Census hypercubes, Statistical Disclosure Control of frequency tables and the Census Hub[1]

Eric Schulte Nordholt

Statistics Netherlands, P.O. Box 24500, 2490 HA The Hague, The Netherlands
Tel.: +31 70 337 4931; Fax: +31 70 387 7429; E-mail: e.schultenordholt@cbs.nl

Abstract: *The programme for data dissemination that Eurostat has implemented for the Census Round of 2011 is based on an innovative approach. The basic data is in the form of hypercubes (high-dimensional tables). The size of these hypercubes has a relevant impact on confidentiality issues: while for a set of predefined common two- or three-dimensional tables the disclosure control for census data could be feasible to implement, such control becomes a real challenge as more dimensions are added. If the confidentiality methods applied differ from one country to another, the comparability of the data might be affected. It has thus been considered worthy to explore the margins of action for a common approach at EU level for disclosure control of census data.*

Statistical information is nowadays available for the public in tabular and microdata form. These microdata can be conveyed with CD-ROMs, USB sticks and other means. Recently, other possibilities of getting statistical information have become more popular as on-site facilities, remote access and remote execution. With remote 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 National Statistical Institutes (NSIs) have the possibility to let bona fide researchers work on-site within the premises of the NSI. In Europe the software packages τ-ARGUS and μ-ARGUS were developed for the Statistical Disclosure Control of tabular data and microdata. With these packages the protection processes have been automated to a high degree.

[1] The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

At the EU level so far for the Census only tabular data are produced and compared between countries. In this paper ideas are sketched how modern ideas on Statistical Disclosure Control can be of help to the challenge to provide comparable and safe hypercubes for the Census Round of 2011.

Keywords: *Census data; hypercubes; legal issues; microdata; on-site facility; remote access; remote execution; Statistical Disclosure Control; tabular data*

1. Introduction

The compilation of the set tables of the 2001 Census was based on a gentlemen's agreement with Eurostat. A general feeling after the 2001 Census in Europe was that a gentlemen's agreement was not enough to continue the ten-yearly censuses in Europe. This was the reason to give the 2011 Census Round a broader basis with four Regulations (European Commission, 2008, 2009b, 2010a and 2010b). With these four regulations the population definitions, census variables and their categories, census hypercubes (high-dimensional tables) and metadata are harmonised within the EU. Moreover, the technical format in which the data have to be delivered has been specified and all countries will produce a quality report in which the methodology used is described.

Regulation 763/2008 on population and housing censuses (European Commission, 2008) is acknowledging the Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing. This Regulation states that "Member States shall take all measures necessary to meet the requirements of data protection. The Member States' own data protection provisions shall not be affected by this Regulation." The transmission of data subject to statistical confidentiality is governed by specific EU regulations, ensuring the physical and logical protection of confidential data and that no unlawful disclosure or non-statistical use occurs when Community statistics are produced and disseminated. In particular, the new "EU Statistical Law" devotes an entire chapter to statistical confidentiality (European Commission, 2009a, chapter V). In other words, what is considered confidential at national level, it remains such also once transmitted to Eurostat and, if a country wants to transmit confidential data, this has to be done in accordance with the EU regulations in force.

The programme for data dissemination that Eurostat has implemented for the Census Round of 2011 is based on an innovative approach. The basic data is in the form of hypercubes. The size of these hypercubes has a relevant impact on confidentiality issues: while for a set of predefined common two- or three-dimensional tables the disclosure control for census data could be feasible to implement, such control becomes a real challenge as more dimensions are added.

If the methods applied differ from one country to another, the comparability of the data might be affected. Moreover, for the users it would be easier to understand the constraints deriving from the application of one single method for all countries rather than the consequences of several national methods for confidentiality. It has thus been considered worthy to explore the margins of action for a common approach at EU level for disclosure control of census data.

Normally, there are no specific provisions for the confidentiality of census data, as this is covered by the restrictions applied to all statistical data. If mention is made in the census law, this is usually recalling the more general national regulations on the subject. The practical implementation of these general provisions is in general a task

of the national statistical offices; however, a few years ago not many of them had already defined a methodology for disclosure control for the next census. There was therefore a large support for an action by Eurostat for analysing the feasibility of a common methodology for disclosure control of the census data.

Up to the late 1980's microdata were rarely sent to Eurostat, the European statistical office. There was a general reliance on submission by National Statistical Institutes (NSIs) of agreed tabular data. National confidentiality rules in some of the European countries made it impossible to harmonise European statistics. This was an unwanted situation for all NSIs, and especially for Eurostat. Therefore, a regulation on the transmission of confidential data to Eurostat has been prepared and was finally adopted by the Council in June 1990 as Regulation 1588/90 (European Commission, 1990).

In January 1994, these measures have been defined and formally adopted by the Member States through the Committee on Statistical Confidentiality (CSC). This Committee met at least once a year at the Eurostat office in Luxembourg. This Committee discussed the implementation and evaluation of European Regulations on the dissemination of microdata and tabular data. Also revisions to the basic statistical legal framework were considered.

Another relevant Council Regulation is No 322/97 of February 1997 (European Commission, 1997). This Regulation defined the general principles governing Community statistics, the processes for the production of these statistics and established detailed rules on confidentiality. This Regulation could be considered as the general statistical law of the European Union.

A new statistical legal framework was introduced in 2009 (European Commission, 2009a). One of the new aspects concern statistical confidentiality: the need to enhance the role of the NSIs and Eurostat for organisational, co-ordination and representation purposes was noted. In this context the former Statistical Programme Committee was replaced by a new Committee, the European Statistical System Committee (ESSC). This new Committee is also entrusted with the functions of the CSC, which thus ceased to exist. The last meeting of the CSC was held in December 2008.

The European Statistical System (ESS) is defined by Regulation 223/2009 on European statistics (European Commission, 2009a) as the partnership between the Community statistical authority (Eurostat) and all national authorities responsible for the development, production and dissemination of European Statistics (ES).

The availability of confidential data for the needs of the ESS is of particular importance in order to maximise the benefits of the data with the aim of increasing the quality of European statistics and to ensure a flexible response to the newly emerging Community statistical needs.

The transmission of confidential data between ESS partners is allowed if necessary for the production, development and dissemination of ES and also for increasing the

quality of these statistics. The conditions for their further transmission, in particular for scientific purposes, are also strictly defined.

The ESSC is consulted on all draft comitology measures submitted by the Commission in the domain of statistical confidentiality.

This paper gives some background on statistical data availability in section 2. The release of safe tabular and microdata is described in sections 3 and 4. For producing safe tables the software package τ-ARGUS was developed; microdata for research and public use microdata files can be produced using the software package μ-ARGUS. Also other methods exist that allow use of microdata. The option for bona fide researchers to work on-site at Statistics Netherlands on richer microdata files is explained in section 5. Remote facilities are discussed in section 6. The challenges of an EU harmonised approach for Census disclosure control and technical aspects of an EU Census methodology are the topics of sections 7 and 8. Finally, some conclusions are drawn in section 9.

2. Statistical Data Availability

Given the ethical codes (ISI, 1985 and 2010 and European Statistics Code of Practice, 2011), UNECE Principles and Guidelines (Trewin et al, 2007 and Pink et al, 2009) and laws described in the previous section, 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 CD-ROMs, USB sticks and other means. Recently, also other possibilities of getting statistical information have become more popular: 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 release as much detail in these data as possible without disclosing individual information (Willenborg and De Waal, 1996 and 2001).

This section and the next sections on Statistical Disclosure Control discuss the available methods to protect sensitive information. The microdata of surveys have to be protected against the risk of disclosure. The software package μ-ARGUS (Hundepool et al, 2010a) was developed to handle this protection process. The tables produced by statistical offices on the basis of the microdata have to be protected as well. Therefore the software package τ-ARGUS (Hundepool et al, 2010b) can be applied on the tables produced.

The software packages μ-ARGUS and τ-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 was 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 had new objectives. It concentrated 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) worked closely together. One of the main tasks of this consortium was to further develop the ARGUS-software which has been put in the public domain by the SDC project consortium. The CASC project involved both research and software development. As far as research is concerned the project concentrated on those areas that could be expected to result in practical solutions, which were then being built into the software. The CASC project had been designed around the software twin ARGUS. This made 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). After the SDC and CASC projects the software packages were further developed in CENEX and ESSnet projects on Statistical Disclosure Control. Statisticians from many countries attended courses on Statistical Disclosure Control and became users of μ-ARGUS and τ-ARGUS.

3. The release of tabular data

3.1 Primary suppressions

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, 2010b) 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 *k* 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. Often used values for *n* and *k* are 3 and 70 %, but also dominance rules with other parameter values can be used in τ-ARGUS. 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. As the p-percent rule and the pq rule are available in τ-ARGUS, it can be expected that these rules will become more popular.

3.2 Secondary suppressions

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. These further suppressions are called secondary suppressions. 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 options 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 options 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 fourth option one can take the hierarchy of the table into account and then software tailored to the specific needs is required. In τ-ARGUS the first three options are available. These three implemented options may lead to different resulting groups of secondary suppressions. The different results can then 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, k 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. Newer versions of τ-ARGUS support linked tables. This implies that τ-ARGUS has been extended 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 first version of τ-ARGUS that supports linked tables was released in the CASC (Computational Aspects of Statistical Confidentiality) project.

4. The release of microdata

4.1 The release of microdata for researchers

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, 2010a) 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.

4.2 The release of public use microdata files

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 several times 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, 2010a) 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.

5. Other methods that allow use of data

Data manipulation or suppression 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 years 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: The Hague in the west of the Netherlands and Heerlen in the south of the Netherlands. However, the possibility to export any information is 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 Centre for Policy Related Statistics, a unit within Statistics Netherlands, runs the on-site facility of the office. The researchers who work on-site on Statistics Netherlands' data have to take the rules of the Centre for Policy Related Statistics into account. The most important rules are:

- researchers must be associated with a recognised research institute (e.g. a university);
- 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 direct identifiers as name and address information;
- 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.

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. The researchers do not have to pay the much larger costs of producing microdata as these costs have already been paid by the Dutch tax payers.

Finally, an option is to allow remote access. This access modality combines the advantage of licensing and microdata for researchers (under contract) 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' Centre for Policy Related Statistics is running both the on-site and the remote facility. The remote facility offered is limited in the sense that employees of Statistics Netherlands still check manually the results before they can be released, just like in the case of on-site analyses. More information about remote access in the Netherlands can be found in the next section.

6. Remote access at Statistics Netherlands

Statistics Netherlands has a longstanding tradition of releasing safe microdata to researchers. This dates back to the beginning of the nineties of the previous century. The microdata files were made available to researchers at universities under a strict contract. These files were protected against statistical disclosure using a specific set of disclosure control rules that were presented in subsection 4.1. The researchers could analyse the microdata files on their own computers. These microdata under contract still exist and can be ordered via the institute DANS (Data Archiving and Networked Services), see http://www.dans.knaw.nl/en/.

However, the level of detail in these microdata files made it impossible for researchers to perform some of the analyses they wanted. The Statistical Disclosure Control restrictions, enforced by Law in the Netherlands, did not allow more detailed microdata files to be made available to researchers outside the premises of Statistics Netherlands. The law demands that the use of and the results from analyses based on detailed microdata files should be under strict control of Statistics Netherlands.

Bona fide researchers who want to make more detailed analyses can work on-site at the premises of Statistics Netherlands. The detailed microdata files are then made available to selected researchers in a controlled setting. The selected researchers can perform their desired analyses, but their results are checked by Statistics Netherlands' staff for possible disclosure risk, before the researchers are allowed to bring the results outside the controlled setting.

The on-site facility has proven to be very successful. Many researchers have been using the facility and from time to time a number of researchers are working at the facility simultaneously. A major drawback of this facility is that the researchers have to travel to the premises of Statistics Netherlands, in order to be able to do their analyses. Even in a small country like the Netherlands this proved to be inefficient in many situations. Moreover, Statistics Netherlands has to organise specially equipped offices for the researchers. As more and more facilities became available to use safe internet connections, the question has risen whether an equivalent of the on-site facility could be built over the internet. This has led to the current remote access facility. First a life test of this system was executed as a pilot project with the University of Tilburg as partner. After this pilot turned out to be successful almost all other research organisations in the Netherlands and even some abroad were connected to this service.

The main idea is that the remote access facility should resemble the 'traditional' on-site situation as much as possible, concerning confidentiality aspects. Moreover, it should resemble the look and feel of the remote access facility without the aspect of having to travel to the premises of Statistics Netherlands.

At the remote access facility access of authorised users only is ensured because researchers cannot enter the premises of Statistics Netherlands unaccompanied. Moreover, only a selected group of researchers working at universities and research institutes is allowed to utilize this facility. The remote access facility is making use of a citrix connection and biometric identification, to ensure that the researcher who is trying to connect to the facility is indeed the intended person. Whenever the researcher wants to access the facility, he will be identified by his fingerprint. Thus biometric identification is used, in combination with PKI (Public Key Infrastructure) certificates.

The network that is used by the facility is not connected to the production network. Moreover, the computers that the researcher can use are such that no removable media can be used (no CD-ROMs, USB sticks or other means) and no internet connection. This means that the microdata used by the researcher can only be accessed using a special computer at the premises of Statistics Netherlands and that the researcher cannot take a copy of the data to the institute where he is working. He is able to view the (intermediate) results of his analyses on the screen, but he is not able to send those results to his institute by e-mail or otherwise. Moreover, he is not allowed to take a printout of the results to his institute either, without having it checked by a member of Statistics Netherlands' staff for confidentiality. This ensures that the microdata and the intermediate results remain at Statistics Netherlands. It is virtually impossible to directly connect from an external computer to the production network of Statistics Netherlands.

For both the on-site and the remote facility, legal measures are taken to prevent misuse of the microdata. To that end, a contract will be signed by the institute where the researcher is working. Moreover, a statement of secrecy is signed by the researcher as well as the institute he works for.

The check on the output for confidentiality is done by hand. Obviously, this is very labour-intensive. In the future, this should ideally be facilitated by some software. However, since the output of the results can be very diverse in format (R, SAS, SPSS, Stata, etc.) the development of such software is very difficult. Moreover, at Statistics Netherlands, no automated checks of the rules are available to decide whether or not general analysis' results breach confidentiality. To make things easier researchers are nowadays stimulated to write their complete papers on the special PCs used for remote access from their own institute. This way, they only have to ask permission at the end of their project and all the labour-intensive checking of preliminary results can be prevented.

7. The challenges of an EU harmonised approach for Census disclosure control

The release of microdata of the 2011 Census on a European scale is infeasible. Therefore, the 2011 Census Round is based on four Regulations (European Commission, 2008, 2009b, 2010a and 2010b) and aims at comparing census hypercubes (high-dimensional tables) within the EU. As it is not yet clear how these hypercubes can be protected, Eurostat set up a Task Force on "EU Methodology for Census Data Disclosure Control" (CENSDC), composed by experts in the field of disclosure control from Germany, Estonia, Italy, the Netherlands, Portugal and the United Kingdom. Its specific objective was to identify and resolve areas of difficulty relating to the confidentiality data treatment of population and housing census data, adopting or developing a harmonised methodology which respects the national regulations. The Task Force CENSDC met two times, and presented the results of its work at the Eurostat Working Group on Demography and Census held in 2010.

The Task Force CENSDC had to deal with several conceptual challenges. First of all, the relatively high number of dimensions of the hypercubes complicated the application of standard methods of disclosure control. Secondly, given the approach of the Eurostat Census Hub, it had also to be decided where these controls should take place, in the national databases or "on the fly". Third, consistency of the tables' results should be ensured between hypercubes and between extractions. Fourth, as the expertise and tools for disclosure control available in each Member State can be rather different, a common approach should be as easy as possible to implement. Fifth, the method should possibly be easy to understand for the common user. Sixth, as one of the added values of a census is the availability of detailed information, the loss of data should be minimised. Last, but certainly not the least, each country has its own regulation on confidentiality that had to be respected.

For the 2011 Census Round, there is a more ambitious programme of census data dissemination than ever before. More information than in past rounds on all the EU Member States (and other countries willing to be part of the census dissemination programme) will be put at disposal of the users by means of a single interface, the Eurostat Census Hub. This system transmits any user's query to the national databases, retrieve the information, and display it all together. From the confidentiality point of view, the other side of the medal is that, given the high number of dimensions and the freedom of the user to build the tables of interest (including repeated queries), the risk associated with standard methods for disclosure control need to be carefully assessed.

One basic choice is if the national data on which the extraction is made have to be already "cleaned" for confidentiality, or if the disclosure control can be made "on the fly", just before the data are displayed to the user. The latter option would be justified by the fact that the number of dimensions displayed to the user would be much less than the total number of dimensions available in the related hypercube, thus reducing the risk of disclosure. On the other side, such an approach would mean that confidential data are somehow being transferred from the national database

(with all the implications from the IT security point of view), and that the risks connected to multiple queries, helpful for potential intruders, are increased. In fact, the levels at which the data can be treated for confidentiality are three: microdata, hypercube or extraction.

Whatever the level at which the disclosure control is implemented, it is considered important that the results disseminated to the users are consistent between selected tables and between extractions. The large use is made of census data and the fact they are queried for a long period require that the users will not be confronted with different results depending on the time of the extraction or on the hypercube of reference. Although some confidentiality methods could be very effective, it should also be assessed whether they generate undesired consequences in terms of data comparability.

Methods and tools for disclosure control have reached a certain level of complexity. Several European projects have been devoted to this domain, among which the CASC project (2000-2003, see http://neon.vb.cbs.nl/casc/), the CENEX project (2006) and the ESSnet project (2008-2009), and UNECE and Eurostat organise regular Work Sessions on Statistical Data Confidentiality (see http://www.unece.org/stats/archive/04.06.e.htm , http://www.unece.org/stats/documents/2009.12.confidentiality.html and http://www.unece.org/stats/documents/2011.10.confidentiality.html). However, it cannot be assumed that the same level of expertise is available in all statistical offices. If the statistical disclosure control has to be applied at national level based on a common approach, then the harmonisation of the confidentiality methods has to take into account additional requisites such as the easiness of implementation and the ready availability of tools (possibly without excessive costs). In periods of scarcity of resources, countries cannot be asked to sustain relevant additional expenses.

Besides the above challenges pertinent to a harmonised approach, there are also other elements to be considered part of the exercise, regardless if the method is applied to all countries or if it is country-specific. Whatever the disclosure control applied, the user should be informed of its characteristics and consequences on the data. The easier a method, the easier for the common user to understand the implications (and likely the easier to communicate this information). For the sake of transparency and overall data quality, this aspect should not be totally neglected.

One of the major features of the census is that it provides information at small geographical level and for small groups of persons, sometimes even being the only available source. Such a richness should be preserved as much as possible vis-à-vis the need to ensure the data confidentiality. Although the extensive application of rigorous methods of statistical disclosure control may help preventing (to the possible extent) confidentiality breaches, at the same time it can significantly reduce the availability of information. It is in the interest of the users to try to minimise the impact of disclosure control methods on data availability. The filters for

confidentiality should be applied to a reasonable extent, bearing in mind the related unavoidable loss of (detailed) information.

Finding a satisfactory solution for each of the above-listed challenges risk to be a hopeless task, and adding the national constraints makes things even more complicated. However, despite of the large number of national requirements on data confidentiality, it still makes sense to look for a common solution because the national regulations are often only setting general principles, leaving in many cases the practical implementation (and related methodological choices) to the national statistical offices. If these statistical offices agree on an harmonised approach, there is no infringement of the national provisions, as these bodies are the technical responsible of the appropriate disclosure control to the national data. The wide support expressed by the statistical offices to a joint action at EU level on confidentiality of census data can be seen as an expression of the need of exchange of experiences and assistance on technical issues: on this, the Task Force CENSDC can certainly play an important role.

8. Technical aspects of an EU Census methodology

The Regulation (EC) No 763/2008 (European Commission, 2008) is output oriented, i.e. it is open to the use of different data sources, but requires the respect of the essential features of population and housing censuses, the use of harmonized definitions, technical specifications, topics and breakdowns. The Census regulation foresees unified reporting years (the first being 2011), a common EU dissemination programme, technical standards for the data transmission and the establishment of quality reports for European purposes. Concerning the statistical confidentiality, the following aspects are of particular importance:

Article 4 (2) foresees that the "Member States shall take all measures necessary to meet the requirements of data protection. The Member States' own data protection provisions shall not be affected by this regulation". That means that the protection of census data comes under the responsibility of the Member States, and has to be done at their level rather than by the Commission. Article 4 (2) provides further that the European Commission is not entitled to issue legislation on the disclosure protection of census data on the basis of the Census regulation. However, Article 6 (4) stipulates "The Commission (Eurostat), in cooperation with the competent authorities of the Member States, shall provide methodological recommendations designed to ensure the quality of the data and metadata produced, acknowledging, in particular, the Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing". Consideration 3 to the regulation explains that "in view of methodological and technological developments, best practices should be identified and the enhancement of the data sources and methodologies used for censuses in the Member States should be fostered".

Article 5 (2) of the Census regulation foresees that the "Member States shall provide the Commission (Eurostat) with final, validated and aggregated data (…)". This

excludes the transmission of microdata. Although aggregated data are not necessarily protected against disclosure of sensitive data, the spirit of Article 5 (2) implies that no confidential data shall be transmitted to Eurostat.

Considerations 5 and 7 stipulate that the Statistical Law, respectively the European Statistics Code of Practice, constitute the framework for the Census regulation, both containing provisions on statistical confidentiality.

Consideration 6 recalls the regulations on the transmission of data subject to statistical confidentiality. This means that, if Member States transmit data they feel is subject to statistical confidentiality, Eurostat has to ensure the physical and logical protection and that no unlawful disclosure or non-statistical use occurs when Community statistics are produced and disseminated. However, the census regulation does not foresee the transmission of confidential census data from the Member States to Eurostat. In a broad sense, Consideration 6 reminds indirectly that everything must be done to avoid inadvertent disclosure of any confidential data.

In principle, the Task Force CENSDC followed up two major branches of thinking:

A recommendation on the pre-tabulation noise protection at the microdata level. This seems to have advantages in the context of both a national and a European dissemination of 2011 Census results. However, this protection can only be done at the NSI (National Statistical Institute) level and Eurostat would have no means of even verifying that such a protection has been executed.

A recommendation on post-tabulation protection (hypercube level). For the time being, the work is split into "cell suppression" and "post-tabulation noise protection". A simple solution would be to check which cells cannot be published (the so-called primary suppressions) and protect in addition a number of cells to prevent recalculations from the margins (the so-called secondary suppressions). However, the Task Force also considered whether synergies between these two methodologies are achievable — given that the objective is limited to preventing the identification of individuals, i.e. to prevent certainty about cell values in frequency tables. This prevention action should ideally take place with minimum information loss.

As the real data of the 2011 Censuses are of course not yet ready, test hypercubes of a few countries were being used by the Task Force CENSDC. In the two examples below two dimensional subtables of higher dimensional Italian test hypercubes are shown[2]. For obvious reasons variable names are replaced by names like var2, var3 and so on. In the first example (Figure 1) some cells have to be protected, but the confidentiality problems seem to be solvable. If even more cells contain no observations a proper protection strategy will probably lead to many cells without a real frequency score.

[2] These pictures were produced by Sarah Gießing (Destatis, Germany).

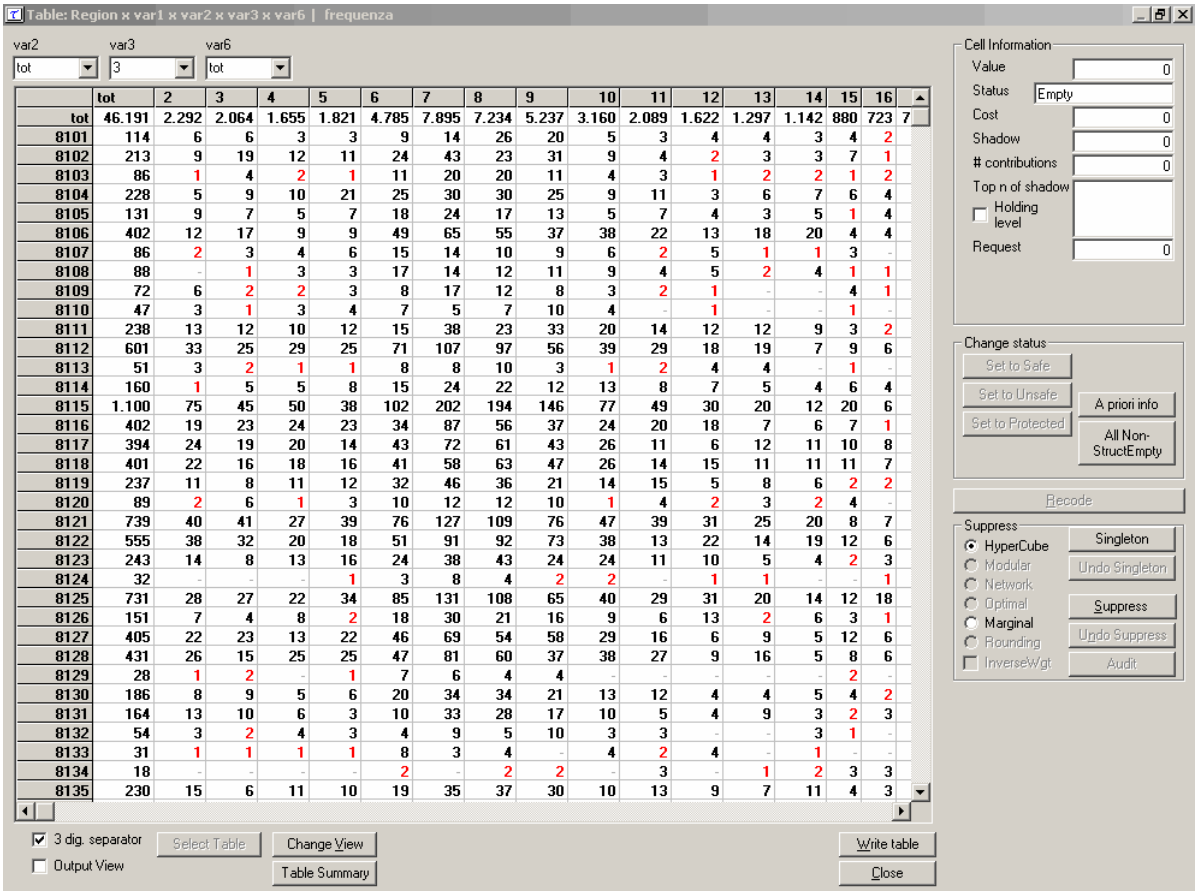

Figure 1. A solvable confidentiality problem.

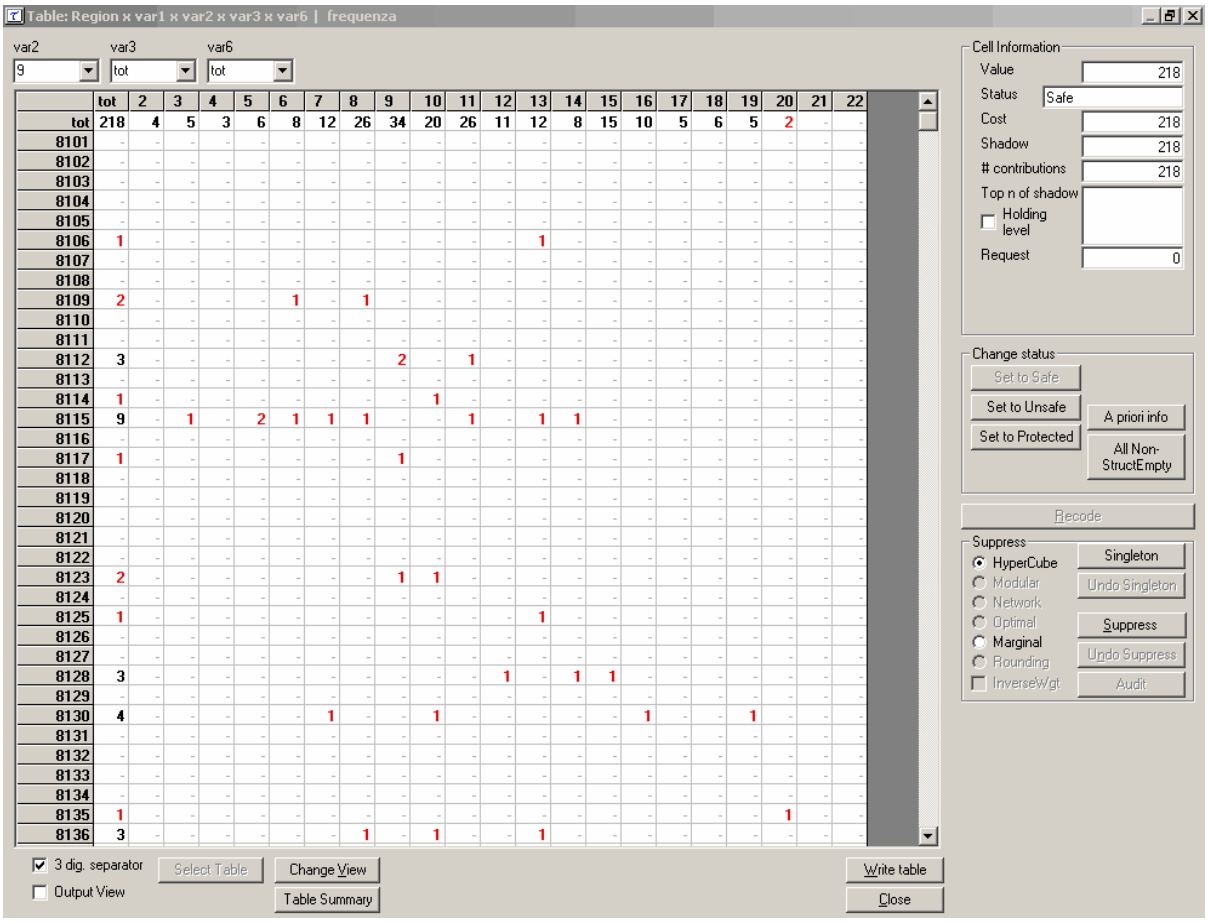

Figure 2. An unsolvable confidentiality problem.

In the second example (Figure 2) even all interior cells are zeroes or have to be suppressed. It is clear that this problem will occur more often for countries that make use of survey information (instead of complete enumeration or register information) for some of the 2011 Census variables.

What is the lesson we can learn from these pictures? For all Census hypercubes it will be important to verify whether most cells can be published. It does not make sense to produce and publish hypercubes with only or mainly zeroes and suppressions. Although confidentiality rules may differ between countries, this problem plays a role in all European countries. Larger countries tend to have more regions and thus face the same problem at a more detailed level as smaller Member States at their national levels.

The Task Force compared different confidentiality rules that countries applied and analysed the effects on the primary suppressions. To prevent recalculating primary suppressions from marginal totals some extra cells have to be suppressed. These extra suppressed cells are called secondary suppressions. A common method to decide on the secondary suppressions in an optimal way is the so-called hypercube method which is implemented in τ-ARGUS (Hundepool et al, 2010b) and used by many countries. The software package τ-ARGUS can be downloaded free of costs from the website http://neon.vb.cbs.nl/casc/. On that website also test data and the manual of the software can be found and downloaded.

9. 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.

It can be concluded that there is still a lot of research to be done in the field of Statistical Disclosure Control. In even years results of pure research in Statistical Disclosure Control are demonstrated in the so-called Privacy in Statistical Databases (PSD) meetings. In odd years Joint UNECE/Eurostat Work Sessions on Statistical Data Confidentiality are held where applied research in Statistical Disclosure Control is promoted. New versions of the ARGUS packages (that include results of the on-going research) are regularly released to the user community. Some of these new versions were part of the CASC project (Hundepool, 2001) and the CENEX on Statistical Disclosure Control (Hundepool, 2006).

New manuals for μ-ARGUS and τ-ARGUS (Hundepool et al, 2010a and b) are disseminated whenever new versions of the software are released. The software packages have been tested intensively as part of the CASC project and later on in the CENEX and ESSnet projects on Statistical Disclosure Control. Both manuals were of great help to the testers. In the newer versions of τ-ARGUS hierarchical tables

can be dealt with as well. In the newer versions 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. The most recent information is published at the CASC website: http://neon.vb.cbs.nl/casc.

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 was funded under the Fifth RTD Framework Programme. Many courses and conferences were organised, among other topics, about Statistical Disclosure Control. These activities stimulated the progress in the implementation of Statistical Disclosure Control methods and techniques in many different countries. Also in the CENEX and ESSnet projects on Statistical Disclosure Control a number of courses and conferences were organised.

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. SDC, CASC, AMRADS, CENEX and ESSnet projects) were subsidised by the European Union. Eurostat has stimulated the forming of consortia of researchers from Universities, NSIs and Market Research Bureaus for the Fifth RTD Framework Programme of the European Union. 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 international 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.

The remote facility has become a promising counterpart of the ‘traditional’ on-site facility. Concerning confidentiality issues, both facilities appear to be comparable. The remote facility allows researchers to perform their analyses on microdata from a computer at their own desk, so they can work any time they want. Moreover, no travelling is needed whenever they want to perform additional research.

The technical implementation of the remote facility tackles most of the confidentiality issues: the microdata remain at Statistics Netherlands, it is not possible to print or download any results and the final results will be checked for

confidentiality before being released to the researcher. So far, no real problems have been encountered with the facility. Both the performance of the system and the look and feel resemble that of working on a state of the art workstation. I.e., it feels like working on the own computer.

This facility can also be used to provide access to microdata files under contract. Currently, those kinds of microdata files are protected using statistical disclosure control methods as well as legal measures. These files are provided using CD-ROMs. Using the remote facility, these files do not leave Statistics Netherlands; hence the dissemination of the microdata is much more under control.

On the basis of the outcomes of the first two meetings of the Task Force CENSDC, and considering the timetable of the censuses, the current orientation is towards a simplified approach. As a clear and full agreement could not be reached by all Member States on a proposed methodology, the EU implementing regulation regarding the statistical data to be transmitted to Eurostat does not contain any provision on the disclosure control method to be applied to census data. This means that the countries can send cells of the hypercubes blanked for confidentiality reasons.

However, it should be taken into account that not all topics may be considered by the Member States as confidential. For instance, thoughts should be given to assess whether characteristics like sex or age have to be subject to disclosure control according to the national regulations. It may well be indeed that some topics are more "sensitive" than others. Considering that the topics listed in the EU regulation – thus mandatory for the Member States - are in fact the CES core topics, which do not include any characteristic on - e.g. - health or income, it may be worthy to consider – already at national level and respecting the national provisions – whether confidentiality applies to all topics and all enumeration units[3].

The work to find a harmonised SDC approach for the Census 2011 hypercubes is continued. This approach is to be recommended for adoption to the countries, taking into account to the possible extent the constraints expressed above. Whether this will lead to an increased comparability of the census data will depend also on the degree of flexibility the national statistical offices will apply in considering a "fit-for-all" proposal. In any case, lots of discussions in this domain can be expected.

References

European Commission, 1990. Council Regulation No 1588/90 of 11 June 1990 on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities. Official Journal, L151, pp. 1-4.

[3] If confidentiality is clearly an issue for persons, this is not straightforward for data on other kinds of enumeration units, like households, dwellings, etc.

European Commission, 1997. Council Regulation (EC) No 322/97 of 17 February 1997 on Community Statistics. Official Journal, L052, pp. 1-7.

European Commission, 2008. Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on population and housing censuses. Official Journal of the European Union, L218, pp. 14-20.

European Commission, 2009a. Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics. Official Journal of the European Union, L87, pp. 164-173.

European Commission, 2009b. Commission Regulation (EC) No 1201/2009 of 30 November 2009 implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses as regards the technical specifications of the topics and of their breakdown. Official Journal of the European Union, L329, pp. 29-68.

European Commission, 2010a. Commission Regulation (EU) No 519/2010 of 16 June 2010 adopting the programme of the statistical data and of the metadata for population and housing censuses provided for by Regulation (EC) No 763/2008 of the European Parliament and of the Council. Official Journal of the European Union, L151, pp. 1-13.

European Commission, 2010b. Commission Regulation (EU) No 1151/2010 of 8 December 2010 implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses, as regards the modalities and structure of the quality reports and the technical format for data transmission. Official Journal of the European Union, L324, pp. 1-12.

European Statistics Code of Practice, 2011. http://ec.europa/eu/eurostat/quality

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., 2006. "The ARGUS software in CENEX", *Privacy in Statistical Databases, Lecture Notes in Computer Science 4302,* Berlin: Springer-Verlag (J. Domingo-Ferrer and L. Franconi, eds.), pp. 334-346.

Hundepool, A., A. van de Wetering, R. Ramaswamy, L. Franconi, S. Polettini, A. Capobianchi, P.P. de Wolf, J. Domingo, V. Torra and S. Giessing, 2010a. *μ-ARGUS, user's manual, version 4.3*, The Hague, The Netherlands: Statistics Netherlands.

Hundepool, A., A. van de Wetering, R. Ramaswamy, P.P. de Wolf, S. Giessing, M. Fischetti, J.J. Salazar, J. Castro and P. Lowthian, 2010b. *τ-ARGUS, user's manual, version 3.4*, The Hague, The Netherlands: Statistics Netherlands.

ISI, 1985. Declaration on Professional Ethics: http://www.isi-web.org/about/ethics1985

ISI, 2010. Declaration on Professional Ethics: http://www.isi-web.org/images/about/Declaration-EN2010.pdf

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.

Pink, B. et al, 2009. Principles and Guidelines on Confidentiality Aspects of Data Integration, UNECE United Nations Economic commission for Europe: http://www.unece.org/stats/publications/Confidentiality_aspects_data_integration.pdf

Trewin, D. et al, 2007. Principles and Guidelines of Good Practice for Managing Statistical Confidentiality and Microdata Access, UNECE United Nations Economic commission for Europe: http://www.unece.org/stats/documents/tfcm/1.e.pdf

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.