Improvement of data access. On the way to Remote Data Access in Germany

Hochguertel, Tim
Federal Statistical Office, Research Data Center
Gustav-Stresemann-Ring 11
65189 Wiesbaden, Germany
E-mail: tim.hochguertel@destatis.de

ABSTRACT

Since a couple of years access to micro data in Germany is possible through different ways. The researcher can use so-called Scientific Use Files in his own institution. Even for business micro data, the researcher can visit the safe centre or can use remote data execution. User needs are developing towards on-site data access, because researchers would like to work with original data mostly.
This leads to a higher burden for the employees of the Research Data Centre (RDC) as for remote data execution they have to apply the analysis programs and have to deal with the manual output checking of the results. Also the researchers themselves have to wait longer to get their results. The project "An informational infrastructure for the E-Science Age" (infinitE) deals with the improvement of Remote Access in Germany. The project aims to find solutions for better Remote Access in Germany through so-called data structure files and (automatic) output checking procedures.
To reduce the expenditure of time, the Research Data Centre develops in the project infinitE in cooperation with partners data structure files, which should allow the user a syntactic and semantic test of the own code. Based on this semantic and syntactic structure file, in the future it should be possible for the user to obtain hints for the final results. This syntactic and semantic structure file should reduce the frequencies of sending codes from the users to the Research Data Centre. In the project "An informational infrastructure for the E-Science Age" the Research Data Centre develops in cooperation with partners this syntactic and semantic data structure file.
A second goal of infinite is the development of fundamentals for a Remote Data Access in Germany. A full automated Remote Access is a vision for the future. In countries with comparable legal situation, a full automated Remote Access could not be realised until now. For a full automated Remote Access different technical, legal and methodical solutions are necessary.
At the moment there are no tools for an automatic output checking available, which guarantee absolute anonymity for the results of a user. One possible solution for this problem is to generate results automatically which are de-facto anonym. Users can obtain these de-facto anonymous results as intermediate results. Only the results, which the user wants to publish, have to be checked for absolute anonymity.

1. Current access to micro data

Since 2001 the Research Data Centre of the Federal Statistical Office offers the possibility of working with micro data from official statistics. Micro data users are offered several ways of access to the data. This different ways are distinct in their level of anonymisation of the micro data.

Scientific users can obtain micro data on a CD as a Scientific Use File, which are de facto anonymised. De facto anonymisation means, that the costs of a de-anonymisation are much higher than the benefit of a re-

identification of a single observation. Another way of micro data access for scientific users is the safe centre. The user has the possibility to analyse de facto anonymised micro data on a workstation in the accommodation of the Research Data Centre.

Non-scientific users can work with Public Use Files. A Public Use File is an absolutely anonymised file, were a re-identification of a single observation of a statistic is impossible. Just as with the Scientific Use File, the users obtain a micro data CD.

By working via remote data execution, the users have the possibility to work with micro data, which are only formal anonymised. Formal anonymisation requires not more than deleting all variables, which allow the direct identification of an observation (name, address etc.). All other variables are not modified by anonymisation techniques.

Formal, de facto and absolutely anonymisation can be ordered into a hierarchy. Absolutely anonymisation represents the strongest level of anonymisation. De facto anonymisation is a stronger level than formal anonymisation. The level of anonymisation correlates negative with the potential of analysis. To achieve anonymisation, the reduction of information of different variables is necessary. Obviously, a stronger level of anonymisation necessitates a stronger intervention.

The main advantage of remote data execution is the access to micro data that are anonymised on the lowest level. That is why the potential of analysis is higher than for any other way of data access. Many research projects can be realised only with formal anonymised data. Therefore, many users have only the possibility to use the remote data execution. Especially with respect to business data, a significant intervention by anonymisation techniques is necessary to achieve a de facto or absolutely anonymised data file.

By law, however, it is not allowed to grant direct access to formal anonymised micro data. In all other ways of access the users have direct contact to the micro data. A solution for the following problem has to be found: How can users analyse statistics without having direct contact with the formal anonymised micro data?
In remote data execution, users can send Stata-, SAS-, SPSS-, or R-codes to the Research Data Center. The code is run by the staff of the Research Data Center. After a statistical disclosure control of the output, the user obtains the results.
For remote data execution, the users have to develop a code without a direct contact to the micro data. That is why a description of the statistics, which are of the user's interest, and a so-called data structure file are made available for the user. The description contains a list with all variables and their characteristics so the user can see which variables are available for his analyses. The user can develop the code based on this variable list.

To check the code for possible syntactic mistakes, the user can test the code with a data structure file. This structure file has to be absolute anonymous. The Research Data Centre generates the structure files with the

following technique: In a first step, the Research Data Centre draws a small random sample from the original formal anonymised micro data file. In a second step, the order of the values of all variables is randomly rearranged. The rearrangement of a variable is independent of the rearrangement of all other variables. The data structure file is the result of this independent rearrangements. The user gets the data structure file to check the codes for accuracy for the remote data execution. First and foremost, this check can test the syntactic correctness. Results of analyses based on a data structure file give only little information concerning the final results, which can be obtained with the original data.

Based on the structure file, it is not possible to obtain information about the covariance structure of the original formal anonymised micro data file. Hence, based on structure files, a user has no hints for results of an analysis, which is based on the covariance structure. Especially, if a user wants to estimate a model, the current structure file does not provide enough information. The structure file is not adequate to offer information about the model fit. The only solution is sending the code to the Research Data Centre and testing the model by the original formal anonymised micro data. The Research Data Centre runs the user code on the original data and checks the results for statistical disclosure. After this, the user obtains the safe output. Based on these results, the user can modify the model and send the code again to the Research Data Centre. As the user has no information about the covariance structure, it is the only way to send codes often back and forth between the Research Data Centre and the user to obtain satisfactory results. This kind of work is time-consuming and labour-intensive for both, data producers and data user.

2. Improvement of data access: The InfinitE-Project

The intention of the project “An informational infrastructure for the E-Science Age - On the way to remote data access for business data” (infinitE) is to achieve an improvement of the access to economic micro data. The Research Data Center of the Federal Statistical Office participates in the project in cooperation with other data producers and representatives of the user of the micro data. For an improvement of the data access, the project focuses on two goals:

1. Development of syntactic and semantic data structure files
2. Improvements in automatic output control

The first goal is to develop data structure files that preserve the correlations between the variables of th e original data file. The second goal is to shorten the time spend on the process of disclosure control and to make access to micro data easier and faster.

3. Development of syntactic and semantic data structure files

To reduce the expenditure of time, the Research Data Centre develops data structure files, which should allow the user a syntactic and semantic test of the own code. Based on this semantic and syntactic structure file, it would be possible for the user to obtain a first guess of the final results. This syntactic and semantic structure file would reduce the frequencies of sending codes from the users to the Research Data Centre. In the project "An informational infrastructure for the E-Science Age" the Research Data Centre develops in cooperation with its partners this syntactic and semantic data structure file.

These semantic and syntactic structure files should satisfy two particular objectives:

1. All analyses, which are made by original micro data, should also be available by data structure files.
2. The analytical result, obtained by a data structure file, should give a good idea how the results based on the formal analysed micro data will look like.

In cooperation with partners who are part of the scientific community, some criteria were set up for the semantic and syntactic structure files:

- The data structure file should contain the same variables as the original micro dataset.
- Structural dependencies and relationships should be preserved.
- The data structure file should achieve the range of metric variables.
- Categorical, ordinal and nominal values should be preserved.
- The dimension of the data structure file should be comparable to the original micro data.
- The frequencies of discrete variables should be approximately preserved.
- The logical data structure should be preserved (e.g. the total sale is the sum of the partial sales).
- Structural zeros should be remained (e.g. if the company does not operate on trade, it should also been shown in the data structure file).
- Descriptive statistics (e.g. mean, median, etc.) should be preserved approximately.
- Correlations, especially the sign of significant correlations, should be preserved.
- Regression coefficients based on the data structure files should retain the same signs as regression coefficients based on the original micro data

At present, the members of the project "An informational infrastructure for the E-Science Age" investigate, which anonymisation techniques are adequate to generate semantic and syntactic structure files. Different perturbation methods come into consideration for the development of such structure files. Methods, which might be applied to produce such data structure files, are in particular the data perturbation methods of multiplicative stochastic noise and multiple imputation. First investigations of the implementation of these methods are initiated. At this stage, it is not possible to decide which of the methods is the most promising.

It is intended to develop semantic and syntactic structure files, which are absolutely anonymous. With respect to absolutely anonymous semantic and syntactic structure files, it is possible to offer these files as a free download on the web site of the Research Data Centre.

Now, it is not foreseeable, whether semantic and syntactic structure files, which fulfil the aforesaid criteria, can be generated as absolutely anonymous files. If these files have to be classified as de facto anonymised files, it is necessary to conclude a contract with the user before he or she can obtain a file.

4. Improvements in automatic output control

The results of analyses, which are realized by remote data execution, have to be checked by the Research Data Centre for statistical confidentiality. It has to be impossible to obtain information about a single observation. That is why a user cannot get any results of an analysis which allows drawing conclusions to single units.

In general, all results of an analysis have to be checked from the Research Data Centre. Especially magnitude and frequency tables, which are based on a small number of cases, can violate the statistical confidentiality. However, multivariate analysis can violate the statistical confidentiality as well.

To guarantee the statistical confidentiality, the staff of the Research Data Centre checks all outputs manually. For an anonymisation of the outputs, a cell suppression technique is adopted. All cells which are identified as ‘unsafe cells’, become suppressed (“primary suppression”). Additionally, cells must be suppressed (“secondary suppression”), if they would allow a re-identification of the primary suppression. This kind of output checking is highly difficult for complex tables and large estimation outputs and is very time consuming and labour-intensive.

Some tools for an automatic output checking already exist. But these tools are not adequate for an application in context of a remote data access. For example, the software tau-argus executes an automatic output checking. For using tau-argus as a tool for output checking, it is necessary to prepare the output files. Unfortunately, tau-argus cannot process the output files of SPSS, SAS, Stata or R.

It will be a task for the future to develop a software tool for an automatic output control, which can be used for the flexible results of remote data execution. Scientific users are used to work with the standard software programs for statistical analysis, like SPSS, SAS, Stata or R.

Therefore, it is necessary to find a way that allows users to work with their standard statistic software and to check the output automatically with adequate software tools.

The project infinitE analysed, whether the concept of “de-facto anonymisation” can help to generate intermediate results for the scientific users in a shorter time. Until now, the concept of de-facto anonymisation is only applied on micro data. So, scientific user can obtain de-facto anonymous micro data files for research. In comparison to other user groups scientific users obtain micro data files with less information reduction by anonymisation techniques. A similar strategy can be applied on results of scientific user. Scientific User can obtain intermediate results, which are de facto anonym. As the production of de facto anonym micro data, the production of de facto anonym results need less intervention by anonymisation techniques because more information can be conserved in the results. Therefore, it might be easier to find an automatic solution for the production of de-facto anonymous results.

The de-facto anonymous results can only be used by the user as intermediate results. All results, which a user wants to publish, have to be checked for absolutely anonymisation by the Research Data Center before a publication takes place.

5. On the way to a remote data access

The real remote data access via a remote server is a vision for the Research Data Centre. For realising this kind of access, which allows a data access to formal anonymised data from an external computer and an automatic output check in real time, some methodical problems have to be solved.

For the establishment of a remote server, it is necessary to develop tools, which allow an automatic start of the users' transmitted code without any intervention from the staff of the Research Data Center.

The development of semantic and syntactic structure files can be used as a temporarily solution to establish a faster way of generating results via remote data execution. In the final solution of a remote data access system, the semantic and syntactic structure files are an important component, too. By law, it is not possible in Germany to visualise the formal anonymised micro data on the users' monitor. As a substitute for the formal anonymised micro data, the semantic and syntactic structure files can be displayed. However, for the calculation of the results the original data are used.

Before the user obtains the results, it is essential to check the results for statistical disclosure. For this purpose it is necessary to develop tools, which allow a (semi)automatic output control. At present, there are no suitable automatic methods available.

REFERENCES (RÉFERENCES)

Brandt, M. / Zwick, M. 2009: Improvement of the informal infrastructure – on the way to Remote Data Access in Germany, http://www.unece.org/stats/documents/ece/ces/ge.46/2009/wp.16.e.pdf.

Gomatam, S. et.al. (2005) *Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers*. In: Statistical Science 20, pp. 163-177.

Hundepool, A. et al. (2010) *Handbook on Statistical Disclosure Control*.

http://neon.vb.cbs.nl/casc/.%5CSDC_Handbook.pdf

Zühlke, S., Zwick, M., Scharnhorst, S. and Wende, T. (2004) The research data centres of the Federal Statistical Office and the Statistical offices of the Länder. In: Journal of Applied Social Science Studies 124 (4), pp. 567-578.