CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Access to different kinds of Statistics Netherlands' microdata.

Eric Schulte Nordholt (Statistics Netherlands)

e.schultenordholt@cbs.nl

Abstract

At Statistics Netherlands different options exist to get access to microdata. In the eighties of last century hardly any access was given to researchers. Gradually the number of facilities has grown. Since the nineties of last century microdata are released as public use file (PUF) or microdata under contract (MUC). MUCs are also called Scientific Use Files. More detailed data can be analysed via the on-site and remote access facilities. Datasets used in those facilities are called Secure Use Files. Special co-operation used to take place via remote execution but nowadays it has become more common to define so-called co-operation projects in which Statistics Netherlands works closely together with one or more external partners. Such projects only run under strict conditions and if they are profitable for all partners involved. As different access facilities in the Netherlands have different confidentiality risks, different Statistical Disclosure Control (SDC) rules are applied for different facilities. The more detail in the data provided implies the stricter the access is organised. A special unit within Statistics Netherlands is responsible for giving access to the different microdata for research purposes.

Access to different kinds of Statistics Netherlands' microdata

Eric Schulte Nordholt [1]

[1] Statistics Netherlands, P.O. Box 24500, 2490 HA The Hague, The Netherlands, e.schultenordholt@cbs.nl*

Abstract. At Statistics Netherlands different options exist to get access to microdata. In the eighties of last century hardly any access was given to researchers. Gradually, the number of facilities has grown. Since the nineties of last century microdata are released as public use file (PUF) or microdata under contract (MUC). MUCs are also called Scientific Use Files. More detailed data can be analysed via the on-site and remote access facilities. Datasets used in those facilities are called Secure Use Files. Special co-operation used to take place via remote execution, but nowadays it has become more common to define so-called co-operation projects in which Statistics Netherlands works closely together with one or more external partners. Such projects only run under strict conditions and if they are profitable for all partners involved. As different access facilities in the Netherlands have different confidentiality risks, different Statistical Disclosure Control (SDC) rules are applied for different facilities. The more detail in the data provided implies the stricter the access is organised. A special unit within Statistics Netherlands is responsible for giving access to the different microdata for research purposes.

Keywords. access; confidentiality; microdata; Statistical Disclosure Control (SDC)

1. Introduction

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 (Hundepool et al, 2012).

The information from statistics becomes available for the public in tabular and microdata form. Historically, only tabular data were available and National

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

Statistical Institutes (NSIs) had a monopoly on the microdata. Since the eighties of last century the PC revolution leaded to the end of this monopoly. Now other users of statistics have the possibility of using microdata as well. These microdata can be conveyed with CD-ROMs, USB sticks and other means. Recently, in several countries also other possibilities of getting statistical information have become more popular. With that technique researchers can get access to data that remain in a statistical office and can execute set-ups without having the microdata 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 or in their own institutes via remote access.

In the eighties of last century less and less microdata became available for research purposes in the Netherlands because of the grown risk awareness. That period was classified by some people as a cold war between Statistics Netherlands and the academia. Since that time microdata access facilities have improved enormously in the Netherlands.

This paper gives some background on microdata availability in the Netherlands. In sections 2 and 3 the releases of public use files (PUFs) and microdata under contract (MUCs) are described. Both PUFs and MUCs can be produced using the software package μ-ARGUS (Hundepool et al, 2014). Nowadays, 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 4. Remote facilities are discussed in section 5. Criteria for co-operation projects are given in section 6. Finally, some conclusions are drawn in section 7.

2. The release of public use 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 under contract (that are described in the next section) are released. For a few surveys also public use files are produced. The software package μ-ARGUS (Hundepool et al, 2014) is of help in producing both public use files and microdata under contract.

For the public use files (also known as PUFs) 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.

As this set of rules leads to heavily protected files, researchers are often not very interested in these files. They prefer the less strictly protected microdata under contract that are described in the next section. For educational purposes public use files play an interesting role. Some of these files (based on data from the Employee Register, Labour Force Survey and Housing Survey) are being used during mathematics lessons in Dutch high schools. More public use files for educational purposes are foreseen in the near future.

A positive exception where public use files play an important role in research is the IPUMS (Integrated Public Use Microdata Series) project, see http://www.ipums.org/international. Protected samples of the microdata of the Dutch censuses of 1960, 1971, 2001 and 2011 were disseminated via this project. These datasets contain a number of demographic and economic variables. The microdata of the first three years have also been archived at the institute DANS (Data Archiving and Networked Services) in the Netherlands, see http://www.dans.knaw.nl/en/.

Statistics Netherlands selected the Dutch censuses of 1960, 1971, 2001 and 2011 to be part of the IPUMS project. The censuses of 1960 and 1971 were traditional censuses, of which most of the micro data records have been recovered. The 2001 and 2011 censuses were virtual censuses, which means that they were composed of available register data and existing surveys at Statistics Netherlands. Unfortunately, this results in not having all variables available for all individual records. As a consequence it was not possible to release the complete set of microdata, but only a sample of individual personal records for which all demographic and economic variables are available. The sampling fractions vary between 1 percent and 2.5 percent of the total population in the Netherlands.

The first stage in the cooperation of Statistics Netherlands in the IPUMS project has been the release of the 2001 census microdata. The selection of variables of the 2001 census has been leading in the selection of the variables of the censuses of 1960 and 1971. Due to differences in variable definitions, classifications and variable availability over time, differences among the three microdata sets remain. For 1960 and 1971 anonymised balanced 1 percent samples of the total population were released. Some more background and documentation of the historic 1960 and 1971 censuses can be found at the following web site: http://www.volkstellingen.nl/en/documentatie/. At a later stage 2011 census microdata were added to the IPUMS project. This concerned a 2.5 percent sample of the total population on Census Day (1 January 2011).

3. The release of microdata under contract

For the microdata under contract (also known as MUCs or Scientific Use Files) 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.

The software package μ-ARGUS (Hundepool et al, 2014) is of help to identify and protect the unsafe combinations in the desired microdata file. Thus the rules 7 and 8 for the public use microdata files and rule 2 for the microdata for researchers 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.

Public use microdata files contain much less detailed information than 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 microdata under contract, 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.

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 is now included in the Statistics Netherlands act 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 under contract may be released from these surveys. Statistics Netherlands can therefore provide two kinds of information from these surveys: tables and public use files.

4. On-site access

Global recoding and local 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.

An 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 public use files or microdata under contract. 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. The researchers can apply standard statistical software packages and also bring their own programmes. 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. At the on-site 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.

Microdata Services, a unit within Statistics Netherlands, runs the on-site facility of the office (https://www.cbs.nl/en-gb/onze-diensten/customised-services-microdata/microdata-conducting-your-own-research). The researchers who work on-site on Statistics Netherlands' data have to take the rules of Microdata Services into account. The most important rules are:

- researchers must be associated with a 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. Given the current Covid crisis the on-site facility is currently closed.

5. Remote access

Finally, an option is to allow remote access. This access modality combines the advantage of licensing and microdata 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. Statistics Netherlands' unit Microdata Services 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.

In the past 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 on-site facility without having to travel to the premises of Statistics Netherlands.

Only a selected group of researchers working at universities and other research institutes is allowed to utilize this facility. The remote access facility is making use of a VMware connection. To ensure that the researcher who is trying to connect to the facility is indeed the intended person, TAN (Transaction Authorisation Number) codes are sent to the mobiles of the researcher.

The network that is used by the facility is not connected to the production network of Statistics Netherlands. 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 via 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.

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 by 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 stimulated to write their complete papers on the 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.

6. Co-operation projects

Special co-operation used to take place via remote execution but nowadays it has become more common to define so-called co-operation projects in which Statistics Netherlands works closely together with one or more external partners. Such projects only run if they are profitable for all partners involved. Statistics Netherlands has co-operation projects with a number of research institutes to do scientific and statistical research together. Co-operation partners of Statistics Netherlands cannot be commercial companies as otherwise they would get a benefit compared to their competitors. Moreover, one has to realise that the results of the co-operation projects will be published for the general public and that is often unwanted by commercial companies.

In recent years Statistics Netherlands has invested a lot in good relations with a growing number of co-operation partners. This has given Statistics Netherlands a stronger and more relevant position in Dutch society. More and new kinds of output have been produced without many extra costs. Ministries are excluded from co-operation projects as they often want individual information (instead of statistical information) for their administrative tasks. Separate research departments of ministries form an exception to this rule. With whom Statistics Netherlands starts co-operation projects depends on capacity and needs and is thus a matter of policy. In some projects Statistics Netherlands is (also) responsible for drawing the sample.

Different kinds of co-operation can be distinguished. In some of these projects complete microdata files are shared, whereas in other projects only protected files are made available to other partners in the project. Such specially protected files are called Controlled Circulation Files (CCFs). A few standard variables in a limited number of regular categories are included in these CCFs (e.g. gender in two categories and age in a few age groups). The product of the number of categories of the different identifying variables is limited to reduce the risk of identification. For a sample survey of a certain reference period either a MUC or a CCF is made. If both would be made the protection of each could be broken by combining information from the two protected files. It is clear that it is easier to produce a CCF than to produce a MUC. However, rare combinations are better protected in a MUC. Both kinds of files can only be made available under contract to bonafide researchers.

Concerning the output three different situations exist:

1. Only Statistics Netherlands is publishing the output whereby credits are given to the co-operation partner(s).
2. Also other partners publish output, but they have to stick to the protection rules of Statistics Netherlands.

3. Other partners also publish and have their own rules (respecting the European General Data Protection Regulation).

Depending on the actual research to be conducted and the relevant partner(s) a specific contract is drawn up for every co-operation project. In some of these contracts rules are given about giving microdata access to organisations outside the co-operation project. After an agreement has been reached the contract, in which the responsibilities and rules are made clear, is signed by all partners. The policy of Statistics Netherlands is to limit the number of co-operation projects in the sense that remote access is the preferred option.

7. Conclusions

In this paper access facilities to microdata have been described. They 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.

Public use files and microdata under contract can be produced with the software package μ-ARGUS (Hundepool et al, 2014). Microdata under contract are currently protected using statistical disclosure control methods as well as legal measures. The ARGUS packages have moved towards interfaces with several state of the art engines produced by statisticians from many different countries. More information is published at the following websites: http://neon.vb.cbs.nl/casc and https://github.com/sdcTools.

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 with data that are stored on the own computer.

In recent years Statistics Netherlands has invested a lot in good relations with a growing number of co-operation partners. This has given Statistics Netherlands a stronger and more relevant position. More and new kinds of output have been

produced without many extra costs. After an agreement has been reached the contract, in which the responsibilities and rules are made clear, is signed by all partners.

References

Hundepool, A., P.P. de Wolf, J. Bakker, A. Reedijk, L. Franconi, S. Polettini, A. Capobianchi and J. Domingo, 2014. *μ-ARGUS, user's manual, version 5.1*, The Hague, The Netherlands: Statistics Netherlands.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer and P.P. de Wolf, 2012. *Statistical Disclosure Control, Wiley Series in Survey Methodology*, Chichester, United Kingdom: Wiley. ISBN 978-1-1199-7815-2