the solution. The result is a set of seven linearly independent equations in eight unknowns with as general solution
$x_{111} = \lambda$, $x_{211} = 14 - \lambda$, $x_{112} = 12 - \lambda$, $x_{212} = \lambda - 4$,
$x_{121} = 13 - \lambda$, $x_{221} = \lambda - 6$, $x_{122} = \lambda - 6$, $x_{222} = 9 - \lambda$,
where λ is a real number. Suppose that we have as additional conditions on the unknown cell values that $x_{ijk} \geq 0$, with $i,j,k = 1,2$. This leads to $6 = \max\{0,6,4,6\} \leq \lambda \leq \min\{12,13,14,9\} = 9$.
Since more than one value of λ is possible the set of equations does not have a unique solution and none of the cell values can be determined exactly.
Note that the two-dimensional aggregations in this example were obtained from the cell values given in Example 1. If we substitute $\lambda=8$ in the general solution we get the original cell values.
One may wonder whether it is possible to find a unique solution if in addition to the given tables, the tables with the aggregations for one variable are also presented. This is not the case in this example. In fact in this case the one-dimensional tables are nothing but the marginal totals of the tables given above. The one-dimensional tables could play a role if some of the cell values in the two-dimensional tables were suppressed. We shall come back to this in Section 5.
In general, if a table is released in which the variables are a subset of the variables in another table, which is also released, then no additional information can be obtained from this table (unless the higher-dimensional table contains suppressed cells). The reason for this is that the table can be seen as a aggregation from the higher-dimensional table.

Example 3: Consider the following three two-dimensional aggregations of a three-dimensional table

Table 4

Variable 1	Variable 2	
	1	2
1	4	4
2	5	3

Table 5

Variable 1	Variable 3	
	1	2
1	3	3
2	5	3

Table 6

Variable 2	Variable 3	
	1	2
1	2	7
2	6	1

Again we aim at determining one or more of the values of the original three-dimensional table. As in Example 2 we obtain seven independent equations with eight unknowns with as general solution

$x_{111} = \lambda$, $x_{211} = 2 - \lambda$, $x_{112} = 4 - \lambda$, $x_{212} = 3 - \lambda$,
$x_{121} = 3 - \lambda$, $x_{221} = 3 + \lambda$, $x_{122} = 1 + \lambda$, $x_{222} = - \lambda$,

If we require that $x_{ijk} \geq 0$ for all *i,j,k*, then it follows immediately that λ=0. Hence, we can find a unique solution and the three-dimensional super-table can be reconstructed completely.
Note that without the third table we obtain a system of six independent equations in eight unknowns with as solution

$x_{111} = \lambda$, $x_{211} = \mu$, $x_{112} = 4 - \lambda$, $x_{212} = 5 - \mu$,
$x_{121} = 3 - \lambda$, $x_{221} = 5 - \mu$, $x_{122} = 1 + \lambda$, $x_{222} = \mu - 2$.

The conditions $x_{ijk} \geq 0$ imply $0 \leq \lambda \leq 3$ and $2 \leq \mu \leq 5$. So, in this case no unique solution exists. Without the first or the second table it is also impossible to obtain a unique solution.

Example 4: Consider the following three two-dimensional tables of a three-dimensional table

Table 7

Variable 1	Variable 2	
	1	2
1	4	2
2	2	2

Table 8

Variable 1	Variable 3	
	1	2
1	6	0
2	2	2

Table 9

Variable 2	Variable 3	
	1	2
1	5	1
2	3	1

Cell (1,2) in the second table yields $x_{111} + x_{122} = 0$. With the conditions $x_{ijk} \geq 0$ this implies. The values of the other unknowns follow immediately. Note that when the tables represent frequency count data, then the simple fact that $x_{111} = x_{122} = 0$ only reveals that the scores 112 and 122 on the combination of the three variables do not appear in the original data set. But in combination with the other information in the tables, it leads to the disclosure of the other cell values.
Zeros in an aggregation may form a potential risk for disclosure of information: although the fact that a certain variable equals zero does not in itself necessarily disclose sensitive information, it may help disclose information in other cells.
In general, let us consider the following two-dimensional aggregations of a three-dimensional table.

Table 10

Variable 1	Variable 2	
	1	2
1	a_{11}	a_{12}
2	a_{21}	a_{22}

Table 11

Variable 1	Variable 3	
	1	2
1	b_{11}	b_{12}
2	b_{21}	b_{22}

In the same way as in Example 2 we can derive seven linearly independent equations with the eight unknown cell values (i,j,k=1,2) from the original table with as general solution

Table 12

Variable 2	Variable 3	
	1	2
1	c_{11}	c_{12}
2	c_{21}	c_{22}

$x_{111} = \lambda$, $x_{211} = c_{11} - \lambda$,
$x_{112} = a_{11} - \lambda$, $x_{212} = a_{21} - c_{11} + \lambda$,
$x_{121} = b_{11} - \lambda$, $x_{221} = b_{21} - c_{11} + \lambda$,
$x_{122} = a_{12} - b_{11} + \lambda$, $x_{222} = a_{22} - b_{21} + c_{11} - \lambda$,

If we require that $x_{ijk} \geq 0$ for all *i,j,k*, then we obtain for x_{11} that
$\max(0, b_{11}-a_{12}, c_{11}-a_{21}, c_{11}-b_{21}) \leq \lambda \leq \min(a_{11}\ b_{11}, C_{11}, a_{22} - b_{11}+c_{11})$ (3.1)
We conclude from this expression that $x_{111} = \lambda$ can be reconstructed exactly if one of the terms from the left-hand side in (3.1) is equal to a term from the right-hand side. It follows immediately that in that case all the other cell values of the three-dimensional table can be reconstructed exactly as well. The given condition is satisfied in Example 3, in which $0 = a_{22} - b_{21} + c_{11}$, and in Example 4, in which b_{11} $a_{12} = a_{11}$. The condition is not satisfied in Example 2. Note that a different formulation of the condition can be obtained if we had set another unknown equal to λ.

4. Some theoretical considerations

The system of equations which can be derived from the aggregations (the lower-dimensional tables) will be denoted in the sequel by $Ax=b$, where A is an $m \times n$ matrix, x is the vector with the unknown cell values and b the vector with the known cell values from the aggregations. In this section we shall study the equation $Ax=b$ and concentrate on the question of when (at least) one of the entries of x can be determined exactly. Before that, we shall describe the structure of the equation in Section 4.1. In Section 4.2 we propose linear programming as a possible tool for settling the question.

The structure of the equation Ax=b

Consider an N-dimensional table which is represented by M aggregations and where variable j is divided into m_j categories (j=1,...,N). Suppose that one of the tables that we intend to release represents the aggregations for the first k variables. (There is no loss in generality in this, since this can always be achieved by rearranging the variables.) Let $b_{i_1,\ldots,i_k}$ (i_j = 1,...,m_j; j=1,...,k) denote the value of cell (i_1,...,i_k)(i_j = 1,...,m_j; j=1,...,k) in this table. If $x_{i_1 \ldots i_N}$ (i_j = 1,...,m_j; j=1,...,N) denotes the unknown value of cell (i_1,...,i_N) (i_j = 1,...,m_j ; j=1,...,N) of the original table, then we can derive the following equations from the table with the aggregations for the first k variables

$$\sum_{j_{k+1}=1}^{m_{k+1}} \cdots \sum_{j_N=1}^{m_N} x_{i_1 \ldots i_k j_{k+1} \ldots j_N} = b_{i_1 \cdots i_k}$$

where

and l=1,...,k. The number of equations which can be derived from a aggregation is equal to the number of cells of the aggregation, which is equal to the number of cells of the aggregation, which is equal to

$$\prod_{j=1}^{k} m_j$$

Similarly the equations are derived tem of equations $Ax=b$, where A is an *$m \times n$ (0,1)-matrix where m is equal to the total number of equations determined by a family of aggregations, and n* equals the total number of unknowns which is equal to the number of cells in the original table, i.e.

$$\prod_{j=1}^{N} m_j \cdot$$

It is clear from the equations that each table yields a set of equations in which each unknown cell value appears exactly once. The number of entries in each column of A that are equal to one is therefore equal to the number of lower-dimensional tables published. It is easy to see that the number of linearly independent equations that can be derived from the presented tables is strictly less than the number of unknown cell values of the original super-table. Stated in terms of the coefficient matrix A this means: $m<n$. However, in practice cell values are non-negative or integer valued by definition. Therefore, we turn to linear programming as a possible way of solving the problem. This is the subject of the next subsection.

A linear programming approach

Consider the equation $Ax=b$ with the constraints $x_j \geq 0$, j=1,...,n. The range of values that x_i can assume equals [min x_i, max x_i], where min min x_i and max x_i are the solutions of the following linear programming problems

minimise x_i	and	maximise x_i
subject to $Ax=b$, $x_j \geq 0$, (j=1,...,n)		subject to $Ax=b$, $x_j \geq 0$, (j=1,...,n) (4.1)

When min x_i = max x_i the unknown x_i is determined exactly. So, with the linear programming problems stated above it is possible to determine whether an unknown cell value can be reconstructed exactly.
Standard routines, which make use of the simplex algorithm, are available for solving the linear programming problems. Without loss of generality we assume that A is an *$m \times n$ matrix with $m < n$. We give two simple examples.*

Example 5: Consider the following three two-dimensional aggregations of a four-dimensional table.

Table 13

Variable 1	Variable 2	
	1	2
1	50	47
2	59	47

Table 14

Variable 1	Variable 3		
	1	2	3
1	37	30	30
2	38	39	29

Table 15

Variable 2	Variable 4			
	1	2	3	4
1	37	30	23	19
2	33	19	20	22

The number of cells (and hence the number of unknowns) in the original table equals the product of the number of categories (j=1,...,4) of the four variables, which is equal to 2×2×3×4=48. From the three tables we can derive 18, not necessarily linearly independent, equations in the 48 unknown cell values. By more careful counting the number of linearly independent equations is 14 [1)]. To determine whether at least one of the cell values can be reconstructed exactly, we compute the range of possible cell values for each of the cells from the original table by solving the linear programming problems formulated in (4.1). The linear programs are solved with standard routines for the simplex algorithm. In this case a total of 2×48=96 linear programs have to be solved. It turns out that for none of the unknown cell values are the lower and upper bound equal. The lower bounds are all equal to zero, while the upper bounds varied from 19 (for 12 cells) to 37 (for 3 cells). All the bounds

were integer-valued. Therefore, none of the cell values can be reconstructed exactly.
We have solved 96 LP-problems. This is the maximum for the given problem. In principle we could stop as soon as we find that for one of the cell values the lower and upper bound are equal, since then at least one of the cell values can be determined exactly.
Next, we add the following table to the three tables above and we are interested in how much tighter the ranges of possible values for the cells in the original table become.

Table 16

Variable 2	Variable 3		
	1	2	3
1	33	43	33
2	42	26	26

We still have 48 unknown cell values, but now we have 24 linear relations between the unknowns, of which 16 are linear independent. Again we use a standard LP-routine to determine lower and upper bounds for the cell values. In this case, too, for none of the cell values are the lower and upper bounds equal. Again all lower bounds are equal to zero. The upper bounds varied from 19 (for 12 variables) to 37 (for one cell). Six cells had lower upper bounds than in the case without the fourth table. All bounds were integer-valued.

Lastly, we add the following table

Table 17

Variable 3	Variable 4			
	1	2	3	4
1	26	17	21	11
2	20	19	14	16
3	24	13	8	14

The five tables yield 36 linear equations – of which 19 are linear independent - in 48 unknowns. Solving the LP-problems again yields the lower and upper bounds for each of the cell values. There are still no cell values with equal lower and upper bound, but the ranges are considerably smaller than without the fifth table. The lower bounds are all equal to zero while the upper bounds varied from 8 (for 4 cells) to 26 (also for 4 cells). Again all bounds were integer-valued.

Example 6: Suppose that from the same data set as in Example 5 we publish the following two tables.

The lower bounds for the unknown cell values in this case are again equal to zero. The upper bounds vary from 12 to 25. All bounds are integer-valued. Note that we obtain tighter bounds than when we publish the following four tables (the first four tables in Example 5): variable 1 by variable 2, variable 1 by variable 3, variable 2 by variable 3 and variable 2 by variable 4.

Table 18

Variable 3	Variable 3	Variable 4		
		1	2	3
1	1	16	18	16
1	2	21	12	14
2	1	17	25	17
2	2	21	14	12

Table 19

Variable 2	Variable 4			
	1	2	3	4
1	26	17	21	11
2	20	19	14	16

In general, linear programming problems do not necessarily yield integer-valued solutions. In the light of the application discussed in this article, this is a problem when the tables represent frequency count data, which are by definition integer valued. Suppose that one found the interval [2.3, 3.7] for one of the unknown cell values. If it is known that this cell value must be an integer, then it is clear that it equals 3. It is known that the linear programming programs (4.1) yield integer valued solutions if the constraint matrix A is a totally unimodular matrix (i.e. the determinant of each square submatrix of that matrix equals -1, 0 or 1; see Nemhauser and Wolsey, 1988). One might therefore hope that the matrices in the problem are totally unimodular. Unfortunately, this is not the case, as De Waal (1998) shows by giving a simple counterexample.
A computational problem is that the number of unknowns may be large, while for every cell two LP problems must be solved to determine whether $\min x_i = \max x_i$, hence computation time may therefore become quite large. It may be possible to reduce the computation time by making use of a modification of the simplex algorithm proposed by Cox (1980).
This modification is based on the observation that all variables appear in the optimisation of each variable. Therefore, we can examine the simplex tableau at each step to see whether one of the variables, not necessarily the objective function, has been optimised.

5. Extensions

Suppressed cells

Tables with cell suppressions can in principle be dealt with in a similar way to tables without cell suppressions. Because of the cell suppression, the information we can deduce from tables with suppressed cells may be less than from tables without suppressed cells. We illustrate this by the following example.
Example 7: Consider Example 2 in Section 3. Suppose we publish the same tables but with suppressed cells in the first one. This table is published in the following way, where suppressed cells are indicated by 'X'.

Table 20

Variable 1	Variable 2	
	1	2
1	12	x
2	10	x
Total	22	10

Note that because of the presence of the marginal, we in fact publish an additional table, namely the aggregation of the original data on variable 2. In the examples in Section 3 the marginal totals were not published, because in those examples the marginal totals only contained redundant information. In this example the marginal does contain additional information.
We can derive the following equations (cf. the equations in Example 2) from the aggregation for variables 1 and 2,
$x_{111} + x_{112}=12$,
$x_{211} + x_{212}=10$,

and the following equation from the aggregation for variable 2
$x_{121} + x_{122} + x_{221} + x_{222} = 10$.

The other equation that can be derived from the marginal total, is just the summation of the first two equations and therefore does not provide any additional information.

With the suppressed cells we can derive only three equations from this table instead of four from the table without suppressed cells (where we do not count the dependent equations which can be derived from the marginal totals and that do not yield extra information about the unknown cell values). We end up with a total of 11 equations instead of 12 in Example 2. The solution is the same as in Example 2.

Functional dependencies between variables

So far, we have assumed that there are no relations between the variables. In practice however, there will often be so-called functional dependencies between certain variables. For instance, it may be known that certain unknowns equal zero (e.g. the number of four-year-olds with a university degree) or that a certain inequality between two variables holds (e.g. between age and number of years of working experience). In general, we will speak about functional dependencies between variables if the set of values that a certain variable can assume is restricted by the score on another variable. When such functional dependencies hold, the extra information which we obtain from them can be used, in combination with the information from the tables, to determine the values of one or more unknowns. A simple example:

Example 8: If we add the condition $x_{122} \geq x_{112}$ in Example 2, it follows immediately that $\lambda = 9$ and hence every unknown can be determined exactly. The condition $x_{221} = 0$, for instance, also leads to a unique solution of the system of equations. Recall that without these conditions no unique solution exists.
So, in combination with information obtained from the tables, functional dependencies between variables may lead to the unique value of one or more cells in the original super-table. Therefore, when aggregations of a high-dimensional table are released, one should be aware of possible functional dependencies between the variables. The linear relations between the cell values, which can be derived from the linear functional dependencies between the variables, can be included in the linear programming problem.

6. Open questions and concluding remarks

In this article we have described the problem of reconstructing information in a table from aggregations of this table. This problem is of interest for the disclosure control of tabular data. Our aim is to avoid disclosure of information from a table by releasing lower-dimensional tables (aggregations of the original table).
The problem of choosing which aggregations of the original table to release is not addressed here. It is clear, however, that one should not release too many aggregations, since this may increase the risk of disclosure of cell values from the original table or data set (cf. Example 3). Also the appearance of zeros in certain aggregations (see Example 4) or functional dependencies between variables (see Example 8) may increase the disclosure risk.
Linear programming can be used to calculate lower and upper bounds for the cell values of the original table. A problem with this approach is that the number of LP-problems to be solved may be large, due to the possibly large number of cells in the original table.

References

Agresti, A. 1990. *Categorical Data Analysis*. Wiley, New York.

Cox, L.H. 1980, Suppression Methodology and Statistical Disclosure Control. In: *Journal of the American Statistical Association*, **75**, pp. 377-385.

De Vries, R.E. 1994. *Disclosure Control of Tabular Data Using Subtables*. Statistics Netherlands Report.

De Waal, T. 1998. *On Two Conjectures Concerning Linked Tables*. Statistics Netherlands Report.

Nemhauser, G.L. and L.A. Wolsey. 1988. *Integer and Combinatorial Optimization*. Wiley, New York.

Willenborg, L. and T. De Waal. 1996. *Statistical Disclosure Control in Practice.* Lecture Notes in Statistics, **111**, Springer-Verlag, New-York.

Notes

[1] The counting required is the same as for determining the degrees of freedom of loglinear (or other) models for tables (see e.g. Agresti, 1990).

Information loss through global recoding and local suppression

Ton de Waal and Leon Willenborg [1)]

1. Introduction

Before a microdata set can be disseminated safely by a statistical office it has to be protected against disclosure. More precisely, the risk of disclosure of confidential information should be sufficiently low before a microdata set may be released. Statistical disclosure control (SDC) aims to limit the disclosure risk. This can be done, as is the case at Statistics Netherlands, by checking whether certain combinations of scores occur frequently enough in the population (see e.g. De Waal and Willenborg, 1996b, Willenborg and De Waal, 1996). If such a combination does not occur frequently enough in the population the disclosure risk of the microdata set under consideration is considered too high and appropriate SDC measures should be taken.
A number of SDC measures can be taken to protect a microdata set: recoding, suppression and perturbation, for example. Recoding is collapsing categories of a variable, suppression is the replacement of a value in a record by a missing value, and perturbation is the replacement of one value by another one. Protecting a microdata set by one of these measures leads to a loss of information. Our aim is to retain as much information in the microdata set while making this data set safe. When microdata are interactively protected using μ-ARGUS the data protector employs an intuitive meaning of information loss. If the automatic mode of μ-ARGUS is used – in which the best combination of local suppressions and global recodings has to be found by the package itself – a quantification of information loss should be used that allows the package to make the necessary decisions and trade-offs. To achieve this it is necessary to quantify the information loss due to an SDC measure.
In this article we consider two methods to quantify the information loss in a microdata file due to local recoding, global recoding, or local suppression. One is objective and uses the entropy concept. The other is subjective, and uses weights specified by the data protector. These weights express the predilections of the data protector (who should be enlightened by the users' needs!) for preserving the information of certain variables over others and certain predefined codings for a variable over others.
The basic idea to evaluate the information loss formally is to use the entropy function for some suitably chosen probability distribution. This probability distribution is a stochastic model for the changes attributable to the applied SDC measures. The basis for the probability function is a 'transition probability matrix' for a variable. Such a matrix gives (an estimate of) the probability that an 'old' value in a record is changed to a particular 'new' one as a result of a modification of the microdata. The matrix for a particular variable in a file may be estimated by assuming a model for the changes due to the SDC measures and by estimating the corresponding probabilities. To estimate the probabilities it is necessary to make an assumption about the available information on which these estimates are based. For instance, these probabilities may be estimated by comparing the old and the new files and counting the number of changes that have occurred in the file with respect to the variable(s) under consideration.
With our discussion of information loss in terms of entropy we start by considering information loss caused by a technique that can be seen as more elementary than both global recoding and local suppression, namely local recoding. A subjective method for measuring information loss is described briefly in Section 5. Both these information loss measures are used in μ-ARGUS (cf. Hundepool et al., 1998) in the 'automatic mode' when using a mixture of global recoding and local suppression. The article concludes with a short discussion in Section 6.

2. Information loss due to local recoding

In order to be able to define the information loss due to global recoding and local suppression we first consider a related but simpler action, which we shall refer to as local *recoding*. We define local recoding as recoding a variable for *one* record only, whereas by global recoding we mean that a variable is recoded for *all* records in which one of the recoded categories occurs.
Suppose that a certain combination of scores in the file, e.g. 'age = 17' and 'marital status = widowed' does not occur frequently enough in the population. The records in which this combination occurs have to be protected. This can be achieved, as far as this particular combination is concerned, by recoding the variable 'marital status' in these records. For instance, the value 'marital status = widowed' may be replaced by 'marital status = widowed or divorced', assuming that the combination 'age = 17' and 'marital status = widowed or divorced' occurs frequently enough. In this case there is some uncertainty about the original value of 'marital status' for a user of the microdata set.
Now suppose that we can assign a probability p'_W to the event that the original value of 'marital status' equals 'widowed' given that the new, recoded, value equals 'widowed or divorced', and a probability p'_D that the original value equals 'divorced'. That is,

$$p'_W = \frac{p_W}{p_W + p_D} \tag{2.1}$$

and

$$p'_D = \frac{p_D}{p_W + p_D} \quad , \tag{2.2}$$

where P_w is the probability that the original value of 'marital status' in the record under consideration equals 'widowed', and P_d is the probability that the original value of 'marital status' equals 'divorced'.

In this case the entropy H_{ms} , i.e. the information loss due to local recoding of 'marital status', is given by

$$H_{MS} = -p'_W \log(p'_W) - p'_D \log(p'_D). \tag{2.3}$$

In general, when categories C_1, C_2,..., C_n of a variable V in a particular record are combined into a single one, denoted as $C_1+C_2+...+C_n$, then the information loss H_v due to this local recoding is given by

$$H_V = -\sum_{i=1}^{n} p'_i \log(p'_i), \tag{2.4}$$

where p'_i is the conditional probability that the original value of V in the record under consideration is equal to C_i given that the recoded value equals $C_1+C_2+...+C_n$. That is,

$$p'_i = \frac{p_i}{\sum_{i=1}^{n} p_i} \quad , \tag{2.5}$$

where p_i is the probability that the original value of V in the record under consideration equals C_i.

So far we have only considered the situation where one variable in one record is recoded. Now we consider the case where several variables are recoded in one record. When variables V_1, V_2,..., V_m are recoded in a particular record *k*, then the information loss H_k^r in record *k* due to these local recodings is measured by

$$H_k^r = -\sum_{i_1,i_2,\ldots,i_m} P(C_{1i_1}, C_{2i_2}, \ldots, C_{mi_m}) \log(P(C_{1i_1}, C_{2i_2}, \ldots, C_{mi_m})) \quad (2.6)$$

where $P(C_{1i_1}, C_{2i_2}, \ldots, C_{mi_m})$ is the simultaneous probability distribution of V_1, V_2,..., V_m. When we make the simplifying assumption that the variables V_1, V_2,..., V_m, are independent, then the information loss H_k^r in record *k* due to the local recodings is measured by the sum of the information losses due to the individual local recodings, i.e.

$$H_k^r = \sum_{j=1}^{m} H_{V_j}^r, \quad (2.7)$$

where $H_{V_j}^r$ denotes the information loss due to the local recoding of variable V_j, which is given by (2.4).

Lastly, we consider the case where several variables are recoded in several records. When variables are recoded in records 1, 2,..., *K* then we measure the total information loss H_{tot}^r due to local recodings in these records by the sum of the information losses in the individual records, i.e.

$$H_{tot}^r = \sum_{k=1}^{K} H_k^r, \quad (2.8)$$

where H_k^r denotes the information loss due to the local recodings applied to record *k*, which is given by (2.7). In fact, we assume here that local recodings in different records are independent.

In order to be able to calculate the information loss one needs to estimate the probabilities that appear in the entropy expression. We consider a crude model for this. Suppose that some of the old categories C_1, C_2,..., C_n of a variable *V* are combined in such a way that the new categories are given by D_1, D_2,..., D_m (m ≤ n), where each D_j is a combination of one or more C_i's. To evaluate the information loss due to this recoding we need to estimate the probability p_{ij}^k that the old, original, category equals C_i given that the new, recoded, category equals D_j in a particular record *k*. The index *k* indicates that
depends on (the values in) the record under consideration. We shall assume that D_j is obtained by collapsing C_1, C_2,..., and C_s. The probability p_{ij}^k *can then be estimated by*

$$\hat{p}_{ij}^k = \frac{n_i}{\sum_{t=1}^{s} n_t}, \quad (2.9)$$

where n_i denotes the number of times that C_i (i=1,...,s) occurs in the original, unprotected, microdata set. Note that this estimate does not depend on (the values in) record *k*.

3. Information loss due to global recoding

In practice, local recodings are hardly ever applied, because they lead to a rather odd kind of microdata set in which the categorisation of each variable can differ per record. Local recoding is used as a stepping stone to global recoding. Instead of local recodings one usually applies global recodings. This means that when a variable *V* is recoded in a particular record, then this recoding is applied to variable *V* in all records in the microdata set, thus achieving a uniform categorisation, i.e. the categorisation of a variable is the same for each record. For instance, when the categories 'widowed' and 'divorced' of the variable 'marital status' are collapsed into the single category 'widowed or divorced', then this is done for all records in which the value of 'marital status' equals 'widowed' or 'divorced'.

Measuring the information loss due to a global recoding is rather easy once the information loss due to local recodings has been defined, because a global recoding can be seen as a number of local recodings that have been applied to all records in the microdata set. Suppose that a variable *V* is recoded by combining some of the old categories C_1, C_2,..., C_n such that the new categories are given by D_1, D_2,..., D_m (m≤n), where each D_j may be a combination of several C_i's. In that case the information loss H_V^r due to recoding *V* can be measured, in each record in which the value of *V* equals D_j, by

$$H_V^r = -\sum_{i=1}^{n} p_{ij} \log(p_{ij}), \quad (3.1)$$

where p_{ij} denotes the probability that the original category of *V* in the record under consideration is equal to C_i given that the new category equals D_j. The total information loss in all records due to global recodings can be measured again by (2.8).

4. Information loss due to local suppression

A combination in a microdata set that does not occur frequently enough can also be protected by local suppression, i.e. one or more values in this combination can be deleted. For example, when the combination 'age=17' and 'marital status = widowed' does not occur frequently enough, then this combination can be protected by replacing the value of 'marital status' by a missing value. Local suppression of the value of a variable *V* is not applied to all records in a microdata set, but to some of the records only.
The information loss due to local suppressions can be measured in different ways. When local suppressions are not applied in combination with global recoding then the situation can be treated relatively easily. The information loss may be expressed as a weighted sum of the numbers of locally suppressed categories. When local suppressions are applied in combination with global recodings then the situation is more difficult. In this case the entropy should again be used to measure the information loss. Both situations, local suppressions not in combination with recodings and local suppressions in combination with recodings, are examined below. It should be noted that the entropy could also be used to measure the information loss when only local suppression is applied. Using the entropy is better from a theoretical point of view, but is more difficult to apply in practice.
De Waal and Willenborg (1994, 1998) considered the problem of finding the minimum number of local suppressions to eliminate a set of unsafe (=rare) combinations in a microdata file. The number of local suppressions was taken as a measure for the information loss due to the suppressions. In these papers this problem was extended to eliminating a set of unsafe combinations while minimising a more general linear target function. For instance, to each variable V_i a weight w_i can be assigned. The information loss in the microdata set due to the local suppressions is then measured by $\sum_i w_i s_i$, where the sum is taken over all variables and s_i equals the number of times that a value of variable V_i is suppressed. Using this linear target function instead of the non-linear entropy has the advantage that the problem of determining the optimal local suppressions is reduced to solving a 0-1 mixed integer programming problem, for which several algorithms are available.
The situation is different in for example Hurkens and Tiourine (1998a,b), where the goal is to find the optimum mix of local suppressions and global recodings to eliminate a given set of unsafe combinations. In this case it was necessary to find a trade-off in information loss due to either type of action. To be able to do this the entropy was introduced.
It is very simple to quantify information loss due to local suppression by means of the entropy once one realises that local suppression is an extreme form of local recoding, namely all categories of a variable are collapsed into a one single category. Information loss for the suppression of a value of variable *V*, having *n* categories, in a particular record is

$$H_V^s = -\sum_{i=1}^{n} p_i \log(p_i) \qquad (4.1)$$

where p_i denotes the probability that the original value of variable V in the record under consideration equals C_i
The total information loss due to local suppressions in record k, H_k^s, is given by formula (2.7) with $H_{V_j}^r$ replaced by $H_{V_j}^s$. The total information loss due to local suppressions in all records, H_{tot}^s, is given by (2.8) with H_k^r replaced by H_k^s. We measure the total information loss due to global recodings and local suppressions by

$$H_{tot}^{r+s} = H_{tot}^r + H_{tot}^s. \qquad (4.2)$$

Suppose that a value C_1, C_2,..., or C_n of variable V is suppressed in a record k. To evaluate the information loss due to this local suppression we need to estimate the probability p_i^k that the original category equals C_i. Again the index k indicates that, in principle, this probability p_i^k depends on the record under consideration. The probability p_i^k can be estimated in the following crude way.

$$\hat{p}_i^k = \frac{m_i}{m_V}, \qquad (4.3)$$

where m_i equals the number of times that C_i has been suppressed and m_V equals the number of times that a value of variable V has been suppressed. Note again that this estimate does not depend on the particular record k.

5. Subjective information loss measure: weights

In μ-ARGUS it is possible to use an automatic mode for protecting a microdata set. In this case the program searches for an optimal mix of global recodes and local suppressions to protect the microdata set. If the program is to perform this task, the data protector should make the necessary preparations, including specifying for each identifying variable a set of possible predefined codings (e.g. for a regional variable codings at the municipality, county, province and area are possible). Then for each variable he should indicate how important each variable is for him, by specifying weights for each variable. Also, for each variable he should indicate how important each of the alternative codings is for him, again by specifying weights. The weight that the system then uses for a particular coding of a particular variable is (proportional to) the product of the variable weight and the coding weight. Furthermore, the user should specify a weight for each identifying variable indicating how important it is that a value of this variable is suppressed.[2)] Unfortunately, the data protector is offered no guidance in specifying these weights in the current version of μ-ARGUS, but he has the full power of this weights approach available.

6. Discussion

The operational definition of the entropy-based model that we propose is based on the point of view of a statistical office, not that of the user of the published data. This is a consequence of the fact that to make the probability estimates one needs the original data as well, and these are obviously only at the disposal of a statistical office and not of data users. This is not restrictive in view of the particular application that we have in mind for this measure: as a kind of steering wheel in μ-ARGUS when the system is being used in automatic mode, finding the optimum mix of global recodes and local suppressions. In this mode safety of the resulting microdata is a goal, but no more modifications should be carried out than necessary. The information loss function is defined to steer the data modification process so as to make sure that enough interesting information will be left in the data.

The attractive thing about this entropy-based information loss approach is that it is general and versatile, deriving the information loss for various modification techniques such as global recoding and local suppression from a common principle, thereby making direct comparison in information losses due to different data modification techniques possible. As is shown in De Waal and Willenborg (1996a) - the paper from which the present article is derived - this method can also be used to quantify information loss caused by a perturbative method (such as PRAM, see Gouweleeuw et al., 1998).
The entropy-based measure of information loss is objective, as it is only based on objective information such as variables, domains, and objective probabilities over these domains. There is no option for a data protector to express his preferences for certain variables or certain recodings. Therefore we have discussed a second type of information loss model in the present article, which can actually be called entirely subjective. The model allows the user to express his preferences (over variables, values or codings) in terms of weights. The disadvantage of this approach – which is not shared by the entropy-based approach – is the difficult intercomparability for different data modification techniques. In practice this is likely to be only achievable by a certain degree of experimenting and fine-tuning, through judgement and valuation of the resulting safe microdata. In μ-ARGUS both information loss measures are used in the automatic mode: the entropy-based method at initialisation and the weight method as an option to change the initial weights. It must be admitted, though, that up to now there has been relatively little experience in setting these subjective weights satisfactorily. A lot of experimenting and testing is needed with real data to develop some intuition in applying this method. This is not unlike the problem a data protector faces when defining suitable cell weights in a table that is to be protected by secondary cell suppression.

References

De Waal, A.G. and L.C.R.J. Willenborg. 1994 *Minimizing the Number of Local Suppressions. Statistics Netherlands Report.*

De Waal, A.G. and L.C.R.J. Willenborg. 1996a *SDC Measures and Information Loss for Microdata Sets. Statistics Netherlands Report.*

De Waal, A.G. and L.C.R.J. Willenborg. 1996b. A View on Statistical Disclosure Control for Microdata. In: *Survey Methodology, Vol. 22, No. 1, pp. 95-103.*

De Waal, A.G. and L.C.R.J. Willenborg. 1998. Optimal Local Suppression in Microdata. In: *Journal of Official Statistics, Vol. 14, No. 4, pp. 421-435.*

Gouweleeuw, J, P. Kooiman, L.C.R.J. Willenborg and P.P. de Wolf. 1998. *The Post Randomisation Method for Protecting Microdata. In: Qüestiió, Vol. 22, No. 1, pp. 145-156.*

Hundepool, A.J., L.C.R.J. Willenborg, A. Wessels, L. van Gemerden, S.R. Tiourine and C.A.J. Hurkens. 1998, *μ-ARGUS User's Manual (Version 3.0), Report, Department of Statistical Methods, Statistics Netherlands.*

Hurkens, C.A.J. and S.R. Tiourine, 1998a, *On Solving Huge Set-Cover Models of the Microdata Protection Problem. Paper presented at SDP'98, 25-27 March, Lisbon, Portugal.*

Hurkens, C.A.J. and S. R. Tiourine, 1998b, *Models and Methods for the Microdata Protection Problem. In: Journal of Official Statistics, Vol. 14, No. 4, 437-447.*

Willenborg, L.C.R.J. and A.G.de Waal, 1996, *Statistical Disclosure Control in Practice, Lecture Notes in Statistics, Vol. 111, Springer-Verlag, New York.*

Notes

1) This article is an update of the report ‘SDC measures and information loss for microdata sets’ by the same authors that appeared in 1996.

2) In principle it is possible to use a more refined weighting system, which allows a user to specify weights for the identifying variables and also weights for each of the categories of a variables, similar to the global recoding case. But this is not implemented in the current version (3.0) of μ-ARGUS.

Statistical data protection at Statistics Netherlands

Peter Kooiman, Joris Nobel and Leon Willenborg

1. Introduction

The task of a national statistical office is to produce and publish statistical information pertaining to about society. When publishing statistical information, the statistical office has to assure itself - and its respondents - that no there be no disclosure of individual information is disclosed in any way. While traditional statistical disclosure protection techniques focused on tabulations, especially for economic statistics, changing practices in statistical analysis meant that statistical data protection also had to become more refined. The spectacular increase in computing capacity has stimulated a growing academic interest in microdata, initially for those on first for persons and households social statistics, but and more recently also for those on companies firms and institutions economic statistics. The development of new data protection techniques has followed this shift of focus.

Is the production and release of microdata part of the mission of national statistical offices? From a traditional perspective the answer is clearly negative: only statistical aggregates should be released. But one might also argue, from a perhaps more modern perspective, that a set of microdata represents the set of all possible tabulations, and therewith produces the best information and the maximum value added for users. In the Netherlands the legislator decided on the traditional perspective in 1996: the production of statistical, i.e. aggregated, information is defined as the task of Statistics Netherlands (Official Statistics Act 1996, section 3). The explanatory memorandum accompanying the law states literally that 'The release of microdata does not follow directly from the task assigned to Statistics Netherlands.' As a general rule, individual data are to be used by Statistics Netherlands exclusively for statistical purposes exclusively (ibid., section 11/1). They may not be distributed to anyone but the persons who perform the task of Statistics Netherlands (ibid., section 11/2). The resulting information may be published only in such a way that no information recognisable data can be inferred about an individual person, household, company, or institution, unless there is good reason to assume that they do not object (ibid., section 11/3).

As an exception to the general rule of statistical confidentiality the release of microdata is authorised for research purposes only as an exception to the general rule of statistical confidentiality, under certain, well-defined conditions. These conditions pertain to four relevant dimensions:

- the amount of detail in the data
- the users involved
- the conditions of use
- the subject matter of the data.

We shall elaborate on these in the second section referring explicitly to the legal mandate; Marsh et al. (1991) uses the terms safe data versus safe settings in this respect.

Several actors are involved in the application of the Statistics Netherlands microdata release policy. The third section contains our *tableau de la troupe*, sketching the respective roles, as we would like to see them, in the current organisational structure. The combined experience of organisational decentralisation, management rejuvenation, and a stronger general feeling for accountability has made clear - at least to the authors - the need to tighten up our microdata release practices.

At the same time Statistics Netherlands believes strongly in the adage 'No rules without tools'. The development of software to facilitate the assessment of statistical disclosure risk as well as the implementation of data protection measures has received much support, not only from Statistics Netherlands but also from the Statistical Office of the European Union (Eurostat). This software is described briefly in the fourth section.

The fifth and final section reviews a number of issues that are becoming increasingly relevant and must be solved in the near future. These issues concern our relationship with sponsors of statistical data collection efforts, the internationalisation of official statistics and research, the effects for our microdata release policy of the increasing use of administrative data for statistical purposes, and the growing popularity of geographical information systems.

2. Safe data and safe settings at Statistics Netherlands

The legal basis for Statistics Netherlands' microdata release policy is literally formulated as a set of exceptions to the general rule of statistical confidentiality. Statistics Netherlands is allowed to release microdata only in aid of statistical or scientific research (Official Statistics Act 1996, section 13/1). It is obliged to take adequate measures to prevent recognition of individual persons, households, companies or institutions (ibid., section 13/1). And users have to take sufficient measures to ensure that microdata will not be used for other purposes than for statistical or scientific research (ibid., section 14). For certain kinds of data the conditions are stricter.

The law defines which *institutions* are considered to perform statistical or scientific research. Apart from the legally recognised universities, a couple of governmental think tanks (called 'planning offices' in Dutch, for historical reasons), and Eurostat are mentioned explicitly under this heading. Other services, organisations, and institutions may apply; the Central Commission for Statistics, the independent supervisory body for Statistics Netherlands, has to agree with the application and has to authorise Statistics Netherlands to release microdata to the applicant. Applicants have to convince the Commission that they have a primary research function, that their results are targeted at the forum of science and the public domain in general, and that their data cannot, and will not, serve administrative objectives targeted at individual data subjects. Only government sector research departments can hope to be considered as a statistical enclave in this respect. Research units of private companies are not entitled access to Statistics Netherlands microdata. In the past two years, several requests from such private companies have been turned down by the Commission, although a number of private organisations specialising in policy analysis, for example, do have access to Statistics Netherlands microdata.

The law also defines which *data* may be released: adequate measures have to be taken to prevent disclosure. Two aspects are relevant here. First, the notion of adequacy entails a certain degree of judgement. And this judgement is the exclusive authority of the Director-General of Statistics, the chief statistician in the Netherlands. The explanatory memorandum accompanying the law states literally that 'Statistics Netherlands is allowed a certain leeway to determine the amount of protection of the files. An exact amount of protection cannot be defined in the law.' Secondly, this judgement must necessarily be dynamic, as circumstances may change. In this respect one must obviously think of the exponential growth of computer capacity and the expanding public and digital availability of individual-level databases. The department of statistical methods at Statistics Netherlands is the Director-General's prime centre of expertise to put the criteria for his judgement into practice.

The law requires Statistics Netherlands to ensure that the applicant has taken sufficient measures to preserve statistical confidentiality. Statistics Netherlands must be convinced that the microdata will not

be used for any other purpose than statistics or research. The main instrument to determine the *conditions of use* is a contract signed by Statistics Netherlands and the applicant. The terms of this contract refer to the storage (not on open networks), the use (matching is not allowed without explicit consent from Statistics Netherlands) and the preservation (deletion of data after the research project is finished) of the microdata, the publication of results (only after they have been screened for safety), etc. Statistics Netherlands has the right to review the measures taken by researchers, by inspecting hardware configurations on site and by previewing draft publications.

The subject matter of the data and accompanying legal provisions may call for tailor-made approaches. The regime sketched so far applies to Statistics Netherlands microdata in general; some data and laws call for more restrictive approaches. In particular, business microdata are exempted from the general approach: the 1936 Economic Statistics Act is not overruled by the 1996 Act (Official Statistics Act 1996, section 11/5). The older law - whose roots go back to 1917 - is quite strict in its clauses on statistical confidentiality. Individual company data may never leave Statistics Netherlands in an identifiable form. A combination of reasons may explain this strictness, one obvious one being the accompanying response obligation for companies. Another reason is the fact that business microdata are much more easily identified than household microdata: the population is smaller, identifying variables have more values, their distributions are much more skewed, knowledge of the composition of specific sub-populations is more widespread, etc. One should also be aware that the business microdata may represent strategic information or even company secrets for the respondents. There are other formal provisions that require a restrictive approach to particular sets of data: the causes-of-death microdata, for example. Many data that Statistics Netherlands receives from administrations, also call for restrictions. The holders, i.e, the original collectors - such as the fiscal administration - wish to avoid a perception among the general public that their personal data are, as it were, in the public domain.

The access options for Statistics Netherlands data are summarised in Table 1. The vertical dimension refers to the type of data, type being a crossing of level of detail and subject matter area. The other dimension refers to the screening of users alluded to above. Within the cells, the methods and conditions of use of the data are given.

Everybody has access to safe tables. Traditionally these were published on paper, today tables are even more easily accessible through digital databases, such as Statistics Netherlands' StatLine. If microdata can be protected to such an extent that statistical disclosure is guaranteed to be impossible, such microdata may be distributed to anyone as well. Although their analytical value for research purposes will be limited because of the lack of detailed variables, they may still be useful for teaching purposes. In practice such public use files are released only occasionally.

If statistical disclosure cannot be ruled out completely, additional measures have to be applied, such as screening applicants and restrictions on the conditions of use. In the Netherlands, only researchers will have access to such data. In general, they will have access to standard microdata under contract (MUC) for analysis purposes. A contract covers the remaining risks. The files are standardised for several reasons. First this reduces cost, as repeated disclosure protection analysis for different versions of the microdata is avoided. Secondly, disclosure protection is considerably more difficult if it has to take account of the extra risk of disclosure through matching different versions of the same microdata. Thirdly, the disclosure analysis can be more thorough if it has to be performed only once, increasing the safety of the microdata files. Fourthly, one standard delivery can be planned better than ad hoc deliveries, thus improving user services such as documentation and timeliness. The release of MUCs was speeded up considerably in 1994 when Statistics Netherlands and the Netherlands foundation for scientific research NWO entered into a four-year agreement to this effect. NWO agreed to pay one million guilders per annum and Statistics Netherlands agreed to release annual MUCs for all major social surveys. NWO even founded a small brokering statistical agency (WSA) to enhance contacts between providers and users. It now organises regular producer-user meetings on specific files. The agreement largely ended the long-running cold war between academia and Statistics Netherlands. Its success was recognised by an evaluation committee and led to its prolongation for another four-year period.

For specific levels of detail (e.g. regional detail), subject matter (e.g. causes of death), and sources (e.g. administrative data) microdata files are not distributed to the users. However, they may be made available on site. It goes without saying that on site analysis is available only for those who cannot work with the standard MUCs, and only if they are qualified according to the Central Commission for Statistics.

The success of the household MUCs and the progress of economic research have stimulated academic interest in access to business microdata enormously. The academic community has made this interest explicit in *Kijk op Economische Kennis (View on Economic Knowledge)*, a report on the state and future of economic research in the Netherlands. Both legal clauses and the opinions of leading businesses necessitated a very cautious approach by Statistics Netherlands, and it took some time before we could convince the business world of the great public benefit, and the relative operational safety of on site access to business microdata for a select group of researchers. There are special provisos for business microdata: only Dutch, verifiably academic researchers have access to the on site facility. The extremely cautious approach is based on a real fear of industrial espionage and government curtail. For the same reason an initial three-year pilot period has been agreed upon.

As of the beginning of 1998 researchers may analyse available business microdata at Ce*R*eM, the Centre for Research of Economic Microdata, at Statistics Netherlands. They have a computing environment at their disposal with even more software than is usually available to Statistics Netherlands' own researchers. In the interest of statistical confidentiality this system has no digital connections (no diskette stations or modems) with the outside world. Furthermore, Statistics Netherlands staff always screen

Table 1. Current Modalities of Access to Data by Type of Data and Type of User

Type of data	Type of user	
	everybody	Researchers only
Safe tables	paper, desk top	
Public use files	desk top	
Standard microdata under contract		desk top
Specific subject-matter microdata		on site
Business microdata		on Site, NL only, 3 years

printed results before these may leave the premises. NWO is investing heavily in the further development of this Ce*R*eM facility.

3. Actors in statistical data protection within Statistics Netherlands

This section briefly discusses the various tasks involved in statistical data protection at Statistics Netherlands and the individuals or departments who are in principle responsible for them. In some cases the description of the tasks is not always in exact agreement with current practices, but refers to a situation that we hope to reach in the near future.

Director-General

The Director-General is responsible for the policy regarding the release of microdata. In the first place he sets the rules and conditions determining which microdata can be considered safe enough to be released under contract. In the second place he sets the conditions under which data may be made available for research, the terms of the contract.
The Director-General's policy staff advise him on legal and corporate aspects of statistical confidentiality policy as they may occur both in draft and implemented legislation and in the contracts. The policy staff also assist or even represent Statistics Netherlands directors in European Union meetings such as those of the EU Committee on Statistical Confidentiality. These are mentioned explicitly because of the increasing importance of Community legislation within the EU member states, even to the extent where Community legislation may overrule or substitute national legislation and policies. The policy staff also has an internal audit section. At some stage of its programme this section may look into statistical data protection practices within the various subject matter departments.
Another tool that can be employed to further discipline the statistical departments is the management contracts between the Director-General and lower management levels. These contracts specify the results that the various units within Statistics Netherlands are expected to achieve in the course of a certain year. The Central Staff assist the Director-General in drafting and guarding these contracts, and are thus instrumental in setting out general policy lines in an administrative setting. Clauses within these internal management contracts may help implement statistical data protection.
There is only one respect in which the Director-General is not responsible for statistical data protection policy. It is the Central Commission for Statistics that determines to whom he may grant access to microdata. Since proposals for this Commission are usually drafted by the Bureau, and the Commission's Secretary is on the Director-General's policy staff, one would expect that Commission decisions in this respect be in agreement with Bureau policies.
And if the Commission were too liberal in granting access to microdata, the Director-General might still decide to apply further restrictions on the data and the settings as a counterweight.

Department of Statistical Methods

The rules and conditions are proposed by the department of statistical methods, which is part of the Research and Development Division. Among other areas of work, this department is engaged in research, consultancy and education in the field of confidentiality. The major portion of research work was published in statistical journals, conference contributions or books (e.g. the special issues of *Statistica Neerlandica* (1992), *Journal of Official Statistics* (1998), as well as the present issue of *Netherlands Official Statistics*, and the monograph by Willenborg and De Waal (1996)). In the past, research in the disclosure control area focused on such topics as the estimation of the number of population uniques in a microdata file (Bethlehem et al., 1990), the assessment of the risk of spontaneous recognition of a data subject being known to a user of the microdata file, the modelling of an intruder's knowledge of a population in terms of circles of acquaintances (cf. Mokken et al., 1992) the formulation of optimisation models for the protection of a microdata set by the optimum application of certain data modification techniques (in terms of information loss), the protection of sampling weights in microdata files (cf. De Waal and Willenborg, 1997). More recently, research has been directed at a perturbation method for adding stochastic noise to categorical data called PRAM (cf. De Wolf et al. (1998), Gouweleeuw et al. (1998), Kooiman et al. (1997), and at a method to identify risky records in a microdata set called fingerprinting (cf. Willenborg and Kardaun, 1999), and at the development of heuristics to carry out cell suppression and rounding in hierarchical tables.
Other departments and staff - including the Director-General - can consult the department of statistical methods for specific applications. Such advice may be requested both on general policy issues and on specific methodological problems concerning the protection of statistical data.
At present the department of statistical methods also hosts the *CeReM* on-site research facility mentioned above. It is worth mentioning that the department bears no direct responsibility for the safety of the data to be released by the subject matter departments. It is only responsible for developing and maintaining an infrastructure of knowledge, policies and tools within the Bureau. It is up to the subject matter departments, who own the data, to conform with the official policies imposed when they prepare statistical data for release.
The educational task of the department of statistical methods is primarily directed at the internal needs of Statistics Netherlands itself, but not exclusively. Relevant courses have also been given by department staff for European statisticians (through TES), and over the years a sizeable group of students has worked as trainees at the department, working on their master's theses on subjects in the field of statistical confidentiality. This provides an interesting and much appreciated link with academia.

Subject-matter departments

The most important actors, for all practical purposes, are the statistical subject-matter divisions. They release publications and microdata; they enter into contractual arrangements with third parties, etc The 1994 reorganisation of Statistics Netherlands (known under the acronym TEMPO) introduced so-called integral management within the various statistical and supporting divisions. In particular, the earlier right of veto for the department of statistical methods for the release of microdata was abolished. Checks against official policy are still carried out within the statistical departments, but there is no longer a formal requirement to involve the department of statistical methods, or even to inform them. Moreover, increasing budgetary pressure went hand in hand with a growing desire to operate in more client-friendly ways. Fast and substantive turnover of staff in management positions have not added to the management awareness of statistical confidentiality issues either. In certain cases the combination of such factors may have been to the detriment of statistical confidentiality as a policy or principle. Although in practice no major incidents have surfaced so far, there is an apparent need for a better grounding of confidentiality protection policies within the subject matter departments. This is to some degree an educational problem.
At the same time the agreement with NWO (1994), the new Dutch statistics law (1996) (cf. also Nobel 1994, 1996), and the start of Ce*R*eM (1998) have created opportunities to promote the cause of statistical confidentiality. The success of the Statistical Agency and Ce*R*eM has led external researchers to perform research rather than to oppose Statistics Netherlands' policies. Moreover, the resulting standardisation of microdata content has made statistical

disclosure protection somewhat less of a burden for Statistics Netherlands staff. And as the amount of available microdata and the scale of use increase, the on-site facilities are becoming more professional too.

4. No rules without tools

The rules that have been formulated at Statistics Netherlands for statistical data protection almost require that specialised software is available to apply them comfortably in practice. The reasons for this are manifold. In the first place the calculations that have to be performed are too difficult to carry out by 'homebrew' software produced by a non-specialist in statistics or operations research. Besides, standard statistical packages do not offer several of the options required. With a specialised tool it should be fairly easy to help produce safe data in a relatively short period of time, and generate the corresponding documentation. Lastly, the use of a standard tool for data protection enhances the correct application of the data protection rules, rules which are a 'corporate issue' for Statistics Netherlands, not a local one, i.e. they transcend the level of divisions.

The need for data protection software triggered the development of a specialised data protection package called ARGUS. In fact ARGUS actually consists of two packages, namely μ-ARGUS for the protection of microdata, and τ-ARGUS, for the protection of tables (cf. Hundepool et al., 1998a,b), The development of μ-ARGUS started in the early nineties as a program, then simply called ARGUS, while τ-ARGUS is the successor to a program called SUPPRESS developed at the department of statistical methods in the eighties.

The rules applied at Statistics Netherlands to protect microdata and tables were taken as a starting point in the design of ARGUS. The ambition with the current version of ARGUS - which was developed in the SDC project sponsored by the EU through the Esprit Programme - was to develop a tool that is useful for the application of those *and similar rules*, so that it would be of more general interest, i.e. not only for Statistics Netherlands but also for other statistical offices as well (this was also a requirement for the funding of the project). For tabular data the rules that Statistics Netherlands applies require the use of concepts and techniques that are generally accepted (e.g. dominance rule, secondary cell suppression, controlled rounding, etc.). The only limitation here was in terms of the dimensions of the tables and the number of tables that could be handled (i.e. only single tables and no linked and hierarchical tables). These limitations were mainly inspired by practical considerations and they are likely to be eliminated in future releases of the program. For μ-ARGUS the case is different, in the sense that there did not seem to be much else to take as a starting point. At Statistics Netherlands quite a lot of research effort was put in the development of rules for two sets of microdata, namely MUCs and Public Use Files. The feeling was that (straightforward generalisations of) these rules would be interesting and general enough to warrant their implementation. For a description of some applications of the current version of ARGUS within Statistics Netherlands, see Schulte Nordholt (1999).

5. Future issues

In trying to strike a balance between the need for statistical data protection and that for socially relevant research on microdata, the importance of a major deal with, in our case, NWO can hardly be overestimated. Currently we can envisage three devices to improve safe access to microdata. They would all require further research and/or budgets (and a basic trust in the intentions and practices of researchers, we are tempted to add). These three devices would be to:

1 make current on-site facilities within Statistics Netherlands more professional;
2 open up local data centres within major research centres under the authority of Statistics Netherlands; and to
3 open up possibilities for off-line and on-line analysis of micro data.

The first option has the distinct advantage of the closeness to the official statisticians who probably know more about the data than anyone else. The disadvantage of travelling to the Statistics Netherlands establishments (Voorburg and Heerlen; two and a half hours apart from each other by train) can hardly be considered prohibitive given the small size of the Netherlands.

The second has already been cautiously adopted or will be in the near future in Canada, Denmark and the USA. The costs are considerable, however, and should be weighed against those of researchers travelling to the Statistics Netherlands' premises. If there is a relation with a local data collection centre or with a joint research programme, initial costs for such a solution might be reduced.

The third option allows a much better control of the use and access rights to the data. Microdata are not distributed: there is one and only one authentic set. All activities performed on this data can be logged, and later analysed. And it is possible to deny and prohibit access to specific parties. But the physical and logistic problems involved are heavy indeed. Off-line analysis gives more safeguards than on-line analysis, although the use of cryptographic techniques may provide sufficient protection for the communication channel used. Statistics Netherlands experimented with an off-line approach recently when a researcher had analysed his microdata for a particular year quite extensively and wanted to perform the same analyses in equivalent ways on the corresponding microdata for other years.

A matter of prime concern is our relationship with sponsors of statistical data collection efforts. Often when additional financing outside the central budget is required - and found - for specific data collection efforts, the sponsor wants his share of the microdata. The interests involved - both in terms of information and in terms of money - can then be quite substantial. The situation becomes even more complicated when a government department sponsors a specific data collection effort, since such departments are usually involved in administrative, as opposed to statistical, tasks as well. Not only does this pose a question of the boundaries between statistical and administrative use of the data, but also the departments concerned have all kinds of administrative data themselves to match with. Our present position on this is that no sponsor of a data collection receives special treatment: data dissemination policies are exactly the same as for any other third party. In particular they will not receive more detailed microdata than the standard microdata sets that we prepare for academic research. Also, they will only obtain access to these standard sets if they satisfy all the conditions discussed in Section 2.

Another concern is the internationalisation of official statistics and research. There is an obvious need to find new ways for EU research on microdata. We feel that for the time being we should look for the highest, rather than the lowest level of protection. And from this perspective Eurostat and the European Commission may still have a few things arrange.

One explicit element of our long-term strategy involves the use of administrative data as a source for statistics. This has a number of consequences for our microdata release policy. First, surveys will increasingly be designed as supplementary to administrative sources already available to the Bureau. Survey information in itself will therefore increasingly be partial, reducing its usefulness for secondary analysis by researchers. The richness of our current microdata sets can, as we see it, only be preserved in the long run if we include administrative data in these sets as well, properly matched to the survey data. It will require a considerable effort on our part to convince the owners of the administrative data that this does not violate their privacy concerns. Secondly, when we release matched administrative/survey microdata sets the suppliers of the

survey data, i.e. our respondents, have to comply as well. At present it is our feeling that Dutch public opinion does not yet favour such policies. However, things may gradually change as we continue to demonstrate that their concerns are also our concerns, and that no one is better equipped than we are, both in terms of privacy awareness and expertise on disclosure risks, to prepare safe data sets for secondary analysis. Ultimately, the function of data broker for secondary statistical analysis of (primarily) administrative data sources might develop into a logical component of the mission of national statistical offices.

Another issue is the risks inherent in the growing popularity of geographical information systems. While detailed regional information may be extremely useful for government, the media, and academic research, from the perspective of statistical confidentiality it may entail serious problems. Regional detail is considered to be the most dangerous re-identification key of any variable in a data file. Traditionally, the statistical offices applied their dominance rules to tabulations and checked the number of establishments within a certain area, just to make sure. By projecting statistical data onto maps, and linking such maps it would become increasingly easy for a user to identify

- areas where illegal products are being dumped; or
- areas where certain rare species of butterflies, plants etc. can be found.

Public opinion in itself may even be at stake. If the production and dissemination of a large amount of junk mail is enabled by detailed statistical publications from statistical offices (detailed regional information about population composition and income distribution), the public might begin to wonder what purpose our efforts to produce detailed statistical information serve.

The final point concerns scientific collaboration in the statistical confidentiality area. An interesting feature of the SDC project was the co-operation between official and academic statisticians. Although such co-operation seems to be quite normal within, for example, the United States of America and the United Kingdom, statistical data protection has hardly generated any interest from Dutch academic statisticians so far. The only interaction with the academic world in the Netherlands on confidentiality research up to now has been through traineeships of students working on their master's degree. We feel that it would be beneficial to both the academic and official statistical communities in the Netherlands if the former were more actively involved in confidentiality research. It is likely that this can be achieved through the collaboration in joint projects in this area that are subsidised by institutions like NWO or the EU.

References

Bethlehem, J.G., W.J. Keller and J. Pannekoek. 1990. Disclosure control of microdata. In: *Journal of the American Statistical Association*, Vol. 85, pp. 38–45.

De Waal, A.G. and L.C.R.J. Willenborg. 1997. Statistical disclosure control and sampling weights. In: *Journal of Official Statistics*, Vol. 13, No. 4, pp. 417–434.

De Wolf, P.-P, J.M. Gouweleeuw, P. Kooiman and L.C.R.J. Willenborg. 1998. *Reflections on PRAM*, Paper presented at the Statistical Data Protection'98 conference, 25–27 March, Lisbon, Portugal.

Gouweleeuw, J.M., P. Kooiman, L.C.R.J. Willenborg and P.-P. de Wolf. 1998. The post randomisation method for protecting microdata. In: *Questiió*, Vol. 22, No. 1, pp. 145–156.

Hundepool, A.J., L.C.R.J. Willenborg, A. Wessels, L. van Gemerden, S.R. Tiourine and C.A.J. Hurkens. 1998a. μ*-ARGUS user's manual (Version 3.0)*, Statistics Netherlands.

Hundepool, A.J., L.C.R.J. Willenborg, L. van Gemerden, A. Wessels, M. Fischetti, J.-J. Salazar and A. Caprara. 1998b. τ*-ARGUS user's manual (Version 2.0)*, Statistics Netherlands.

Fienberg, S. and L.C.R.J. Willenborg (eds.) *Journal of Official Statistics.* 1998. Vol. 14, No. 4, Special issue on confidentiality

Keller, W.J. and P. Kooiman (eds.). 1992. *Statistica Neerlandica*, Vol. 46, No.1, Special issue on statistical confidentiality.

Kooiman, P., L.C.R.J. Willenborg and J.M. Gouweleeuw. 1997. *PRAM: a method for disclosure limitation of microdata.* Statistics Netherlands Report.

Marsh, C., A. Dale, C. Skinner. 1991. *Safe data versus safe settings: access to customised results from the British Census.* Proceedings 48th Session of the International Statistical Institute, Cairo, Egypt.

Mokken, R.J., P. Kooiman, J. Pannekoek and L.C.R.J. Willenborg. 1992. Disclosure risks for microdata. In: *Statistica Neerlandica*, Vol. 46, No. 1, pp. 49–67.

Nobel, J.R. 1994. *Data confidentiality and data access- practical and legal issues in the Netherlands*, Proceedings of the International Seminar on Statistical Confidentiality, 28–30 November, Luxembourg.

Nobel, J. 1996 A law on official statistics. In: *Netherlands Official Statistics*, Vol. 11, Summer, pp. 48–53.

Schulte Nordholt, E. 1999. Statistical disclosure control of the Statistics Netherlands employment and earnings data, *this issue.*

Willenborg, L.C.R.J. and A.G. de Waal. 1996. *Statistical disclosure control in practice,* Lecture Notes in Statistics, Vol. 111, Springer-Verlag, New York.

Willenborg, L.C.R.J. and A.J. Hundepool. 1998. *ARGUS for statistical disclosure control.* Paper presented at the Statistical Data Protection'98 conference, 25–27 March, Lisbon, Portugal.

Willenborg, L.C.R.J. and J.W.P.F. Kardaun. 1999. *Fingerprints in microdata sets,* Paper presented at the Eurostat/UN-ECE Seminar on Statistical Confidentiality, Thessaloniki, Greece.

Disclosure limitation practices and research at the U.S. Census Bureau

Laura Zayatz, Paul Massell and Phil Steel[1)]

1. Introduction

The United States Census Bureau collects data via censuses and surveys. The data are collected under Title 13 of the U.S. Code which prevents the Census Bureau from releasing any data '...whereby the data furnished by any particular establishment or individual under this title can be identified.' At the same time, the agency has the responsibility of releasing data for the purpose of statistical analysis. The desire, then, is to release as much statistically valid and useful data as possible without violating the confidentiality of the data as stated in Title 13. Disclosure limitation techniques are applied to the data prior to their release in an effort to protect confidentiality. This contribution discusses disclosure limitation practices currently in effect at the Census Bureau, as well as current Census Bureau research into alternative disclosure limitation procedures.

2. Publicly released Census Bureau data

If the Census Bureau releases a set of data to any outside data user (anyone not employed at the Census Bureau), then those data are publicly available. Anyone can request a copy of them, and the Census Bureau will supply them, perhaps for a small fee. In other words, the Census Bureau cannot choose to release a data set to some outside users and not to others.

The three most commonly used forms of data release include microdata, frequency count data, and magnitude data. A microdata file consists of records at the respondent level. Each record represents one respondent and consists of values of characteristic variables for that respondent (Federal Committee on Statistical Methodology, 1994). Typical variables for a demographic microdata file include age, occupation, and income of a responding individual. Variables for an economic microdata file might include employment size and value of shipments of a responding establishment. Because economic data are highly skewed and establishments are often easily identified by just a few characteristics, the disclosure risk of economic microdata files can be quite high. For this reason, the Census Bureau releases very few economic microdata files, and those that are released have extremely little detail.

The second form of data release is frequency count data. Tables of frequency count data present the number of units of analysis in a table cell. For example, you might have a table where the columns represent categories of race and the rows represent categories of sex, and the table cells show the counts of the number of people having the various combinations. Tables of frequency count data may relate to people or establishments. The Census Bureau does not consider frequency count data for establishments to be sensitive because so much information about an establishment, particularly classifications that would be used in frequency count tables, is publicly available. But disclosure limitation techniques are applied to tables of frequencies based on demographic data. In many cases, sampling takes care of the disclosure limitation for this type of data. The values shown in the cells are estimates based on weighted observations and pose no threat to confidentiality. But when the tables of counts are based directly on decennial census data, disclosure limitation procedures must be applied.

The third form of data release is magnitude data. Magnitude data is another form of tabular data. But instead of counts, magnitude data present the aggregate of a 'quantity of interest' over all units of analysis in the cell. The quantity of interest must measure something other than membership in the cell. For example, tables presenting the total value of shipments within the manufacturing sector by Standard Industrial Classification group by county-within-state are tables of magnitude data. Magnitude data are generally non-negative quantities reported in Census Bureau surveys or censuses of business establishments or farms. As stated previously, the distribution of these reported values is likely to be skewed, with a few entities having very large values. Disclosure limitation in this case concentrates on making sure that the published data cannot be used to estimate an individual establishment's data too closely. For magnitude data it is less likely that sampling alone will provide disclosure protection because most sample designs for economic surveys include a stratum of the larger volume entities which are selected with certainty. Thus, the units which are most visible because of their size, do not receive any protection from sampling.

3. The Disclosure Review Board

The Census Bureau has a Disclosure Review Board which has to approve data before they are publicly released. Six members on the Board, representing the economic, demographic, and decennial program areas, serve four-year terms, while there are three permanent members representing the research and policy areas. Board members use a checklist they have developed when reviewing a proposed data release. The checklist, filled out by staff members who wish to release a file, asks for information about file contents and disclosure limitation procedures applied to the file and is quite a useful tool. The Interagency Confidentiality and Data Access Group under the U.S. Office of Management and Budget has taken the Census Bureau's checklist and generalized it for use by other federal statistical agencies.

After reviewing a request, the Disclosure Review Board may approve or deny it. If the request is denied, the Board will state the reasons for denial and may offer suggestions for changes that could be made in order to obtain approval. If Census Bureau staff members are not satisfied with the Board's decision, they may appeal to a Steering Committee consisting of several Associate Directors. Thus far, there have been few appeals, and the Steering Committee has never reversed a decision made by the Board.

4. Restricted access procedures

Sometimes users need more information than the Census Bureau can publicly release without violating the confidentiality pledge. This is particularly true for users who want economic microdata. In such cases, researchers can submit a research proposal to the Census Bureau stating very specifically what research they want to conduct, what data they will need, and what type of results they will want to publish. The proposed research should benefit the Census Bureau in some way. If their proposal is accepted by the Census Bureau, they can be assigned *special sworn status* and come to the Census Bureau to work with data that cannot be publicly released due to confidentiality concerns. Researchers with special sworn status are bound by law to maintain confidentiality just like any Census Bureau employee.

In addition to the Census Bureau, which is located in Suitland, Maryland, these researchers can also access data at one of our

research data centers. These are secured sites specifically set up to allow researchers access to data that cannot be publicly released. Currently, the Bureau has research data centers in Boston, Massachusetts and Pittsburgh, Pennsylvania, and there are plans to set up two to four more in the future. Any papers published as a result of this type of research are carefully reviewed to ensure that there is no breach of confidentiality.

5. Current disclosure limitation practices

Current disclosure limitation practices for microdata

The only firm rule for public use microdata files is that all identified geographic areas must contain at least 100,000 persons in the sampling frame. For one survey, the Survey of Income and Program Participation (SIPP), the geographic criterion is 250,000 persons because it is longitudinal and because there is an extremely large amount of very detailed information on the file. Variables are recoded when categories are sparse, and some continuous variables like real estate taxes are put into categories if those values are publicly available elsewhere. Topcoding and bottomcoding are used to protect outliers in the ends of our distributions. For example, a record would not show a wage income of $2,000,000 or an age of 103. Instead, the record would show that the wage income was more than $150,000 dollars and the age was greater than 90. Often the medians or means of all topcoded values are shown. When topcoding, the Census Bureau typically follows what is called the half percent or three percent rule. For variables where almost everyone in the population has a non-zero value, the topcode is set at the top one half of one percent of all values. For variables where many people in the population have a zero value, the topcode is set at the top three percent of all non-zero values.
Survey design variables, weights, and contextual variables are not shown if publishing them could allow someone to identify a geographic area with fewer than 100,000 persons. These potential problems are addressed in the previously described Disclosure Review Board checklist. When reviewing a file, the Board considers any potential outside files with identifiers that could be matched to the file and must feel that such a linkage could not be made.
Noise addition was used once to protect a file. The Census Bureau released a file that linked its March 1991 Current Population Survey (CPS) data with income-related data obtained from the Internal Revenue Service (IRS). The Bureau had to ensure that the IRS could not take this linked file that the Bureau released and use the IRS information on that file to match back to IRS data and then start attaching names onto the Bureau's linked microdata file. The technique used to protect this file is becoming known as the Kim-Winkler approach to data masking. Briefly, noise was added to the data using a method developed by Jay Kim which preserves means and correlations. Then record linkage software developed by Bill Winkler was used to see if any records containing noise could be matched back to their noise-free counterparts. Most could not. For those that could, a sample of them were swapped. This means that for some pairs of records that matched on certain demographic characteristics, quantitative data were exchanged. It was judged that this technique sufficiently masked the IRS data to prevent linkages while preserving data utility.

Current disclosure limitation practices for frequency count data

As previously stated, the Census Bureau does not consider frequency count data for establishments to be sensitive because so much information about an establishment, particularly classifications that would be used in frequency count tables, is publicly available. But disclosure limitation techniques may be needed for tables of frequencies based on demographic data. For tables of frequency counts from demographic surveys, sampling takes care of the disclosure limitation. The values shown in the cells are estimates based on weighted observations. The tables are not at all sparse and pose no threat to confidentiality. But when the tables of counts are based directly on decennial census data, disclosure limitation procedures must be applied.
For tables generated from Census 2000 data, the disclosure limitation technique will be data swapping. This technique will be applied to both the short form (100%) and the long form (sample) data items. In each case, a small percent of household records will be swapped. Data swapping requires that a sample of pairs of household records that match on a cross tabulation of certain (key) variables but are in different geographic areas will be swapped across those geographic areas. All tables to be publicly released will be generated from the data after the swapping procedure has been applied.
In addition to the data swapping, there will be some restrictions on the tables themselves in order to protect confidentiality. Tables should not be too sparse, and tables from the long form (sample) data will be weighted.

Current disclosure limitation practices for magnitude data

Currently, the Census Bureau uses cell suppression for disclosure limitation of tables of magnitude data. This means that any table cell showing a value which could allow users to estimate a respondent's value too closely is suppressed (the value is not shown). Such a table cell is called a sensitive cell, or a primary suppression. The Census Bureau uses the p-percent rule to identify sensitive cells (Federal Committee on Statistical Methodology, 1994). This rule is designed to ensure that a user cannot estimate a respondent's value to within p% of the value.
Once sensitive cells have been identified, other cells called complementary suppressions, or secondary suppressions, are selected and suppressed so that the sensitive cells cannot be derived or estimated too closely by addition and subtraction from the published values. The Census Bureau uses software developed by Bob Jewett (Jewett, 1993) for choosing complementary suppressions. This software is based on network flow algorithms and is geared mainly at protecting two dimensional tables with a hierarchical structure on one of those dimensions. An auditing program based on linear programming techniques is applied to three dimensional tables following cell suppression to identify sensitive cells that did not receive adequate protection, and additional complementary suppressions are applied by hand if necessary.

6. Current disclosure limitation research

The Confidentiality Staff in the Statistical Research Division perform research into statistical disclosure limitation methods.

Current disclosure limitation pesearch for microdata

Recent research includes an analysis and an extension of the Kim-Winkler approach to masking microdata (Moore, 1996b). Moore concentrated on the development of a standard procedure for determining an acceptable amount of noise, and on the development of a standard for the maximum percentage of re-identifiable 'high risk' records in a publicly released file. A second paper by Moore (1996a) describes rank-based proximity swapping and discusses its effect on data analysis. The procedure involves sorting the values in each continuous field and swapping pairs of values so that the ranks of exchanged values differ by less than a prescribed amount. Moore's research in this area focused on deriving the 'prescribed swapping difference' for each continuous variable subject to certain constraints on maintaining data quality and on providing sufficient protection.

Staff also recently reviewed the μ-ARGUS software. Currently this software is not used at the Census Bureau, but it may prove useful in the future, particularly as an easy way to discover which categories of which variables pose a high disclosure risk.
Staff are currently investigating the existence of outside microdata files which could potentially be linked to the American Housing Survey (AHS) microdata file. If such files exist and the Census Bureau can obtain them, staff will perform a reidentification study using record linkage software to see whether indeed the files can indeed be successfully matched. If such files exist but cannot be obtained by the Census Bureau, staff will attempt to quantify the disclosure risk in terms of an estimate of the number of records on the AHS file that represent households that are 'population uniques.' That is, staff will identify the variables common to both files. Then they will estimate the number of records that represent households that have a unique combination of these variables when compared with all other households in the sampling frame.

Current disclosure limitation research for frequency count data

Current research focuses on the effects of data swapping on data from Census 2000. Staff are investigating the effects of swapping on variances of estimates for different swapping rates, or percentages.
Staff also recently reviewed the τ-ARGUS software. This software will probably not be used at the Census Bureau for count data because the Bureau wishes to do disclosure limitation for this type of data at the microdata level, rather than on a table by table basis.

Current disclosure limitation research for magnitude data

Staff recently reviewed software developed at Statistics Netherlands.
τ-ARGUS focuses on limiting the disclosure risk of magnitude data through cell suppression. This software is not used at present at the Census Bureau because the Bureau publishes an extremely large number of interrelated, multidimensional tables that τ-ARGUS cannot currently process. However, future versions of τ-ARGUS may be able to handle this type of data, and the use of this software would certainly be considered.

As discussed above, the Census Bureau uses cell suppression for the protection of magnitude data. But cell suppression does present some problems. It limits the amount of information available to data users. In fact, users often complain that too much information is suppressed. In addition to limiting the amount of information provided, using cell suppression also necessitates coordinating suppression patterns between interrelated tables, which can be a very complicated and difficult process. In particular, it makes the production of special requests for tabulations which often follow the main publication extremely tedious. Special tabulations are becoming the rule rather than the exception, making coordinating suppression patterns among all requested tables virtually impossible.
As an alternative to cell suppression, confidentiality staff are currently investigating the addition of noise to establishment microdata prior to tabulation as a disclosure limitation technique for magnitude data. Specifically, for this noise approach, each respondent establishment's data is perturbed by a small amount, say 10%. Again, the noise is applied at the microdata level before the tables are created.
Then, upon tabulation, if a single establishment dominates a cell, the value in the cell will not be a close approximation to the dominant establishment's value because that value has had noise added to it. By adding noise, one avoids disclosing the dominant establishment's true value.
Noise is added to an establishment's data by means of a multiplier. For the 10% example, the multiplier would be near either 0.9 or 1.1 and would be applied to all the establishment's data items prior to tabulation. The parameters used, such as 0.9 and 1.1, would remain confidential within the Census Bureau. Because the same multiplier would be used with an establishment wherever that establishment was tabulated, values would be consistent from one table to another.
That is, if the same cell appeared on more than one table, it would have the same value on all tables. A variety of distributions could be used to generate the multipliers, provided that they were centered at or near 1.1 and 0.9. It is a key requirement, however, that the overall distribution be symmetric about 1. That is, the distribution around 0.9 and the one around 1.1 should be 'mirror images' of each other. In this case, the expected value of any multiplier will be 1, even though in practice no multiplier will ever actually equal 1. Hence the expected value of the *amount* of noise in any establishment will be zero. This prevents the noise from introducing any bias into our level estimates.
In the degenerate case, where a cell contains only a single establishment, the cell value would contain about 10% noise. Other sensitive cells, in which one large establishment dominates the cell value, would also contain large amounts of noise because the amount of noise in the cell total would resemble the amount of noise in the dominant establishment (roughly 10%). The more dominant the large establishment is, the more closely the cell resembles the single-contributor case. These are precisely the cells that are at risk for disclosure and need to be protected. In general, the amount of protection provided to an estimate by the noise would depend on the amount needed. This property, combined with the fact that noise only has to be added once, greatly simplifies the production of special tabulations. An agency could produce as many tabulations as necessary, and for each one the noise would naturally end up being greater in the sensitive cells. There is no longer a need to keep track of suppressed cells between tables. For surveys, the overall percent change to a cell value can be reflected in the variance of the value, thus users can still perform legitimate analyses of the data. This looks like a promising alternative to cell suppression for magnitude data.

7. Conclusion

As described above, a number of different techniques are used to protect the confidentiality of Census Bureau data. Recently at the Census Bureau, there has been an increase in concern about the disclosure risk of data because of the amount of data available that could be matched to ours, the excellent record linkage and linear programming software available that can be used in attempts to thwart disclosure limitation techniques, the DADS, and increased computer power, memory, and speed. In the future, the Bureau will probably make more use of noise addition and data swapping. Many other statistical agencies in the U.S. are doing just that. Also, the Census Bureau will probably make more use of record linkage techniques to analyze the disclosure risk of microdata files and show us where more protection is needed.

References

Evans, B.T., L. Zayatz and J. Slanta, J. 1996. Using Noise for Disclosure Limitation of Establishment Tabular Data, In: *Proceedings of the 1996 Annual Research Conference.* U.S. Bureau of the Census, pp.65-86.

Federal Committee on Statistical Methodology. 1994. *Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology*, Washington, DC: U.S. Office of Management and Budget.

Jewett, R. 1993. *Disclosure Analysis for the 1992 Economic Census.* Unpublished manuscript. Economic Statistical Methods and Programming Division, Bureau of the Census, Washington, DC.

Moore, R. A. 1996a. *Controlled Data-Swapping Techniques for Masking Public Use Microdata Sets.* Statistical Research Division Report Series, RR 96-04. U.S. Bureau of the Census, Washington, DC.

Moore, R. A. 1996b. *Analysis of the Kim-Winkler Algorithm for Masking Microdata Files —- How Much Masking Is Necessary and Sufficient? Conjectures for the Development of a Controllable Algorithm.* Statistical Research Division Report Series, RR 95-05. U.S. Bureau of the Census, Washington, DC.

Note

1) All three authors work at the United States Bureau of the Census. This article reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

Protecting the confidentiality of Eurostat statistical outputs

David Thorogood[1)]

1. The role of Eurostat

Although Eurostat performs many of the functions which would normally be associated with a National Statistical Institute (NSI), there are important differences in role which impact on the treatment of statistical confidentiality. It would be fair to say that Eurostat is a special case with differing roles and responsibilities to those of other statistical institutes. Eurostat itself does not generally become involved in the collection of data, being rather a processor and disseminator of data collected in the main by NSIs and related agencies.

Eurostat has a duty to undertake and to foster research at the pan-European level on a wide range of issues relevant to the development of the European Union. This can include topics as diverse as industrial innovation and social exclusion. Certain patterns and processes are common to all member states, and can most profitably be studied at the European level. To facilitate this, it is important that Eurostat has access to data which would be considered too confidential for general release. For example, table cells which may be confidential at the level of individual member states may be publishable when aggregated with the same cells from other member states.

There are particular difficulties where data are published at both the national and European level. For example, Eurostat does not normally have any control over what is released at the member state level by the NSIs themselves. Frequently data are prepared for release at the European level some time after their release by member states. In a typical cell suppression problem, one seeks to suppress the minimum number of cells compatible with maintaining respondent confidentiality. However, when publishing European level tables, one finds that certain cells which would ideally be used for secondary suppression can be directly derived from tables already published by member states. This limits the choice of cells available for secondary suppression, and may mean that a sub-optimal suppression solution is necessary.

Eurostat has demonstrated a long-term commitment to both the practical application of statistical confidentiality measures and to supporting research into confidentiality and statistical disclosure control. A series of biennial international seminars has been held to encourage the exchange of information developments and best practice in statistical methodology, computing, and legislation relating to statistical confidentiality. Work is on going with the United Nations Economic Commission for Europe to develop joint work sessions which will concentrate on the practical application of statistical confidentiality measures. Eurostat has also been involved in sponsoring research and development into statistical disclosure control, including funding the further development of the ARGUS computer systems under the DOSIS programme.

2. Policy and legislation

Although removed from the data collection process, as a trusted receiver of - often confidential - data from NSIs, Eurostat is bound to uphold the confidentiality promises made by the NSIs. This is required both by European legislation, and by the necessity of maintaining a good working relationship based on trust and mutual co-operation between Eurostat and the NSIs. It is important therefore for Eurostat to work closely with the NSIs and other international institutions to address issues of statistical confidentiality and issues relating to the release of specific datasets or types of output. Two key pieces of European legislation govern the treatment of confidential data in Eurostat. Council Regulation 1588/90 and more recently Council Regulation 322/97 (the Council Regulation on Community Statistics, widely referred to as the *Statistical Law*) define the duties of Eurostat and the NSIs in terms of the protection of European statistical data.

Prior to the implementation of Regulation 1588/90, certain member states had been restricted by national legislation from passing confidential data to Eurostat. Under the regulation, reasons of statistical confidentiality may not be used to prevent transmission of data to Eurostat. Following the established principle of separating the administrative use of data from the statistical use, although confidential data are passed to Eurostat, access may not be allowed to other parts of the European Commission. NSIs are not compelled, however, to transmit to Eurostat data which could directly identify respondents. Therefore, it is normal for details such as names and addresses to be removed by the NSI before data are sent to Eurostat.

Council Regulation 1588/90 also established the Committee on Statistical Confidentiality, which serves as a forum for discussions between Eurostat and the member states regarding statistical confidentiality. This Committee has a designated role to act as the link by which the statistical confidentiality policies to be applied by Eurostat can be discussed and commented upon by the NSIs. Proposals regarding the treatment of confidential data are submitted to this Committee for its opinion. The Regulation makes provision for decisions to be reached by a majority vote. However, the aim of Eurostat would be for any such decisions to be reached by general agreement rather than by a majority vote.

Although Regulation 1588/90 allowed for confidential statistical data to be passed to Eurostat, the decision about whether the data were confidential remained with the member state concerned. Council Regulation 322/97 – the Statistical Law - envisages the agreement of common EU-wide definitions of confidentiality based on the identifiability of statistical units. Article 13 of the Statistical Law states that ‘To determine whether a statistical unit is identifiable, account shall be taken of all the means that might reasonably be used by a third party to identify the said statistical unit’.

Regulation 322/97 also makes special provision for the transmission of confidential data to third parties for scientific use, providing that the confidentiality of the respondent is protected and that the member states give their permission for this use. This is significant because it has opened the prospect for research to be undertaken by external experts in particular fields, who would not otherwise be able to use individual data.

Policies and working procedures have been put in place at Eurostat to implement these pieces of legislation. These procedures, which cover details of the transmission, storage, processing and dissemination of data, have been agreed with member states through the Committee on Statistical Confidentiality. Where appropriate, more detailed discussions have also taken place at topic-specific working groups and task forces. Task forces of experts from the member states and Eurostat have addressed a number of issues relating to the treatment of confidential data. The work of these task forces is continuing, and will help to ensure that, where necessary, procedures first developed under Regulation 1588/90 can be updated in line with the Statistical Law.

3. Case study: The first Community Innovation Survey

An important application of microaggregation techniques has been on the first Community Innovation Survey (CIS) (European Commission, 1994). This is a major study of innovation in European enterprises developed by Eurostat together with the OECD and

independent experts. The first CIS took place in all EU member states as well as Norway during 1992-3 and addressed issues such as expenditure on innovation activities, information sources, the transfer of technology, and factors restricting or promoting innovation. The results of this study have been widely used: the European Commission's Action Plan for Innovation was partly based on information from the study.

Innovation is seen as one of the prime indicators of likely future industrial success, and is therefore an aspect of the economies of the member states and the European Union as a whole which requires extensive study. In conflict with this is the fact that the scale and direction of investment in innovation is frequently one of the more sensitive of industrial secrets. This is particularly the case in high technology or knowledge based industries, which are often the main subjects of innovation research. It was important therefore that enterprises participating in the CIS were reassured that the information provided for the survey would remain confidential. This provides an interesting example of the conflicting needs of researchers and data suppliers.

In common with normal practice, Eurostat was not directly involved in the data collection process itself. Data were collected by NSIs and/or other research organisations working within the participating countries. A total of over 40,000 anonymised records were passed to Eurostat. Although Eurostat acted as the central point at which the data were combined to form a unified dataset, final permission over access to the data remained in the hands of the NSIs and other data collection agencies.

An additional problem here, and one that is common to many studies of enterprises, is that the relatively small number of enterprises in certain sectors, coupled with the great variation in enterprise size, means that particular enterprises are often extremely identifiable. Even though all directly identifying information is removed from a record, an enterprise will frequently be unusual because of, say, its size and sector of operation. Enterprises are often particularly concerned about keeping sensitive information away from their direct competitors, yet it is these competitors who are most likely to have the necessary knowledge of that sector of the economy to be able to disclose the information.

It was also important that the methods of dissemination chosen for the CIS allowed research to be undertaken by external researchers working outside of Eurostat and the member state NSIs. This has helped to maximise the use that was made of the data, ensuring that those working on the study included experts in the field of industrial innovation. It was decided that an appropriate approach would be to microaggregate the data using the methods outlined below. It was also deemed necessary to control the access to the microaggregated data. Although the data were perceived to be safe for release from the premises of Eurostat, they were only available to bona fide researchers who were allowed to access them under contract.

Microaggregation methods adopted for the CIS

Rather than being a single method, microaggregation is a group of related microdata disclosure control methods based on the creation of what is, in practice, artificial data representative of the original data.

The application of microaggregation is discussed more detail by Defays and Nanopoulos (1992) and by Anwar (1993).

As the data collected for the CIS consisted of quantitative (or metric), ordinal, and nominal variables, it was necessary to call upon three different methods of microaggregation to protect all of the dataset variables.

The similarities between these methods are that they each form groups or microaggregates of k cases (k normally being equal to 3), and replace the individual data items with a value calculated as the average of the group values. The differences between the methods are ways in which the groups are formed and the type of average calculated.

Quantitative variables

A method of individual ranking was implemented for quantitative variables. Cases were ranked in ascending order by the quantitative variable, and were placed into groups of three neighbouring cases. The individual observations were then replaced by the arithmetic mean for that group of three. Where the number of cases was not a multiple of three, the final group could consist of 4 or 5 cases.

The example below shows this for the quantitative CIS variable 'Turnover'.

Case no.	Turnover	Group	Group mean (microaggregated turnover)
1	2	1	3
2	3	1	3
3	4	1	3
4	5	2	7
5	7	2	7
6	10	2	7
7	12	3	23
8	17	3	23
9	22	3	23
10	24	3	23
11	40	3	23

This process was repeated for each quantitative variable individually. Therefore, each individual case was normally grouped in different clusters for each variable.

A strength of this method is that the greatest changes are made to those records which are most readily identifiable. Whereas common values receive only minimal change, significant changes are made to outlying values. It may be argued that it is these outliers - possibly signifying enterprises with unusual or effective innovation policies - which are most worthy of study. However, it is a problem common to all disclosure control problems that it is the unusual which requires the greatest protection and which is therefore not available for analysis.

Ordinal variables

The method of individual ranking 'with snake' was adopted for ordinal variables. The ordinal variables were grouped into segments - segments consisting of groups of variables which were closely linked. An example for the CIS would be the segment 'internal sources' which consists of two ordinal variables - internal sources of information within the enterprise; and internal sources of information within the enterprise group. Segments within the CIS consisted of 2 to 8 variables. Cases with neighbouring values in the segment were grouped together in threes. The order in which the enterprises were grouped was defined by the 'snake' whose arbitrary route was defined by the equation

$$v(i,j) = 6*i+j(-1)^{i+1}$$

The median of each group of three was then found and used to replace the original values for cases in that group.

Nominal variables

A relatively simple method was used here to group similar cases. In the example below, the responses for six related variables (a

Case no.	Var A	Var B	Var C	Var D	Var E	Var F	Group
1	1	1	1	1	1	1	1
2	1	1	0	1	1	1	1
3	1	1	0	1	1	1	1
4	1	0	0	1	1	1	2
5	1	0	0	0	1	1	2
6	1	0	0	0	1	1	2

segment) are considered together. Case 1 is selected and all other cases are assessed for their degree of difference from case 1. Cases are then sorted in order of their similarity to case 1. As can be seen from the example table below cases 2 and 3 differ by only one variable from case 1, whereas cases further down the list are more different - cases 5 and 6 differ from 1 by 3 variables. Cases are then divided into groups of 3 according to this sorting process, and original values for the variables are replaced by the group mode.
It will be clear that this method is highly sensitive to the order of application, with the definition of Case 1 being crucial to the resulting calculation of groups based on degree of difference from Case 1.

Evaluating the success of the disclosure control measures

The success of the disclosure control measures put in place for the CIS data can be assessed by the degree of protection conferred on the data, and the extent to which the usefulness of the data is retained. As with other disclosure control problems there is an important balance to be reached between protecting the data and minimising the effects of the disclosure control methods on the results of analyses. The CIS data have been used successfully in a number of research projects on topics such as technology transfer, patterns and costs of innovation, and the application of innovative techniques in a range of industries.
The microaggregated CIS data are additionally protected in that the data are released only to bona fide researchers working under a contract which defines what can be done with the data. This contract restricts access to named individuals working on the research, and forbids any attempt to identify individual enterprises. Data for cells with fewer than three enterprises may not be published in any results of the research, and all data to be published must be cleared by Eurostat before publication. Data may only be published in accordance with national legislation. Lastly, there are provisions to ensure the destruction of the data supplied to the researcher at the end of the project.
Such contracts, which have been used in similar circumstances by several NSIs, provide a powerful additional control over the use of the data. It is strongly in a researcher's interest to comply with the terms of the contract, as any breach could lead to prosecution, professional disgrace and the refusal to supply any future data essential to his or her work. Kooiman et al. (1999) discuss the use of data access contracts by Statistics Netherlands.
If the microaggregation process can be seen as a success in terms of the extent to which data are protected, it is also necessary to assess the extent to which the microaggregated data remain suitable for analysis. A number of studies have addressed the impact of microaggregation generally on the results of analysis (see for example Baeyens and Defays, 1998; Corsini et al., 1998; Mateo-Sanz and Domingo-Ferrer, 1998).
A study undertaken for Eurostat by Hu and De Bresson (1999) compared the results of analyses of the microaggregated CIS data with the results of identical analyses undertaken on the original data. To preserve the confidentiality of the original un-microaggregated data, the analyses on the un-microaggregated data were undertaken by Eurostat staff. The results of this study suggest that microaggregated data can be used with confidence in analyses of means, variance and percentage scores for most variables. Microaggregated data are also suitable for cluster analysis, and simple and logistic regression. There are however a few - mostly percentage - variables for which problems occur. Further research is necessary to understand these problems in more detail.

Work on CIS II

Innovation data are highly sensitive. As mentioned above, this is seen as a topic closely related to the success of both enterprises and economies as a whole. At the request of several member states, a more restrictive approach will be taken with data from the second CIS. This should not be seen as indicating that there have been problems with the implementation of the microaggregation for the first CIS. At the time of writing, exact arrangements for this are still being made, but it is possible that the microaggregated data will not be released to external researchers. An option may be for microdata analyses to be undertaken by Eurostat staff, with external researchers having access only to the results of these analyses.

4. Protecting Eurostat outputs in the future

The vast amount of data held by Eurostat represents a unique information resource on the economic and social situation of all member states of the European Union. These data have been collected at public expense, often with a considerable burden being placed on respondents. It is clearly in the interests of governments, citizens and enterprises that these data are fully exploited.
The extent to which the data held by Eurostat can be analysed is limited by the finite number of expert analytical and research staff available within the organisation. It is desirable therefore to develop ways in which external researchers may make increased use of Eurostat data. To allow maximum flexibility of analysis, it would be better if researchers could use microdata or highly detailed aggregates. However, because of the confidential nature of much of the information, such access must be tightly controlled. Data cannot simply be released from Eurostat premises in their 'raw' confidential form. Eurostat is working towards expanding access to data for external researchers, whilst of course protecting confidentiality. In all such developments, however, there are two overriding concerns. Firstly, that all reasonable care should be taken to maintain the confidentiality of those who respond to statistical enquiries or whose administrative data are used to produce statistics. Secondly, that Eurostat continues to work in close co-operation with the member state NSIs in whose name statistical data are initially collected.
In common with many statistical institutes, Eurostat has come under increasing pressure to produce more flexible statistical products. The prevalence of powerful computing technology has increased the demand for statistical outputs which can be manipulated and analysed more freely. Conversely, however, increased computing power has greatly increased the risks that confidential information may be disclosed, by facilitating techniques such as record linkage and the differencing of statistics for overlapping geographical areas. The majority of statistical outputs are still in the form of aggregate

tabular data. Although it is possible to make such products available in electronic formats, this type of output really belongs to an earlier age - that of the printed volume. Releasing this type of statistical product on, for example, a CD-ROM increases its portability and ease of access, but it does not greatly add to the types of analyses which can be performed. What is required therefore is an increased range of outputs, more closely tailored to the individual requirements of researchers. This forms part of a policy of increasing the use made of Eurostat data, helping to fulfil Eurostat's duty to foster research at the European level.

In certain cases it is likely that it will be difficult or impossible for data to be sufficiently protected to be released from Eurostat premises to external researchers. It is also probable that the level of disclosure control measures which would be necessary in such cases would seriously reduce the usefulness of the data for analysis. Eurostat has therefore proposed the development of a Data Analysis Centre (DAC) - a secure facility where external researchers may come to work on confidential data for a specific project. Such a facility allows a greater level of control over who can access the data and the types of analyses which are performed.

A development such as the DAC would of course have to take into account the requirements of national and European legislation, and would need to take place with the consent of the member states. Certain member states have expressed reservations about granting permission for the operation of the DAC. It has been agreed to look at ways of proceeding with this option on a 'case by case' basis, with permission being sought from member states for particular pieces of research, to allow a better understanding of opportunities and problems of such a facility.

Whilst the DAC might be thought of as 'data in a safe setting', for certain data it may be more appropriate to develop 'safe data' – microdata sets which can be released outside of Eurostat premises. In addition to the example of the Community Innovation Survey detailed above, Eurostat has prepared a safe microdata set based on the European Community Household Panel survey. Modifications to protect the microdata are of the types frequently made by statistical institutes when releasing microdata, including:

- limiting the geographical information;
- reducing the classification detail or limiting the upper codes for certain variables;
- removing sampling information which might be used to define geographical details.

Detailed analysis of the data was first necessary to assess the rate of occurrence of certain combinations of values, and to help target the necessary data modifications. The longitudinal nature of a panel survey creates particular disclosure control risks, as certain changes are particularly visible and increase identifiability - such as change of address, or the birth of children. As with all such Eurostat work, progress has involved detailed discussions with the NSIs who were originally involved in the data collection.

Moving beyond the 'safe data'/ 'safe setting' distinction, it is likely that distributed processing will further enhance the degree to which external researchers can have access to confidential data (Domingo-Ferrer et al., 1998). Such a system would allow the data for analysis to remain on the computer of, say, an NSI, with only the non-confidential processing instructions and results being transmitted to a remote researcher. This has particular potential benefits for research at a European level, as it should allow the creation of 'virtual-European' data composed of individual member state-level datasets held by NSIs. The concept of distributed processing is one which requires much further research.

Eurostat is continuing with research into disclosure control, as well as related issues such as distributed processing, the quantification of disclosure risks, and respondent perceptions about statistical confidentiality. The Statistical Disclosure Control (SDC) project funded through the DOSIS Programme has led to creation of updated versions of the ARGUS software which are able to protect data at both the aggregate and micro levels (Willenborg and Hundepool, 1998). The content of the SDC project illustrated the complexity and breadth of research related to statistical confidentiality and disclosure control. As well as the development of efficient methodologies and tools to protect data, it was also necessary to address how these tools should be applied. This implied work on measures of disclosure risk and a greater understanding of the motives and opportunities for a person intent on disclosing data. There clearly remains further research to be done on these and related issues.

Recently published calls for tender have included further work on disclosure control, including the development of generalised computer systems for microaggregation. It is hoped that further research work can be funded under the European Commission's Fifth Framework Programme. Eurostat also intends to continue to sponsor regular conferences and seminars to encourage the exchange of information on methodology and best practice.

References

Anwar, N.M. 1993. *Microaggregation: The small aggregates method*. Eurostat.

Baeyens, Y. and D. Defays. 1998. Estimation of variance loss following microaggregation by the individual ranking method. In: *Proceedings of the Statistical Data Protection'98 seminar, Lisbon, 25-27 March 1998.*

Corsini, V., L. Franconi, D. Pagliuca and G. Seri. 1998. An application of microaggregation methods to Italian business surveys.In: *Proceedings of the Statistical Data Protection'98 seminar, Lisbon, 25-27 March 1998.*

Defays, D. and Ph. Nanopoulos. 1992. Panels of enterprises and confidentiality. *Proceeding of the Statistics Canada Symposium 92, Design and Analysis of Longitudinal surveys.*

Domingo-Ferrer, J., R. Sanchez del Castillo and J. Castilla. 1998. Dike: a prototype for secure delegation of statistical data. In: *Proceedings of the Statistical Data Protection'98 seminar, Lisbon, 25-27 March 1998.*

European Commission. 1994. *The Community Innovation Survey. Status and perspectives.* Office for Official Publications of the European Communities. Luxembourg.

Hu, X. and C. De Bresson .1999. (forthcoming). *Internal report undertaken on behalf of Eurostat examining the effects of microaggregation on data analyses.*

Kooiman, P., J.R. Nobel and L.C.R.J. Willenborg. 1999. Statistical data protection at Statistics Netherlands. In: *Netherlands Official Statistics; this issue.*

Mateo-Sanz, J. and J. Domingo-Ferrer. 1998. A method for data-oriented multivariate microaggregation. *Proceedings of the Statistical Data Protection'98 seminar, Lisbon, 25-27 March 1998.*

Manual on Disclosure Control Methods. 1996. Office for Official Publications of the European Communities. Luxembourg

Willenborg, L.C.R.J. and A.J Hundepool. 1998. ARGUS for Statistical Disclosure Control. In: *Proceedings of the Statistical Data Protection'98 seminar, Lisbon, 25-27 March 1998.*

Note

1) Eurostat. This paper reflects current legislation and practice. The views expressed are those of the author, and should not be interpreted as official Eurostat policy.

Statistical disclosure control of Statistics Netherlands employment and earnings data

Eric Schulte Nordholt

1. Introduction

Statistical offices collect large amounts of data for statistical purposes. Understandably, respondents are only willing to provide the required information if they can be certain that the offices will treat it confidentially. Therefore, the amount of detail in publications of the data has to be limited. At the same time, there is a growing demand for more detailed information on the part of researchers: data analysis is no longer the sole resonsibility of statistical offices; as more and more researchers can perform their own analysis on personal computers, they want more than only the official tables published by statistical offices. Statistical disclosure control theory is used to solve the problem how to publish and release as much detail in these data as possible without disclosing individual information. Statistical data protection at Statistics Netherlands is discussed by Kooiman in this issue.

In the case of Statistics Netherlands' new Annual Survey on Employment and Earnings (ASEE), companies included in the survey population are obliged by law to provide data to Statistics Netherlands. This law on official statistics dates back to 1936 and was renewed in 1996 without changing the obligation for companies to respond in the ASEE. In publishing the results of the survey, no individual information may be disclosed. As the ASEE is a business survey, the law prohibits the release of microdata for research. Statistics Netherlands therefore provides two kinds of information from the ASEE: tables and a public use microdata file. Public use microdata files contain much less detailed information than microdata for research.

The data of the ASEE are described in section 2. To protect the tables produced by Statistics Netherlands from the ASEE microdata against the risk of disclosure, the software package τ-ARGUS (Hundepool et al., 1998a) is applied to three basic tables. More information about τ-ARGUS and how this package was applied in the ASEE can be found in section 3. Section 4 explains how a public use microdata file has been produced using the software package μ-ARGUS (Hundepool et al., 1998b). The current state and some possible extensions for the ARGUS packages are discussed in section 5. The software packages τ-ARGUS and μ-ARGUS emerged from the Statistical Disclosure Control (SDC) project that was carried out under the Fourth Framework of the European Union. Many ideas for the present report came from Willenborg (1993), Citteur and Willenborg (1993), Willenborg and De Waal (1996) and Groot and Citteur (1997).

2. The ASEE data

The Division of Socio-economic Statistics at Statistics Netherlands recently started a large new Annual Survey on Employment and Earnings (ASEE). Most data are no longer collected by paper forms but by EDI (Electronic Data Interchange). At the moment the percentage of companies responding electronically is still modest, but it is increasing rapidly. Moreover, as it is mainly the larger companies who are quick to switch to EDI, the number of employee records sent electronically to Statistics Netherlands is quite substantial. In principle, companies using EDI no longer send data on only samples of employees, but tapes with all the employee records, so that that much more information is available. More information about the changes in the data collection process of the ASEE can be found in Arnoldus (1997). So we now have much more earnings information than before. For 1995, the first reference year of the survey, the ASEE data set contains approximately one and a half million records with detailed earnings information. The challenge is to expand this number of records within the next few years to cover all six million employees in the Netherlands.

3. The release of tables with τ-ARGUS

Very many tables are produced on the basis of the ASEE. As they have to be protected against the risk of disclosure, the software package τ-ARGUS (Hundepool et al., 1998a) is applied. Two common strategies to protect against the risk of disclosure are table redesign and suppressing individual values. It is necessary to suppress cell values in the tables because publication of (good approximations of) these values may lead to disclosure. These suppressions are called primary suppressions. A dominance rule is used to decide which cells have to be suppressed. This rule states that a cell is unsafe for publication if the n major contributors to that cell are responsible for at least p percent of the total cell value. The idea behind this rule is that in unsafe cells the major contributors can determine with great precision the contribution of their competitors. In τ-ARGUS the default value for n is 3 and the default value for p is 70 %, but these values can be changed easily if the user of the package prefers other values. Using the chosen dominance rule τ-ARGUS shows the user which cells are unsafe. In publications unsafe cell values are normally replaced by crosses (×).

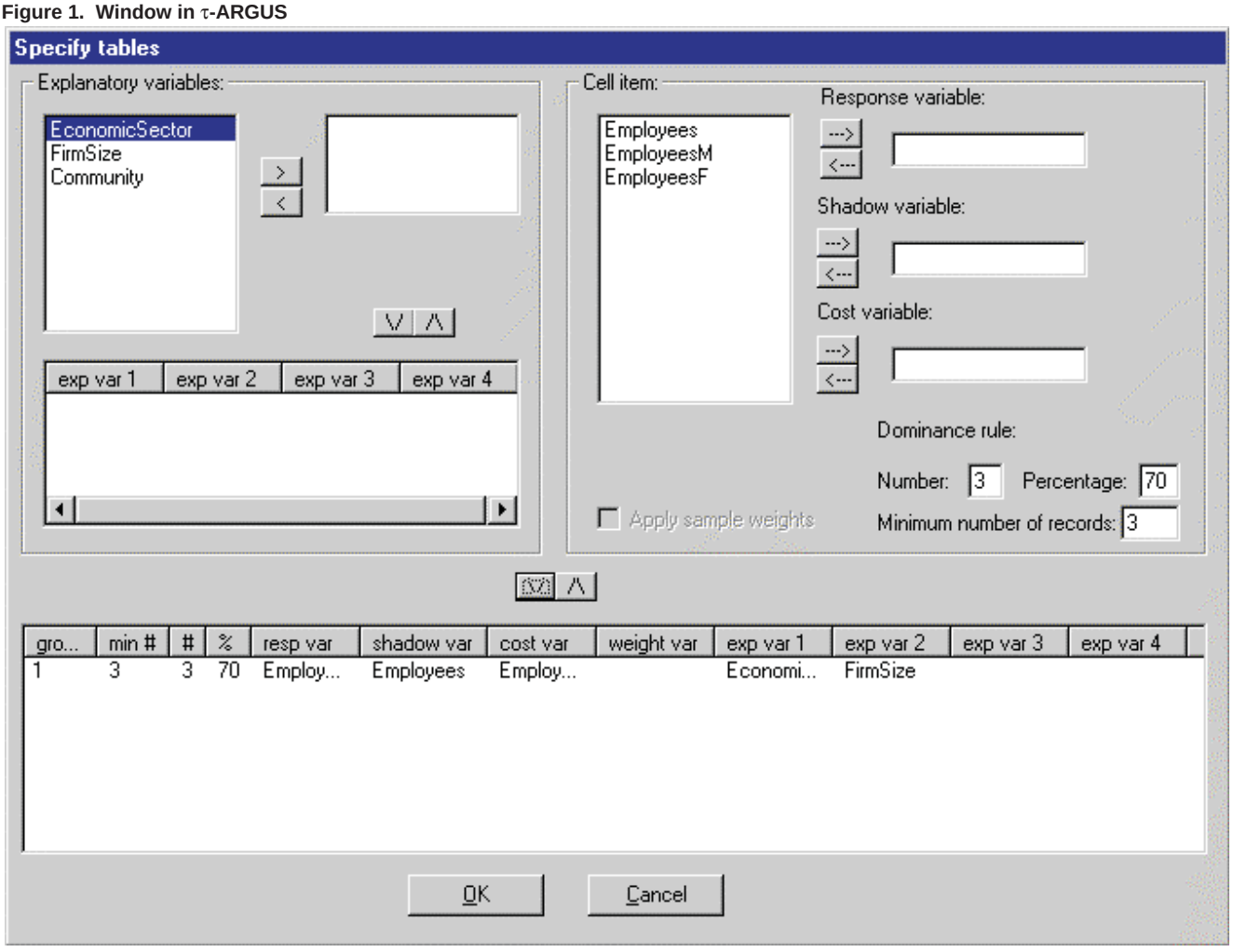
As the marginal totals are given in addition to the cell values, it is necessary to suppress more cells, as otherwise the suppressed cell values may be recomputed using the marginal totals. Even if it is not possible to recalculate the suppressed cell exactly, it is often possible to calculate a sufficiently small interval that contains the suppressed cell value. In practical situations like the ASEE employee tables, for instance, every cell value is non-negative and therefore cannot exceed the marginal totals in the row or column. If the size of such an interval is small, then the suppressed cell can be estimated quite accurately, which is of course undesirable. Therefore, it is necessary to suppress additional cells, making the intervals sufficiently large. The user has to indicate how large an interval should be. This interval is called the safety range; in τ-ARGUS the default safety range has a lower bound of 70 % and an upper bound of 130 % of the cell value. Users can also change these default values in τ-ARGUS. All extra suppressions are called secondary suppressions. Users of the tables cannot tell if a suppression is primary or secondary: normally all suppressed cells are indicated by crosses (×). Not showing why a cell has been suppressed helps preventing the disclosure of information.

Preferably the secondary suppressions will be executed in an optimal way. Just what constitutes 'optimal' is quite interesting in this respect. The chosen definition of 'optimal' is often the minimum number of secondary suppressions. Other possibilities are to minimise the total of the suppressed values or the total number of individual contributions to the suppressed cells. Minimisation of the total of the suppressed values is of course only relevant if all cell values are non-negative (as in the ASEE employee tables). In τ-ARGUS the option of minimising the total of the suppressed values has been implemented as the default. In τ-ARGUS 2.0 it is also possible to minimise the total number of individual contributions to the suppressed cells. If that criterion is desired, a so-called cost variable that is equal to 1 for every record has to be used to execute the secondary suppressions in τ-ARGUS 2.0. However, the option of minimising the number of secondary suppressions itself is not yet

implemented. Future versions of τ-ARGUS will aim to implement more options so that the different resulting groups of secondary suppressions can be compared.
If the process of secondary suppressions is directly executed on the most detailed tables available, large numbers of local suppressions will often result. Therefore it is better to try to combine categories of the spanning (explanatory) variables. This table redesign by collapsing strata leads to fewer rows or columns. If two safe cells are combined a safe cell will result. If two cells are combined of which at least one is not safe, it is impossible to say beforehand whether the resulting cell will be safe or unsafe, but this can easily be checked afterwards by τ-ARGUS. However, the remaining cells with larger numbers of companies tend to protect the individual information better, implying that the percentage of unsafe cells tends to diminish by collapsing strata. Thus a practical strategy to protect the ASEE employee tables is to start by combining rows or columns in the most detailed tables available. This can be executed easily within τ-ARGUS. Small changes in the spanning variables can most easily be executed by editing manually in the recode box of τ-ARGUS, while large changes can be handled more efficiently in an externally produced recode file which can be imported into τ-ARGUS without any problem. Once this redesign process is completed, the local suppressions can be executed with τ-ARGUS given the parameters for n, p and the lower and upper bound of the safety range.
To give an idea of what the current version of τ-ARGUS looks like, a window of this package is shown in Figure 1, in which the table *number of employees by economic sector and company size* is ready for processing using the default values for the parameters. As publishing the used parameters for the data protection could facilitate disclosure of this information, the actually used parameter values for this basic table are kept secret.

As many tables are produced on the basis of the ASEE microdata and the software package used for the data protection is based on individual tables, there is a danger that although all individual tables are safe, combining tables may result in unsafe cells which disclose individual information. This may occur for tables that have spanning and response variables in common.
Although τ-ARGUS has an option to protect linked tables, the current version cannot guarantee protection. However, the aim is to extend τ-ARGUS in such a way that it is able to deal with an important sub-class of linked tables, namely hierarchical tables. Intuitively, a hierarchical table is an ordinary table with its marginals, but with additional subtotals.
The problem is therefore how to approach linked tables now, with the current version of τ-ARGUS. As all tables have to be protected against the risk of disclosure, the current version of τ-ARGUS is applied to three basic tables. This is much fewer than the total number of tables that are published. Many specific tables can be constructed from the protected basic tables and will thus automatically be safe as well. What remains is how to protect the different basic tables simultaneously. As the current version cannot guarantee optimal suppression for two or more tables simultaneously, we had to find some kind of practical protection strategy.
In the ASEE the following three basic employee tables had to be protected: number of employees by economic sector and gender, number of employees by economic sector and region and number of employees by economic sector and company size. The first basic table is constructed by combining economic sectors in such a way that no cell suppressions are necessary. The other two basic tables are protected by first calculating all primary suppressions, then adding to this number by suppressing cells that would otherwise

Figure 1. Window in τ-ARGUS



break the protection of the prior protected basic table. Lastly, safe specific tables are derived from the safe basic tables. This strategy does not always lead to the optimum data protection strategy in the sense that the secondary suppressions are executed in an optimal way. However, it seems to be a reasonable approach that can be executed without too much trouble and leads to tables with only safe cells.

In practice two complications make our data protection process of linked tables a bit more difficult. Firstly, not only cell values and totals are published, but also many subtotals. Therefore the process must be executed at the level of the basic subtable. Secondly, if there is a choice of where to place a secondary suppression cross it is considered to be superior practice to place it in a cell that was also suppressed in the previous year. Otherwise, while for each year a basic subtable may be safe, combining such tables for consecutive years could lead to disclosure of individual information as many cell values do not differ substantially from year to year and also the main contributors to these cells are often the same. Thus good estimates can be made for suppressed cell values if the same cell is not suppressed the in the previous or subsequent year.

Below is an example of the results of the applied data protection strategy to the basic subtables of the ASEE. Table 1 lists codes and names of the economic sectors according to NACE 1. These economic sectors are the row categories in Tables 2a and 2b where the safe table 'number of employees by economic sector and province' can be found. Tables 2a and 2b form one table but are split into two parts because of graphical reasons. This table is derived from the basic table *number of employees by economic sector and region*. In this safe table six crosses are placed to indicate the suppressions. The crosses that represent secondary suppressions are placed in such a way that the suppressed primary cell values cannot be recomputed 'accurately' using the marginal totals. 'Accurately' is defined here as within the chosen - but not published - safety range.

The table *number of employees by economic sector group and province* is another table from *number of employees by economic sector and region*. The variable *economic sector group* is derived from the variable *economic sector* by collapsing the 14 economic sectors into four groups (01-05, 10-45, 50-74 and 75-93). In the table *number of employees by economic sector group and province* no cell suppressions were necessary. Therefore, the secondary suppressions in Tables 2a and 2b are placed in such a way that the table *number of employees by economic sector group and province* will not help to disclose these suppressions. As in Tables 2a and 2b suppressions are found only in economic sectors of the economic sector group 10-45, we actually applied the data protection strategy to the subtable *number of employees by economic sector group 10-45 and province*.

4. The release of a public use microdata file with μ-ARGUS

Many users of the ASEE are satisfied with the released safe tables of Statistics Netherlands. However, some users need more information. As the law does not allow the use of microdata from business surveys for research, a public use microdata file was produced with the software package μ-ARGUS (Hundepool et al., 1998b). Two criteria have to be fulfilled for this public use microdata file: every category of an identifying variable needs to occur frequently enough and every bivariate value combination of two identifying variables needs to occur frequently enough. These criteria depend on the number of individuals in the population with those characteristics. A common threshold value for the number of individuals in the population in a category of an identifying variable is 200,000; a common threshold value for the number of individuals in the population in a bivariate value combination of two identifying variables is 1,000.

One of the most important users of the ASEE is the National Bureau for Economic Policy Analysis (CPB). One of its tasks is to calculate the effects of proposed policy decisions. The CPB needs ASEE microdata to be able to produce estimates of the effects of the proposed decisions quickly.

To calculate these estimates they need only a limited number of variables in a limited number of categories. Therefore, a public use microdata file was prepared with μ-ARGUS to meet their basic needs. For special requests they sometimes need more information than is available in the public use microdata file. In such cases it is possible to work on richer microdata files on site at Statistics Netherlands. Other bona fide researchers may also make use of this provision. Like all employees of Statistics Netherlands, other people who work on-site have to swear an oath to the effect that they will not disclose individual information of respondents (Kooiman et al., this issue).

Table 1. Codes and names of the economic sectors according to the NACE 1 used in Tables 2a and 2b

Economic sector	
Code	Name
01–05	Agriculture and fishing
10–14	Mining and quarrying
15–37	Manufacturing
40–41	Electricity, gas and water supply
45	Construction
50–52	Wholesale and retail trade; repair of motor vehicles, motor cycles and personal and household goods
55	Hotels and restaurants
60–64	Transport, storage and communication
65–67	Financial intermediation
70–74	Real estate, renting and business activities
75	Public administration and defence; compulsory social security
80	Education
85	Health and social work
90–93	Other community, social and personal service activities
01–93	All activities

Table 2a. First segment of a table with number of employees (x 1,000) by economic sector and province, 31 December 1995

Economic sector	Province					
Code	Groningen	Friesland	Drenthe	Overijssel	Flevoland	Gelderland
01–05	1.7	4.5	2.5	3.6	2.1	9.2
10–14	×	0.2	1.9	0.4	×	0.4
15–37	33.7	31.8	27.9	78.5	8.1	115.6
40–41	×	1.5	0.8	3.0	×	4.6
45	10.8	12.8	9.8	27.0	3.8	40.8
50–52	25.8	28.1	22.0	58.8	14.9	109.3
55	4.2	5.3	4.2	10.3	2.1	18.3
60–64	12.5	9.3	5.7	19.6	2.8	32.3
65–67	3.6	8.2	3.0	7.4	1.3	20.1
70–74	28.5	22.8	17.6	44.8	12.0	84.5
75	13.8	12.5	10.7	20.8	6.1	38.9
80	19.0	13.7	8.0	28.3	5.7	45.7
85	30.2	28.2	21.5	48.8	8.0	91.4
90–93	6.3	7.0	5.0	11.6	2.7	23.7
01–93	193.5	185.8	140.5	362.8	69.9	634.8

Source: ASEE, 1995.

Table 2b Second segment of a table with number of employees (x 1,000) by economic sector and province, 31 December 1995

Economic sector	Province						The Netherlands
Code	Utrecht	North Holland	South Holland	Zeeland	North Brabant	Limburg	
01–05	3.1	11.3	26.1	1.9	11.8	6.5	84.3
10–14	x	1.2	2.2	0.0	0.5	0.7	9.2
15–37	46.3	118.1	143.6	23.2	190.8	92.9	910.5
40–41	x	7.4	8.8	1.6	5.4	3.9	42.3
45	26.4	47.1	78.5	7.5	56.1	20.9	341.4
50–52	85.6	178.1	217.4	19.4	148.7	57.4	965.5
55	14.1	39.3	33.0	5.0	25.1	14.1	175.0
60–64	30.3	90.1	100.3	7.2	43.4	21.0	374.4
65–67	24.7	52.3	47.8	2.8	22.9	10.7	204.7
70–74	76.3	155.8	191.6	12.2	103.7	45.1	794.9
75	27.4	65.2	106.3	8.7	40.2	22.3	372.8
80	34.8	59.8	81.6	7.2	52.7	24.3	380.8
85	67.6	127.2	158.9	17.0	101.7	54.1	754.5
90–93	20.2	51.5	47.3	3.7	26.0	11.6	216.7
01–93	459.9	1004.4	1243.4	117.5	829.1	385.3	5627.0

Source: ASEE, 1995.

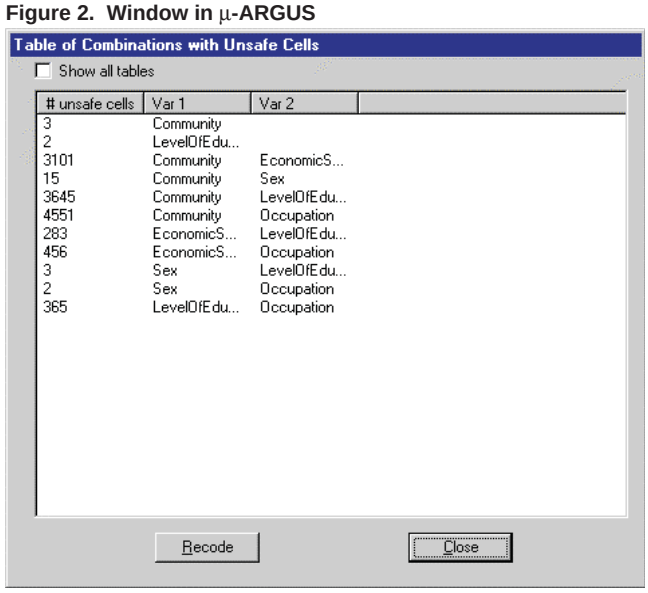
Data from the ASEE were matched with microdata from the social security files and the Labour Force Survey (LFS). This matching was executed because not only the earnings themselves, but also the structure of earnings is a main target of analysis. The social security files were taken into account to enlarge the matching probability with the LFS. The LFS happens to be the only one of these three sources that contains data on level of education and occupation. Thus, Statistics Netherlands obtained census like information without having to set up a separate and elaborate survey. Schulte Nordholt (1998) describes in more detail how the three sources were matched and how some of the remaining missing information was imputed. Actually, the public use microdata file produced contains not only information from the ASEE, but also from the social security files and from the LFS. However, the ASEE remains the main source for these earnings microdata.

The software package μ-ARGUS is used to identify and protect the unsafe combinations in the desired microdata file with information on earnings. Two data protection techniques to produce safe microdata files are global recoding and local suppression. In the case of global recoding several categories of an identifying variable are collapsed into one single category. To obtain a uniform categorisation of each identifying variable, this technique is applied to the entire data set, not only to the unsafe part. Examples of identifying variables in this data set are sex, economic sector and level of education. If an identifying variable is really desired in many categories, this implies that other identifying variables are allowed fewer categories. Ideally, all identifying variables have so few categories that no more unsafe combinations in the microdata exist and local suppressions are not necessary. When local suppression is applied, one or more values in an unsafe combination are suppressed, i.e. replaced by a missing value. These missing values could be imputed, but this is not usually done as bad imputations give misleading information to users and good imputations could lead to disclosure of individual information of respondents. Local

suppressions thus limit the analysis possibilities as no longer rectangular data files to analyse result. However, in the practice of producing a public use microdata file it is hard to limit the level of detail in the identifying variables and one will often need some local suppressions to meet the data protection criteria for a public use microdata file. Therefore, after recoding the identifying variables interactively with μ-ARGUS, the remaining unsafe combinations had to be protected by suppressing some values. The software package μ-ARGUS determined the necessary local suppressions automatically and optimally, i.e. the number of values that have to be suppressed is minimised. That way it was possible to quickly produce a public use microdata file.

To give an impression of the current version of μ-ARGUS, a window of this package is shown in Figure 2 in which the *Table of Combinations with Unsafe Cells* is ready to help to the user decide which variables have to be recoded. In this example it is clear that the variable *community* leads to a large number of unsafe cells and is therefore the first candidate for recoding. Small changes in the identifying variables can be executed most easily by editing manually in the recode box of μ-ARGUS, while large changes can be handled more efficiently in an externally produced recode file which can be imported into μ-ARGUS without any problem. After this global recoding the remaining unsafe combinations will be suppressed by μ-ARGUS to obtain a public use microdata file. No other public use microdata files may be produced from the same data set, otherwise the data protection measures could be circumvented by combining information. Therefore a lot of careful consideration is necessary before a public use microdata file can be released: which variables should be included in the file and how should the identifying variables in the file be recoded.

Figure 2. Window in μ-ARGUS



5. Discussion

The software packages τ-ARGUS and μ-ARGUS emerged from the Statistical Disclosure Control (SDC) project carried out under the Fourth Framework of the European Union. These software packages have proven to be a great help in the practice of statistical disclosure control. Many data protection problems of socio-economic statistics can be solved using the ARGUS packages. A few of these problems concern employment and earnings data and this article describes how these were solved.

The new manuals (Hundepool et al., 1998a and 1998b) will be of great help for the users of the ARGUS packages. However, one can always wish for more! In the case of τ-ARGUS it would be very useful if linked tables, and in particular hierarchical tables, could be dealt with more automatically. This need has been recognised for some time; in fact a preliminary implementation of a linked table option is already available in the current version of τ-ARGUS. A dedicated computer program (using the optimisation DLL of τ-ARGUS) is being developed at the department of statistical methods at Statistics Netherlands to deal with hierarchical tables. Also, more research is needed on how consecutive years of the same survey can be protected against disclosure. Finally, it would be good to have more options available to execute the secondary suppressions. In the case of μ-ARGUS it is important to distinguish more clearly between protecting microdata for research and protecting public use microdata files. As μ-ARGUS can be used with many different protection criteria, it is important to explain to users how different strategies can be executed using this package.

It can be concluded that there is still much research to be done in the field of statistical disclosure control. Hopefully, new versions of the ARGUS packages (that include results of the on-going research) will be released in next few years.

References

Arnoldus, F., 1997. Electronic supply of data for labour statistics. In: *Netherlands Official Statistics*, Vol.12, autumn 1997, pp. 60-68.

Citteur, C.A.W. and L.C.R.J. Willenborg, 1993. Public use microdata files: current practices at national statistical bureaus. In: *Journal of Official Statistics*, Vol. 9, No. 4, 1993, pp. 783-794.

Groot, A. and C.A.W. Citteur, 1997. Accessibility of business microdata. In: *Netherlands Official Statistics*, Vol. 12, winter 1997, pp. 18-32.

Hundepool, A.J., L.C.R.J. Willenborg, L. van Gemerden, A. Wessels, M. Fischetti, J.J. Salazar and A. Caprara, 1998a. *τ-ARGUS, user's manual, version 2.0.*

Hundepool, A.J., L.C.R.J. Willenborg, A. Wessels, L. van Gemerden, S. Tiourine and C.A.J. Hurkens, 1998b. *μ-ARGUS, user's manual, version 3.0.*

Kooiman, P., J.R. Nobel and L.C.R.J. Willenborg, 1999. Statistical data protection at Statistics Netherlands. In: *Netherlands Official Statistics*, Vol. 14, spring 1999.

Schulte Nordholt, E., 1998. Imputation, the alternative for surveying earning patterns. In: *Netherlands Official Statistics*, Vol. 13, spring 1998, pp. 14-15.

Willenborg, L.C.R.J., 1993. Discussion statistical disclosure limitation. In: *Journal of Official Statistics*, Vol. 9, No. 2, 1993, pp. 469-474.

Willenborg, L.C.R.J. and A.G. De Waal, 1996. *Statistical disclosure control in practice, Lecture Notes in Statistics 111,* Springer-Verlag, New York.