WP. 50
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (ix): Statistical disclosure limitation for table and analysis servers: how to make outputs of modern data access infrastructures safe

De facto anonymity in results

Prepared by Tim Hochguertel and Emanuel Weiss, Federal Statistical Office, Germany

De facto confidentiality of results

Tim Hochgürtel and Emanuel Weiss*

* Federal Statistical Office Germany,
e-mail: tim.hochguertel@destatis.de; emanuel.weiss@destatis.de

Abstract: The German Law on Statistics for Federal Purposes distinguishes between different levels of confidentiality. The strongest level is called absolute confidentiality. Absolute confidential microdata and results must be modified in a way that makes it impossible to re-identify a single observation unit. Unfortunately, this comes at the cost of a reduced potential for empirical analyses. In addition to absolute anonymous microdata the German Research Data Centers offer microdata that are de-facto confidential. Concerning this level of confidentiality, the costs of a de-anonymization have to be higher than the utility of a re-identification. According to the German law access to de-facto confidential microdata is restricted to scientists employed by scientific institutions. However, published results based on de-facto confidential microdata have to be absolute confidential. There exists plenty of literature dealing with de-facto confidential microdata. In contrast there is no theoretical foundation for de-facto confidentiality of results. Hence, our paper develops first conceptual ideas of de-facto confidentiality in results and whether the application of the concept of de facto confidentiality to results could reduce the waiting period for researchers when making use of Remote Execution.

1. Access to microdata in Germany

In the last years, the demand for access to microdata of official statistics has significantly increased. The Statistical Offices responded to this increasing demand with the establishment of Research Data Centers. Since 2001, the access to microdata of German official statistics has been possible for external users via those Research Data Centers.

A big challenge for the providers of data access is to balance the interest of researchers to get wide access to microdata and the need of the public for data confidentiality. As a consequence, microdata are accessible via four different ways: Public Use Files, Scientific Use Files, Safe Centres and Remote Execution.
The mentioned methods of data access differ in the level of confidentiality and the group of users who are allowed to make use of the different types of access. The microdata are modified using Statistical Disclosure Control (SDC) techniques, which reduce the risk of disclosing information on individuals or entities. The degree of confidentiality defines the level of modification of the microdata.

Table 1: Different ways of microdata access via Research Data Centers

Way of Access	**Level of Confidentiality**	**Authorized User**
Public Use Files	Absolute confidentiality	Everyone
Scientific Use Files (Off-Site)	De facto confidentiality	German scientific institutions
Safe Centers (On-Site)	De facto confidentiality	German and foreign scientific institutions
Remote Execution	Formal confidentiality	Everyone

The German law distinguishes between three levels of confidentiality of microdata: absolute confidentiality, de facto confidentiality and formal confidentiality.
The top level of confidentiality is **absolute confidentiality**. A microdata file is modified by SDC techniques in such a way that re-identifying a single statistical unit is absolutely impossible. As a consequence, a strong intervention by SDC techniques is necessary to obtain this level of confidentiality.

The second level of confidentiality is **de facto confidentiality**. De facto confidentiality means that the costs of a re-identification of a microdata file are higher than the benefit of this re-identification. However, a re-identification is not completely impossible. Because of the high costs of a re-identification, no rational intruder will try to start an attack. The costs are mainly determined by working hours which an intruder would have to spend for a re-identification. The benefit results from the knowledge, which the intruder generates by the re-identification. In general, de facto confidential files require less intervention by SDC techniques than absolute confidential files.
The lowest level of confidentiality is called **formal confidentiality**. All variables which allow a direct identification are deleted (e.g. name, address, tax identification number). Apart from this, no further SDC techniques are required to generate a formal confidential microdata file.

The different ways of access to microdata are restricted as follows. **Public Use Files** are not limited to a special user group. Anyone has the possibility to request Public Use Files. A user obtains a CD with an absolute confidential microdata file. However, most users do not prefer to work with Public Use Files, because of the strong intervention by SDC techniques for generating absolute confidential files: variables have been deleted or strongly perturbed. This strong intervention comes with a relative low level of information, reflecting the trade-off between confidentiality loss and information loss.

In contrast, de facto anonymised **Scientific Use Files** are only available for scientific institutions. The Research Data Centers provide access to microdata for researchers in the Safe Centers and via anonymised microdata files. In the latter case, users of Scientific Use Files obtain a CD with the microdata. However, the intervention is less strong than in the case of Public Use Files. Therefore, a de facto confidential file contains more information (e.g. more variables and less perturbed variables) than an absolute confidential file.

Another way of access to de facto anonymised microdata for scientific users is possible via **Safe Centers**. By visiting a Safe Center, a researcher obtains access to microdata on a save workstation dedicated for this purposes and located in facilities of the Statistical Office. The

microdata do not leave the Statistical Office. Standard software for econometric and statistic analysis (SPSS, SAS, Stata and R) is available on these workstations, but there is no access to the internet. The protected environment of the Safe Center enables the user to obtain a microdata file which includes more information than a Scientific Use File, but the level of de facto anonymity is still achieved.

If a user wants to work with data that are not modified by SDC techniques, he has the possibility to make use of **Remote Execution**. As in a Safe Center, the data do not leave the Statistical Office. By this way of using microdata, the user sends codes written in SPSS, SAS, Stata or R to a Research Data Center. A staff member of the Research Data Center analyzes the formal confidential data via the codes of the user. The results of this analysis have to be checked twice by at least two members of the staff. The user obtains the results of the analysis, if they do not allow assigning information to a single observation. Only results which are absolutely confidential leave the Statistical Office. To guarantee the confidentiality, the staff of the Research Data Center uses a cell suppression technique to prevent the publication of sensitive information.

Using Remote Execution has an important advantage: The user can work with the "real" microdata. The data is not perturbed by SDC techniques. Therefore, all results from the analysis are "real" results. No modification of microdata has an influence on the results. That is why most users prefer to work via Remote Execution.
Nevertheless, Remote Execution is connected with some disadvantages as well. The user has the possibility to send codes to the Research Data Center. As a result, he does not have direct access to the microdata, which makes it difficult to create an appropriate code. The development of a code is supported by the Research Data Center by providing a so-called "data structure file". Currently, this file does not allow a semantic test of the code. The data structure file enables only a syntactic check without giving any suitable information about the final results. Experience has shown that the users have to send the codes to the Research Data Centers several times until the results are satisfying (Zuehlke 2005).
Therefore, the Remote Execution is very time-consuming for both, data producers and data users. The data producers have to invest a lot of time for the manual output checking of every single result. From the user's perspective, the waiting period between the sending of a code and the receipting of the results can take up to several weeks.

The long waiting period in Remote Execution shows that there is a strong need for a progress of the user service. In cooperation with other data producers and representatives from the scientific community, the Research Data Center of the Federal Statistical Office participates in the project infinitE. The goal of infinitE is to make improvements in microdata access (Hochguertel 2011).

In the course of infinitE it should be considered if the concept of de facto confidential microdata can be fruitfully used for the check of outputs of Remote Execution. At the moment, the check of the outputs has to lead to absolute confidential results. One idea of infinitE is to evaluate, whether an output check can be realized faster if the output is checked for "de facto confidentiality" only. Users often produce a huge number of results. However, only a very small fraction of the results is published. Although the most results which the users generate in Remote Execution are intermediate results which are not published, every result is checked for absolute confidentiality. If it turns out that a check for de facto confidentiality is less time-consuming, users could obtain the results as de facto confidential intermediate results solely for their own use. Only the results which the user likes to publish would have to be checked for absolute confidentiality.

2. De Facto confidentiality

Since 1987 the concept of ”de facto confidentiality“ has been part of the German Statistical Law, § 16.6 (Bundesstatistikgesetz BStatG). This paragraph guarantees scientific institutions a privileged access to microdata. In contrast to other user groups, scientific institutions have access to de facto confidential data. All other user groups only obtain data with a stronger level of confidentiality.
At the moment, the concept of de facto confidentiality is only applied to microdata. De facto confidential microdata files are offered as Scientific Use Files and via Safe Centers as mentioned above. Concerning de facto confidentiality of microdata exists experience for twenty years (Mueller et al. 1991).
De facto confidentiality is not only a result of microdata manipulation. There exist many other components of protection. All components together guarantee de facto confidentiality:

1. Only scientific institutions have access to de facto confidential data.
2. The scientific institutions have to make a contract with the Statistical Office.
3. This contract contains a punishment in the case of violating the rules of confidentiality.
4. The de facto confidential data may only be used for a special project, which is named in the contract.
5. The access to de facto confidential data is limited in time.
6. The scientific institutions have to name all researchers who work with the de facto confidential data.
7. All these researchers have to complete a personal obligation.
8. It is forbidden to allow other people access to the data (in the case of Scientific Use File).
9. After the end of the project, the users have to delete all the files (in the case of Scientific Use File).
10. Modifications of the data file by SDC techniques.

With respect to de facto confidentiality of microdata, the costs of a re-identification of a single statistical unit of a data file have to be higher than the benefit of the re-identification. That is why a modification of a microdata file has to take into account the existing prior knowledge about the statistical units of a microdata file. For example, one potential source of prior knowledge can be found in commercial enterprise databases.
To quantify the level of protection of a microdata file, a matching procedure is an adequate technique. For this purpose two files are necessary: the first file includes the names of all statistical units and the existing prior knowledge about them. The second file consists of the de facto confidential microdata. A number of overlapping variables (key variables) exist in both files. The variables in the confidential microdata file, which are not overlapping, are called “sensitive variables”. The files are matched by making use of “nearest neighbour”-methods. The measurement of confidentiality is determined by the rate of correct matches and the quality of the revealed sensitive variables (see Lenz 2010, p. 39-54).

3. De facto confidentiality in results

One goal of the project infinitE is to find an answer to the question whether the concept of de facto confidentiality can be fruitfully applied to the output checking of results produced via Remote Execution. The idea of applying the concept of de facto confidentiality is based on the following observation: almost every user generates a huge number of results. All of the

results have to be checked for absolute confidentiality in a time-consuming process. However, the users publish only a small fraction of these results. As a consequence, a large part of the generated outputs are unnecessarily checked. Therefore, it could be enough to check the results for de facto confidentiality as long as the results are only intermediate and not published. Only the final results, which the users like to publish, would have to be checked for absolute confidentiality.
The idea of de facto confidentiality of results might be helpful for improving the data access service, if it is possible to find other criteria for output checking which allows a faster checking of the results. In contrast to de facto confidentiality of microdata, there is no experience with de facto confidentiality of results.
Several components of the concept of de facto confidentiality of microdata could be used for producing de facto confidentiality of results. The items 1 to 9 from the list in chapter 2 can be adopted for the production of de facto confidentiality of results.
As in the case of de facto confidentiality, it is necessary to offer the access to de facto confidential results only for scientific institutions. There is no legal basis for the access to de facto confidential data by non-scientific institutions. Furthermore, it is essential to sign a contract with the scientific intuition. The contract contains the punishment rules in case of a violation of the agreement. All intermediate results are only for the user's own purpose. After the end of the project all intermediate results have to be deleted by the user.
As in the case of de facto confidentiality in microdata, an intervention with respect to the output is necessary. Otherwise the disclosure of information on single statistical units would be too easy for an intruder who has suitable prior knowledge. For an application of de facto confidentiality of results, it is necessary to identify rules for the process of output checking. These rules have to guarantee in combination with the items 1 to 9 of the list in chapter 2 that the costs of a disclosure are higher than its benefit.
Chapters 4 and 5 develop first ideas about the application of the concept of de facto confidentiality to results. Chapter 4 discusses frequency tables, chapter 5 deals with magnitude tables.

4. De facto confidentiality of frequency tables

4.1. Rules for absolute confidentiality of frequency tables

Currently, the Statistical Offices realizes output checking of all frequency tables generated by a user. The output checks are subject to the following rules:

1. Primary suppression of small frequencies: if only 1 or 2 respondents contribute to a frequency, the corresponding cell has to be suppressed.
2. Primary suppression of group attribute disclosure: if all respondents of an identifiable group fall into one category for a particular variable the corresponding cells have to be suppressed (Hundepool et. al. 2010, p. 168).
3. Check disclosure risk by differencing or linking: differencing and linking enables the intruder to gain additional information using multiple overlapping tables. For example, if a frequency table exists for Germany and another table for Western Germany with the same variables, an intruder can generate the table for Eastern Germany by subtracting the two existing tables. As a consequence, it has to be checked if the tables are linked by common variables in such a way that disclosure is possible (Hundepool et. al. 2010, p. 169).

Beyond that, the primary suppression has to be complemented by a secondary suppression (Cox 2001, p. 177-181). In general, it is very time-consuming to find manually an adequate pattern of secondary suppression.

4.2. Rules for de facto confidentiality of frequency tables

The concept of de facto confidentiality should improve the user service by reducing the long waiting period in Remote Execution. That is why it is essential to identify other methods of output checking. These rules have to guarantee sufficient data confidentiality and have to allow a faster check of outputs generated via Remote Execution.
If the user obtains the results as de facto confidential results, the costs of the disclosure of a single observation unit have to be higher than the benefit of the disclosure. The rules regarding de facto confidentiality have to prevent a data attack that could be realized without higher costs. An intruder can use a frequency table for an attack. For a successful attack in frequency tables, the following conditions have to be fulfilled (Duncan 2011, p. 30 and Mueller 1991, p. 92):

- A data file includes the record set of a demanded observation unit and an intruder is aware of this.
- The intruder possesses enough prior knowledge about the observation unit of interest.
- It is possible to generate a cross table with a key variable and a sensitive variable. All observations that have the same value in the key variable as the demanded observation unit have the same value in the sensitive variable.

If the mentioned preconditions are fulfilled, the intruder can match the value of the sensitive variable to the demanded observation. Therefore, the application of the second rule from chapter 4.1 is necessary to guarantee de facto confidentiality in results.

An intruder can camouflage the attack. So it is possible to split the attack in several tables and to obtain the information of a single unit from the interaction of these several tables. Therefore, the application of the third rule of chapter 4.1 is essential.

A check for de facto confidentiality can abandon the first rule of the chapter 4.1. A frequency of 1 or 2 does not contain a higher risk of disclosure for the regarding observations (Smith and Elliot 2008, S. 41).

5. De facto confidentiality of magnitude table

5.1. Concentration rules for absolute confidentiality of magnitude tables

The output checking of magnitude tables differs from the output checking of frequency tables. An additional check for concentration is essential with respect to magnitude tables. This concentration check has to prevent a so-called “approximately disclosure”.
To avoid a good estimation of a value of an observation unit, sensitivity rules are needed. If a cell total is dominated by the contribution of one observation unit, an intruder could use the total as estimation for this observation. For example, ten enterprises provide together a certain total of turnovers. Let’s consider nine enterprises to be small and one to be big. The latter one contributes 95 percent to the total of turnovers. An intruder can use the total of turnovers as estimation for the value of the turnover for the big enterprise. Following this strategy, the intruder obtains an estimation that comes very close to the original value.
Obviously, the respondent with the second largest contribution to the cell total is in the best situation to obtain a good estimation. He can use the difference between the total and his own value as an estimator for the turnover of the biggest enterprise.

To avoid such scenarios, the Statistical Offices make use of rules like the (n,k)-dominance rule to identify sensitive cells. A cell is considered as unsafe, when the sum of the n largest contributions exceeds k percent of a cell total. Originally, the value 2 is used as the parameter n by the check to absolutely confidentiality. All unsafe cells have to be suppressed.
By using n = 2 for the (n,k)-dominance rule, an intruder will overestimate the single value from the biggest contributor at minimum by $100\ \frac{100\ -\ k}{k}$ percent even if he has the information of the second biggest value (Gießing 1999, p. 8).

5.2. Concentration rules for de facto confidentiality of magnitude tables

Concerning de facto confidentiality of results, it is not necessary to assume that an intruder can use the own contribution to a given cell total to estimate the value of the biggest contributor. Therefore, an application of an (1,k)- dominance rule would lead to a suitable level of confidentiality. The user obtains the results classified as intermediate results. It is not allowed to publish these results. Obviously, it is not a realistic scenario to assume that a scientist has information about the second biggest contributor.
The (1,k)- dominance rule guarantees that totals, which are based on only one contributor will be classified as unsafe. Therefore, the user obtains only such values as intermediate magnitude results that are based on two or more observation units.

6. Conclusion

If an output check has to guarantee de facto confidentiality of results, a modification of the rules for output checking is necessary. As we have shown, a few rules can be abandoned or changed. But the effect on the duration of output checking is probably quite low. By checking for de facto anonymity, it is not longer necessary to suppress cells with small frequencies, but other aspects of output control have to be performed. In the course of a manual checking of a given output, it is time-consuming to find a suitable secondary suppression pattern.
Furthermore, with respect to de facto confidentiality it is still necessary to make use of the (n,k)-dominance rule.
As long as statistical disclosure control is realized by cell suppression techniques, the process of output checking remains a time-consuming affair even if one checks for de facto instead of absolute confidentiality.

7. References

Cox, Lawrence H. 2001: *Disclosure Risk in Tabular Economic Data*, in: Doyle, Pat et. al. (eds.): Confidentiality, Disclosure and Data Access, Elsevier, Amsterdam, p. 167-183.

Duncan, George / Elliot, Mark / Salazar-González Juan-José 2011: *Statistical Confidentiality. Principles and Practice*, Springer, New York.

Gießing, Sarah 1999: *Statistische Geheimhaltung in Tabellen*, Statistisches Bundesamt (eds.): Methoden zur Sicherstellung statistischer Geheimhaltung, Metzler Poeschel, Stuttgart, p. 6-26.

Hochguertel, Tim 2011: *Improvement of data access. On the way to Remote Data Access in Germany*, http://isi2011.congressplanner.eu/pdfs/950948.pdf.

Hundepool, Anco et. al. 2010: *Handbook on Statistical Disclosure Control. Version 1.2*, http://neon.vb.cbs.nl/casc/.%5CSDC_Handbook.pdf.

Lenz, Rainer 2010: *Methoden der Geheimhaltung wirtschaftsstatistischer Einzeldaten und ihre Schutzwirkung*, StatistischesBundesamt, Wiesbaden.

Mueller, Walter et.al. 1991: *Die faktische Anonymität von Mikrodaten*, Metzler Poeschel, Stuttgart.

Smith, Duncan and Elliot, Mark 2008: *A Measure of Disclosure Risk for Tables of Count*, in: Transactions on Data Privacy 1, p. 34-52.

Zuehlke, Silvia et. al. 2005: *The research data centres of the Federal Statistical Office and the statistical Offices of the Laender*, FDZ-Arbeitspapier Nr. 3, http://www.forschungsdatenzentrum.de/publikationen/veroeffentlichungen/fdz_arbeitspapier-03.pdf.