Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2003: Challenges in Survey Taking for the Next Decade

2003

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2003
Challenges in Survey Taking for the Next Decade

APPLICATIONS OF STATISTICAL DISCLOSURE CONTROL METHODS

Eric Schulte Nordholt[1]

ABSTRACT

National Statistical Institutes (NSIs) and Market Research Bureaus conduct surveys about many different topics. To reach this aim they have developed a fully-equipped statistical production process. It is a long way from collecting raw data to publishing public information. This paper discusses one of the methodological aspects of the statistical production process. The aim of a NSI or Market Research Bureau is always to publish as much as possible. However, the privacy of individual respondents should be respected. Therefore, statistical disclosure control techniques have been developed to protect sensitive information that can be attributed to individual respondents. Statistical Disclosure Control (SDC) is a topic that can be approached from many viewpoints. In this paper it is explained how results of European research projects on SDC can be used in the production of official statistics. Two related software packages are described that can be applied for producing safe data. The package τ-ARGUS is used for tabular data and its twin μ-ARGUS for microdata. The main techniques used to protect sensitive information are global recoding and local suppression. Both τ-ARGUS and its twin μ-ARGUS are products developed in the SDC project under the Fourth Framework Programme of the European Union. New versions (that include results of the on-going research) of both packages have been released in the CASC (Computational Aspects of Statistical Confidentiality) project that is funded under the Fifth Framework Programme of the European Union. Also other methods that allow use of data are discussed. The most important one is that of restricted access sites. Bona fide researchers who need more information have the possibility to visit Statistics Netherlands and work on-site in a secure area within Statistics Netherlands. The AMRADS (Accompanying Measures in Research And Development in Statistics) project pushes forward the transfer of SDC knowledge in many different countries.

KEYWORDS: μ-ARGUS; τ-ARGUS; Microdata; Software; Statistical Disclosure Control; Tables.

1. INTRODUCTION

National Statistical Institutes (NSIs) and Market Research Bureaus conduct surveys about many different topics. To reach this aim they have developed a fully-equipped statistical production process. It is a long way from collecting raw data to publishing public information.

The information from statistics becomes available for the public in tabular and microdata form. Historically, only tabular data were available and NSIs had a monopoly on the microdata. Since the eighties the PC revolution leaded to the end of this monopoly. Now also other users of statistics have the possibility of using microdata. These microdata can be conveyed with floppy's, CD-roms and other means. Recently also other possibilities of getting statistical information have become more popular as remote access and remote execution. With these techniques researchers can get access to data that remain in a statistical office or can execute set-ups without having the data on their own PC. For very sensitive information some NSIs have the possibility to let bona fide researchers work on-site within the premises of the NSI.

The task of statistical offices is to produce and publish statistical information about society. The data collected are ultimately released in a suitable form to policy makers, researchers and the general public for statistical purposes. The release of such information may have the undesirable effect that information on individual entities instead of on sufficiently large groups of individuals is disclosed. The question then arises how the information available can be modified in such a way that the data released can be considered statistically useful and do not jeopardize the privacy of the entities concerned. The statistical disclosure control theory is used to solve the problem of how to publish and

1 Eric Schulte Nordholt, Senior researcher, Statistics Netherlands, Division Social and Spatial Statistics, Department Support and Development, Section Research and Development, P.O. Box 4000, 2270 JM, Voorburg, The Netherlands, ESLE@CBS.NL.

release as much detail in these data as possible without disclosing individual information (Willenborg and De Waal, 1996 and 2001).

This paper is partly based on earlier work on statistical disclosure control (e.g. Schulte Nordholt, 2001). It discusses the available methods to protect sensitive information. The tables produced by statistical offices on the basis of the microdata of surveys have to be protected against the risk of disclosure. Therefore the software package τ-ARGUS (Hundepool et al, 2002a) can be applied on the tables produced. More information about τ-ARGUS and how this package can be applied is given in chapter 2. Chapter 3 explains how microdata for research and public use microdata files can be produced using the software package μ-ARGUS (Hundepool et al, 2002b). The option for bona fide researchers to work on-site at Statistics Netherlands on richer microdata files is explained in chapter 4. Also other methods that allow use of data are discussed in that chapter. Finally, in chapter 5 a discussion follows about the current state and some possible extensions for the ARGUS packages. Some conclusions are also drawn in chapter 5. Many of the ideas for this paper came from Citteur and Willenborg (1993), Groot and Citteur (1997), Willenborg (1993) and Willenborg and De Waal (1996 and 2001).

The software packages τ-ARGUS and μ-ARGUS have emerged from the Statistical Disclosure Control (SDC) project that was carried out under the Fourth Framework Programme of the European Union. The Computational Aspects of Statistical Confidentiality (CASC) project can be seen as a follow-up of the SDC project. The CASC project is funded under the Fifth Framework Programme for Research, Technological Development and Demonstration (RTD) of the European Union. It builds further on the achievements of the SDC project. On the other hand it has new objectives. It concentrates more on practical tools and research needed to develop them. In the CASC project fourteen partners from five different European countries (Germany, Italy, the Netherlands, Spain and the United Kingdom) work closely together. One of the main tasks of this consortium is to further develop the ARGUS-software which has been put in the public domain by the SDC project consortium. The CASC project involves both research and software development. As far as research is concerned the project concentrates on those areas that can be expected to result in practical solutions, which are then being built into the software. The CASC project has been designed around the software twin ARGUS. This will make the outcome of the research readily available for application in the daily practice of National Statistical Institutes and Market Research Bureaus. More information about the CASC project can be found in Hundepool (2001).

2. THE RELEASE OF TABULAR DATA

Many tables are produced on the basis of surveys. As these tables have to be protected against the risk of disclosure, the software package τ-ARGUS (Hundepool et al, 2002a) can be applied. Two common strategies to protect against the risk of disclosure are table redesign and the suppression of 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 often 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 crosses (×) normally replace unsafe cell values. Other rules that can be used to decide which cells have to be suppressed are the p-percent rule and the pq rule. The p-percent rule states that approximate disclosure of magnitude data (business data reporting non-negative quantities about certain establishments or similar entities) occurs if the user can estimate the reported value of some respondent too accurately. Such disclosure occurs, and the table cell is thus declared sensitive, if upper and lower estimates for the respondent's value are closer to the reported value than a prespecified percentage, p. In the derivation for the p-percent rule, one assumes that there was a limited prior knowledge about respondent's values. Some people believe that agencies should not make this assumption. In the pq rule, agencies can specify how much prior knowledge there is by assigning a value q, which represents how accurately respondents can estimate another respondent's value before any data are published ($p<q<100$).

The most widespread technique used to identify sensitive cells is the dominance rule. The p-percent rule can be considered as a special kind of pq rule. The pq rule is intuitively clearer and easier to extend in specific situations than the dominance rule. The pq rule can also be used if we have negative contributions or cell values in the table. When some of the contributors know approximately some of the other contributions to a cell value, this prior information can be taken into account with the pq rule. This is not the case with the dominance rule. An example of such a situation is when permission is obtained from a respondent in a sensitive cell to publish the cell. Such a waiver can be useful for publication purposes and not too demanding for a large public company where similar information is already in the public domain. The pq rule can handle waivers whereas with the dominance rule it is not clear how to continue as it should not be allowed to disclose approximately the value of another contributor to that cell. Finally, the pq rule has the advantage that both upper and lower limits are taken into account whereas when the dominance rule is used, only an upper limit can be deducted. The last mentioned disadvantage for the dominance rule also holds for the p-percent rule. In spite of these disadvantages not many countries have already experience in using other rules than the dominance rule for the identification of sensitive cells in tables. When the p-percent rule and the pq rule will become available in standard software packages for statistical disclosure control it can thus be expected that these rules will become more popular.

As marginal totals are given as well as cell values, it is necessary to suppress further cells in order to ensure that the original suppressed cell values cannot be recalculated from the marginal totals. Even if it is not possible to recalculate the suppressed cell value exactly, it is often possible to calculate it within a sufficiently small interval. In practical situations every cell value is often non-negative and thus 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 with great precision, which is of course undesirable. Therefore, it is necessary to suppress additional cells to ensure that the intervals are sufficiently large. A user has to indicate how large a sufficiently large interval should be. This interval is called the safety range and a safety range could e.g. have a lower bound of 70 % and an upper bound of 130 % of the cell value. A user of a table cannot see if a suppression is a primary or secondary suppression: normally all suppressed cells are indicated by crosses (×). Not revealing why a cell has been suppressed helps to prevent the disclosure of information.

Preferably the secondary suppressions are executed in an optimal way, however the definition of optimal is an interesting problem. Several measures for the loss of information can be defined and then the loss of information according to the measure chosen should be minimised. Four possibilities are:

- the minimisation of the number of secondary suppressions;
- the minimisation of the total of the suppressed values;
- the minimisation of the total number of individual contributions to the suppressed cells;
- the minimisation of a weighted function of scores attributed to cells that symbolise information, where empty cells get weight 0 and neighbouring cells to primary suppressions get lower weights than cells further away from primary suppressions.

Often, the minimisation of the number of secondary suppressions is considered to be optimal. Also the possibilities to minimise the total of the suppressed values or the total number of individual contributions to the suppressed cells are now and then used. The minimisation of the total of the suppressed values is of course only relevant if all cell values are non-negative. For the last mentioned minimisation one can take the hierarchy of the table into account and then software tailored to the specific needs is required. In τ-ARGUS the option of minimising the total of the suppressed values has been implemented as the default. In τ-ARGUS version 2.1 it is also possible to minimise the total number of individual contributions to the suppressed cells. If that criterion is used a so-called cost variable that is equal to 1 for every record is used to execute the secondary suppressions. Also the option of minimising the number of secondary suppressions itself has been implemented in τ-ARGUS version 2.1. This implies that with τ-ARGUS version 2.1 the three options implemented that may lead to 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. A table redesigned by collapsing strata will have a diminished number of rows or columns. If two safe

cells are combined a safe cell will result. If two cells are combined when at least one is not safe it is impossible to say beforehand if 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 enterprises tend to protect the individual information better, which implies that the percentage of unsafe cells tends to diminish by collapsing strata. Thus, a practical strategy for the protection of a table is to start by combining rows or columns. This can be executed easily within τ-ARGUS. Small changes in the spanning variables can most easily be executed by manual editing 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. After the completion of this redesign process, 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.

As normally many tables are produced on the basis of a survey and the software package used for the data protection is based on individual tables, there is the risk that although each table is safe, the combination of the data in these tables will disclose individual information. This may be the case when the tables have spanning and response variables in common. Version 2.1 of τ-ARGUS does support linked tables. An earlier version had an option to protect such tables, but this was not warranted. This implies that the aim is now reached to have extended τ-ARGUS in such a way that it is able to deal with an important sub-class of linked tables, namely hierarchical tables. A hierarchical table is an ordinary table with marginals, but also with additional subtotals. Hierarchical tables imply much more complex optimisation problems to be solved than single tables. Some approximation methods exist for finding optimal solutions for these problems. The extension version 2.1 of τ-ARGUS was released in the CASC (Computational Aspects of Statistical Confidentiality) project.

3. THE RELEASE OF MICRODATA FOR RESEARCHERS AND PUBLIC USE MICRODATA FILES

Many users of surveys are satisfied with the safe tables released by statistical offices. However, some users require more information. For many surveys microdata for researchers are released. The software package μ-ARGUS (Hundepool et al, 2002b) is of help in producing these microdata for researchers. For the microdata for researchers Statistics Netherlands uses the following set of rules:

1. Direct identifiers should not be released.
2. The indirect identifiers are subdivided into extremely identifying variables, very identifying variables and identifying variables. Only direct regional variables are considered to be extremely identifying. Each combination of values of an extremely identifying variable, a very identifying variable and an identifying variable should occur at least 100 times in the population.
3. The maximum level of detail for occupation, firm and level of education is determined by the most detailed direct regional variable. This rule does not replace rule 2, but is instead an extension of that rule.
4. A region that can be distinguished in the microdata should contain at least 10 000 inhabitants.
5. If the microdata concern panel data direct regional data should not be released. This rule prevents the disclosure of individual information by using the panel character of the microdata.

In the case of most Statistics Netherlands' business statistics the responding enterprises are obliged by a law on official statistics to provide their data to Statistics Netherlands. This law dates back to 1936 and was renewed in 1996 without changing the obligation of enterprises to respond. No individual information may be disclosed when the results of these business surveys are published. The law states that no microdata for research may be released from these surveys. Statistics Netherlands can therefore provide two kinds of information from these surveys: tables and public use microdata files. Public use microdata files contain much less detailed information than microdata for research. The software package μ-ARGUS (Hundepool et al, 2002b) is also of help in producing public use microdata files. For the public use microdata files Statistics Netherlands uses the following set of rules:

1. The microdata must be at least one year old before they may be released.
2. Direct identifiers should not be released. Also direct regional variables, nationality, country of birth and ethnicity should not be released.

3. Only one kind of indirect regional variables (e.g. the size class of the place of residence) may be released. The combinations of values of the indirect regional variables should be sufficiently scattered, i.e. each area that can be distinguished should contain at least 200 000 persons in the target population and, moreover, should consist of municipalities from at least six of the twelve provinces in the Netherlands. The number of inhabitants of a municipality in an area that can be distinguished should be less than 50 % of the total number of inhabitants in that area.
4. The number of identifying variables in the microdata is at most 15.
5. Sensitive variables should not be released.
6. It should be impossible to derive additional identifying information from the sampling weights.
7. At least 200 000 persons in the population should score on each value of an identifying variable.
8. At least 1 000 persons in the population should score on each value of the crossing of two identifying variables.
9. For each household from which more than one person participated in the survey we demand that the total number of households that correspond to any particular combination of values of household variables is at least five in the microdata.
10. The records of the microdata should be released in random order.

According to this set of rules the public use files are protected much more severely than the microdata for research. Note that for the microdata for research it is necessary to check certain trivariate combinations of values of identifying variables and for the public use files it is sufficient to check bivariate combinations. However, for public use files it is not allowed to release direct regional variables. When no direct regional variable is released in a microdata set for research, then only some bivariate combinations of values of identifying variables should be checked according to the statistical disclosure control rules. For the corresponding public use files all the bivariate combinations of values of identifying variables should be checked.

The software package μ-ARGUS is of help to identify and protect the unsafe combinations in the desired microdata file. Thus rule 2 for the microdata for researchers and the rules 7 and 8 for the public use microdata files can be checked with μ-ARGUS. Global recoding and local suppression are two data protection techniques used to produce safe microdata files. In the case of global recoding several categories of an identifying variable are collapsed into a single one. This technique is applied to the entire data set, not only to the unsafe part of the set, so that a uniform categorisation of each identifying variable is obtained.

In the field of microdata several new techniques are being investigated in the CASC (Computational Aspects of Statistical Confidentiality) project. New methodologies like post randomisation (PRAM), micro-aggregation and noise-addition will be implemented in new versions of μ-ARGUS that will be released in the near future. PRAM is a perturbative method for disclosure protection of categorical variables (see e.g. Gouweleeuw, Kooiman, Willenborg and De Wolf, 1998). Applying PRAM means that for each record in a microdata file the score on one or more categorical identifying variables may be misclassified into different scores according to a predetermined probability mechanism (Van den Hout and Van der Heijden, 2002). Since the original data file is perturbed, it will be difficult for an intruder to identify records with certainty as corresponding to certain individuals in the population. In other words, the randomness of the procedure implies that matching a record in the perturbed file to a record of a known individual in the population could, with a high probability, be a mismatch. The records in the original file are thus protected, which is the main goal of applying PRAM. On the other hand, since the probability mechanism that is used when applying PRAM is known, characteristics of the true data can be estimated from the perturbed data file. Hence, it is still possible to perform all kinds of statistical analyses after PRAM has been applied. However, using the transition matrix with the misclassification probabilities to take into account the perturbation due to PRAM requires extra effort and becomes more complex when the research questions become more complex. Two key questions concerning PRAM are the following:

- How should the misclassification probabilities be chosen in order to make the released microdata file safe?
- How should statistical analyses be adjusted in order to take into account the misclassification probabilities?

The first question is addressed in Willenborg and De Waal (2001) and the second question in Van den Hout and Van der Heijden (2002).

The implementation of the new methodologies in μ-ARGUS will allow for experimenting with these techniques. In the future a mixture of several disclosure protection methods can be applied, e.g. combining PRAM, global recoding and local suppression. Generally speaking, it is at the moment still unclear what the implications will be for statistical disclosure control rules in official statistics.

To measure the quality of the methods applied disclosure risk and information loss models will be implemented in new versions of μ-ARGUS too. A disclosure risk model is specified to distinguish safe from unsafe microdata. Disclosure models can differ greatly in their levels of sophistication. In a fairly simple disclosure model a combination of values is safe only if the estimated frequency of its occurrence in the population is above a certain threshold value. Which combinations to consider is also part of the disclosure risk model that one applies, and should be specified by the data protector. Fienberg and Makov (1998) and Skinner and Holmes (1998) have proposed more advanced disclosure risk models. They use log linear models for the estimation of the individual risk. Whatever disclosure risk model is used, one always has to make some assumptions on the nature of possible attacks by intruders to the privacy of an individual (see e.g. Keller and Bethlehem, 1992, Mokken et al, 1992, Elliot and Dale, 1999 and Elliot, 2001).

If unsafe microdata are going to be transferred into safe microdata it is necessary to have a measure of information loss at one's disposal. This measure is used to limit the amount of damage done to the microdata when they are being modified by the data protector. In case of applying local suppressions, μ-ARGUS uses the number of local suppressions as measure of information loss. The more suppressions the higher the information loss. The optimisation problem that has to be solved is to select the local suppressions in such a way that the resulting microdata are safe and the associated information loss is minimised. As μ-ARGUS uses the number of local suppressions as measure of information loss, the problem how to suppress in an optimal way can be solved record-wise (De Waal and Willenborg, 1998). Another possibility is to choose the number of different categories affected by the suppression as measure of information loss. The minimisation problem then tends to be more difficult to solve in practice. In some cases this problem can be decomposed into a number of smaller, and therefore easier to solve, problems. In case of global recoding of an identifying variable the information loss depends on the valuation of the importance of the variable and the valuation of each of the possible predefined codings for the variable. Optimisation models for a special form of global recoding have been developed by Hurkens and Tiourine (1998). Both global recoding and local suppression lead to information loss, because either less detailed information is provided or some information is not given at all. A balance between global recoding and local suppression should always be found in order to make the information loss due to the statistical disclosure control measures as low as possible. It is recommended to start by recoding some variables globally until the number of unsafe combinations that has to be protected is sufficiently low. Then the remaining unsafe combinations have to be protected by local suppressions.

4. OTHER METHODS THAT ALLOW USE OF DATA

All techniques described in chapter 3 necessarily involve data manipulation or suppression and are likely to reduce the quality of estimates to be produced from the data. As a result, National Statistical Institutes (NSIs) have begun to investigate other methods that allow use of data while protecting confidentiality of sensitive information given by respondents. These methods allow the data to be used in an environment controlled by the NSI and require that its use be subject to the same legal and ethical protections placed on the NSI itself.

Some NSIs (e.g. in the U.S.A.) have introduced the process of licensing whereby institutions and researchers outside the NSIs temporarily gain access to (a part of the) data at their site by agreement to conform to legal protections surrounding those data that are imposed on the NSI. Data licensing is thus a way to provide access to data when they cannot be released to the public because of confidentiality concerns. It is necessary that periodic inspections are performed of the licensed sites. Also a good organisation of the licensed files within the NSI is a necessity for the agreement to become a success.

Probably the most important access modality developed in the past decade is that of restricted access sites. These sites permit NSIs to respond to the microdata needs of researchers. Some researchers need namely more information

than is available in the released microdata for researchers or public use microdata files. As the releasing of richer data is not allowed, it is then possible for individual researchers to perform their research on richer microdata on the premises of the NSIs. Statistics Netherlands is one of the NSIs that has such a facility. Bona fide researchers have the opportunity to work on-site in a secure area within Statistics Netherlands. Researchers can choose at will between the two locations of Statistics Netherlands: Voorburg in the west of the Netherlands and Heerlen in the south of the Netherlands. The possibility to export any information is however only possible with the permission of the responsible statistical officer. They can apply standard statistical software packages and also bring their own programmes. Like all employees of Statistics Netherlands, these people who work on-site have to swear an oath to the effect that they will not disclose the individual information of respondents (Kooiman, Nobel and Willenborg, 1999).

The researchers who work on-site on Statistics Netherlands' economic data have to take the rules of the Centre for Research of Economic Microdata (CEREM) into account. The most important rules are:

- researchers must be associated with a recognised research institute (e.g. a university);
- there must be a research proposal that conforms to current scientific standards;
- the researcher and his superior have to sign a confidentiality warrant;
- the researcher obtains only access to the data needed for his project;
- the data do not contain information on names and addresses of the enterprises;
- data related to the two most recent years will not be supplied;
- it is forbidden to let data or not safeguarded intermediate results leave the premises of Statistics Netherlands;
- all prospective publications will be screened with respect to the risk of disclosure;
- all publications will be in the public domain;
- a public register contains the researcher's name(s), the research project, the publication(s) and the databases provided.

The facility provided by Statistics Netherlands is not free of charge. As a rule the researcher has to pay the cost for the supply of the required data. In addition, there is a tariff for using the on-site facility.

Finally, an option is to allow remote access. This access modality combines the advantage of licensing that researchers can stay in their own institute and the advantage of working on site that the data stay in the NSI. Normally, researchers get access through an intermediary controlled by the NSI that guarantees that all use conforms to the law. One step further goes the option of remote execution. Then no longer an intermediary is placed between the researcher and the NSI. With remote execution researchers can execute set-ups without having the data on their own PC. Although remote execution is a more efficient option than remote access the question is whether the security systems are strong enough to let this technique become an often used modality. Currently, Statistics Netherlands has a Centre for Policy Research that is running a successful pilot with the Ministry of Social Affairs and Employment. In this pilot the remote execution is limited in the sense that employees of Statistics Netherlands still check manually the set-ups that are sent to the Centre for Policy Research.

5. DISCUSSION AND CONCLUSIONS

The software packages τ-ARGUS and μ-ARGUS have emerged from the Statistical Disclosure Control (SDC) project that was carried out under the Fourth RTD Framework Programme of the European Union. These software packages appear to be of great help in the practice of statistical disclosure control. Many of the protection problems of statistical data can be solved using the ARGUS packages.

The manuals (Hundepool et al, 2002a and b) are of great help for the users of the ARGUS packages. However, there are always additional things to desire. In the case of τ-ARGUS more research is needed into how a set of linked tables and corresponding metadata could efficiently be dealt with in an automated way. More research is also needed into how consecutive years of the same survey can be protected from disclosure. In the case of μ-ARGUS, it is important to clarify in the package the difference between protecting microdata for research and protecting public use microdata files. As μ-ARGUS can be used with lots of different protection criteria, it is important to help the

users to understand how different strategies can be executed using the package. Recently, research has been directed at perturbation methods by adding stochastic noise to microdata. It would be good to have several options in μ-ARGUS to perturb data as a protection technique. Especially interesting for a user is to know how well protected microdata are after the perturbation process.

It can be concluded that there is still a lot of 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 soon be released to the user community. The production of these new versions is part of the CASC project (see Hundepool, 2001). To promote the results of the statistical projects under the Fourth RTD Framework Programme of the European Union the AMRADS (Accompanying Measures in Research and Development in Statistics) project is funded under the Fifth RTD Framework Programme. Many courses and conferences are being organised, among other topics, about statistical disclosure control. These activities will stimulate the progress in the implementation of statistical disclosure control methods and techniques in many different countries.

A couple of National Statistical Institutes (NSIs) and Market Research Bureaus have their own research activities (see e.g. Bethlehem and Pannekoek, 1998). Besides those NSIs the statistical agency of the European Union, Eurostat, has established itself as a main promoter of research in statistics. A dedicated budget for subsidising targeted research and development projects has been available in recent years. Many projects (e.g. the projects SDC, CASC and AMRADS discussed in this paper) were subsidised by the European Union. Before the Fifth RTD Framework Programme of the European Union started, the European Plan for Research in Official Statistics (EPROS) was launched. Statistical disclosure control is one of the topics mentioned in EPROS. Eurostat has stimulated the forming of consortia of researchers from Universities, NSIs and Market Research Bureaus. This way, many ideas have been exchanged and many researchers learnt a lot form each other. Not all subsidised projects always lead to good results that can be implemented in practice. However, it is hard to predict which projects will become most successful. Critical success factors are at any rate a clear aim of the project and an efficient project organisation. Hopefully, Eurostat (and maybe also other international organisations) will continue to find ways to stimulate research in statistics in the future as well. Although one never knows exact outcomes of research projects beforehand, it is clear that subsidising statistical research projects has led to economies of scale and speeded up the process towards better and more comparable statistics.

In this paper methods have been described that have been developed to protect confidentiality, while at the same time providing access to data, through various means that either alter the data or restrict access to them. The balance between data confidentiality and data access is a delicate one. Hopefully, the new research methods and software for statistical disclosure control can help in keeping the right balance.

As part of the CASC project new versions of the ARGUS packages become available for users. Recently new manuals for τ-ARGUS and μ-ARGUS became available (Hundepool et al, 2002a and b) that have been tested intensively as part of the CASC project. Both manuals are of great help to the testers. In the new version of τ-ARGUS HiTaS has been included, so that from this version onwards hierarchical tables can be dealt with as well. In the new version of μ-ARGUS new options are PRAM and individual risk models. The ARGUS packages have moved towards interfaces with several state of the art engines produced by statisticians from many different countries.

REFERENCES

Bethlehem, J. and J. Pannekoek (1998), “Statistical research activities at Statistics Netherlands”, *Research in Official Statistics*, 1, pp. 131-134.

Citteur, C.A.W. and L.C.R.J. Willenborg (1993), “Public use microdata files: current practices at national statistical bureaus”, *Journal of Official Statistics*, 9, pp. 783-794.

Elliot, M. (2001), “Advances in data intrusion simulation: a vision for the future of data release”, *Statistical Journal of the United Nations Economic Commission for Europe*, 18, pp. 383-391.

Elliot, M. and A. Dale (1999), "Scenarios of attack: the data intruder's perspective on statistical disclosure risk", *Netherlands Official Statistics*, 14, pp. 6-10.

Fienberg, S.E. and U.E. Makov (1998), "Confidentiality, uniqueness and disclosure limitation for categorical data", *Journal of Official Statistics*, 14, pp. 385-397.

Gouweleeuw, J.M., P. Kooiman, L.C.R.J. Willenborg and P.-P. de Wolf (1998), "Post randomisation for statistical disclosure control: theory and implementation", *Journal of Official Statistics*, 14, pp. 463-478.

Groot, A. and C.A.W. Citteur (1997), "Accessibility of business microdata", *Netherlands Official Statistics*, 12, pp. 18-32.

Hout, A. van den and P.G.M. van der Heijden (2002), "Randomised response, statistical disclosure control and misclassification: a review", *International Statistical Review*, 70, pp. 269-288.

Hundepool, A.J. (2001), "Computational aspects of statistical confidentiality: the CASC-project", *Statistical Journal of the United Nations Economic Commission for Europe*, 18, pp. 315-320.

Hundepool, A., A. van de Wetering, P.P. de Wolf, S. Giessing, M. Fischetti, J.J. Salazar and A. Caprara (2002a), *τ-ARGUS, user's manual, version 2.1*, Voorburg, The Netherlands: Statistics Netherlands.

Hundepool, A., A. van de Wetering, L. Franconi, A. Capobianchi and P.P. de Wolf (2002b), *μ-ARGUS, user's manual, version 3.1*, Voorburg, The Netherlands: Statistics Netherlands.

Hurkens, C.A.J. and S.R. Tiourine (1998), "Models and methods for the microdata protection problem", *Journal of Official Statistics*, 14, pp. 437-447.

Keller, W.J. and J.A. Bethlehem (1992), "Disclosure protection of microdata: problems and solutions", *Statistica Neerlandica*, 46, pp. 5-19.

Kooiman, P., J.R. Nobel and L.C.R.J. Willenborg (1999), "Statistical data protection at Statistics Netherlands", *Netherlands Official Statistics*, 14, pp. 21-25.

Mokken, R.J., P. Kooiman, J. Pannekoek and L.C.R.J. Willenborg (1992), "Disclosure risks for microdata", *Statistica Neerlandica*, 46, pp. 49-67.

Schulte Nordholt, E. (2001), "Statistical disclosure control (SDC) in practice: some examples in official statistics of Statistics Netherlands", *Statistical Journal of the United Nations Economic Commission for Europe*, 18, pp. 321-328.

Skinner, C.J. and D.J. Holmes (1998), "Estimating the re-identification risk per record in microdata", *Journal of Official Statistics*, 14, pp. 361-372.

Waal, T. de and L.C.R.J Willenborg (1998), "Optimal local suppression in microdata", *Journal of Official Statistics*, 14, pp. 421-435.

Willenborg, L.C.R.J. (1993), "Discussion statistical disclosure limitation", *Journal of Official Statistics*, 9, pp. 469-474.

Willenborg, L.C.R.J. and T. de Waal (1996), *Statistical disclosure control in practice, Lecture Notes in Statistics 111*, New York: Springer-Verlag.

Willenborg, L.C.R.J. and T. de Waal (2001), *Elements of statistical disclosure control, Lecture Notes in Statistics 155*, New York: Springer-Verlag.