WP. 16
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (iii): Research data centres and virtual labs

IMPROVEMENT OF THE INFORMATIONAL INFRASTRUCTURE – ON THE WAY TO REMOTE DATA ACCESS IN GERMANY

Invited Paper

Prepared by Maurice Brandt and Markus Zwick, Federal Statistical Office, Germany

Improvement of informational infrastructure
On the way to remote data access in Germany

Maurice Brandt* and Markus Zwick**

* Federal Statistical Office Germany, e-mail: maurice.brandt@destatis.de
** Federal Statistical Office Germany, e-mail: markus.zwick@destatis.de

Abstract: For a couple of years, access to microdata in Germany is possible through different ways of access. The researcher can use so called scientific use files, even for business microdata, in his or her own institution. Furthermore they can visit the safe centre or can use remote data execution.
There are still reservations in the scientific community concerning the data perturbation methods for business microdata. Unlike household surveys, business surveys are faced with skewness of the distribution. Due to the anonymisation methods the skewness of the original microdata can not be preserved. For this reason the user needs are developing towards original microdata, which can only be analysed via remote data execution or within safe centres inside the research data centres (RDC).
This leads to a higher burden for the employees of the RDC on the one hand. For remote data execution they have to apply the analysis programs and have to deal with the manual output checking of the results. On the other hand the researchers themselves have to wait longer to get their results.

The project "An informational infrastructure for the E-Science Age" deals with the improvement of remote access. The project aims to find solutions for better remote access in Germany through so called data structure files and (automatic) output checking procedures.

Keywords: [Remote Access, Output Checking, Business Microdata]

1 Introduction

Since the research data centres (RDCs) of the statistical offices of the German Federation and federal states were set up in 2001, they have become firmly established and today empirical science is unimaginable without them.

The demand for data of official statistics has reached a dimension which is very difficult to deal with by the research data centres within a reasonable period of time. What is particularly time-consuming is on-site use of the data, i.e. the data are processed at the safe centre of the research data centre or through remote data execution; at the end of either of those processes the tables containing the results are checked for confidentiality.

Another option is to use the scientific use files (SUFs), which can be analysed on the researcher's own workstation in the relevant institution. There is much demand for SUFs with regard to person-related surveys such as the microcensus. For business data, more intensive anonymisation measures are performed because data perturbation methods such as microaggregation or stochastic noise are inevitable to protect the enterprises. Due to reservations regarding data perturbation methods of anonymisation and due to longer waiting times for providing a SUF, demand for data from economic statistics is increasingly shifting towards original microdata through on-site use, and in particular towards remote data execution. That form of access is higher comfortable for researchers as it does not require travelling to, and staying at the RDCs. Upon request, the researcher receives a so called data structure file consisting of a fully anonymised subsample whose structure is identical to that of the original data. They can be used to write the program for data analysis. For the RDC staff, that form of access is time-consuming for two reasons. First, the programs must be adjusted several times because the current data structure files do not reflect very well the original data as they are strongly anonymised (sampling and exchanging). Second, manual checking of results and checking for confidentiality takes much of the overall time needed for a remote data execution order.

Both points are treated by the project "An informational infrastructure for the E-Science Age - On the way to remote data access for business data" described in this paper.

2 Current situation at the research data centres

The research data centres (RDCs) of the statistical offices observe a fundamentally changing demand for their products. It turns out that remote data execution and safe centres have become the most frequently used forms of access to microdata of economic statistics in Germany. Demand thus focuses on on-site data use, which is highly time-consuming for both data producers and data users (Zwick 2006 and Zühlke et al. 2004).

With the new projects *Amtliche Firmendaten für Deutschland (AFiD)* (Official company data for Germany) and *Kombinierte Firmendaten für Deutschland (KombiFiD)* (Combined company data for Germany) the statistical offices, together with other partners, deal with highly complex data sets (cf. Bender et al. 2009). Ensuring de facto anonymity and, at the same time, maintaining a maximum of the analysis potential is a complex job even for cross-section and panel data of economic statistics. Considerable impairment of the analysis potential is inevitable. It is foreseeable that the new data sets, which are currently produced in AFiD and KombiFiD and will be linked in cross-section terms and over time, would – as a result of the anonymisation measures required to produce a SUF for off-site use – lose their information potential that will have been redesigned by such linking. As a

result, demand for time-consuming on-site use will further increase. However, the RDCs of the various data producers have to cope with heavy workload even now as their capacities are widely used especially in remote data execution.

3 Project activities

A real remote access application, which is fully automated and does not require any manual handling, is a vision for the future which has not been achieved in countries with a legal frame like Germany either. A reasonable interim goal for Germany in the medium term is a remote access solution like those implemented in the Netherlands, Denmark or at NORC[1]. The project "An informational infrastructure for the E-Science Age - On the way to remote data access for business data" is necessary to take first steps towards that goal. Furthermore a "RDC in RDC" solution where data of one RDC can be processed in another RDC, using remote access, would be a possible next step for Germany. This can be used as a test implementation for real remote access to be established later and permits shifting activities towards more data exploration, data documentation, internationalisation and less visitor care.

Before a remote access application can completely be implemented, however, many technical, legal and – the focus of the project – methodical problems must be solved.

Although first applications are available – Lissy in Luxembourg[2] and the methods of the Dutch[3] and Danish[4] national statistical institutes –, none of them provides fully automated access routines and part of them, e.g. Lissy, are limited to specific applications. In Sweden, MONA[5] is a feasible remote data access solution which, however, can be applied only because of the particular legal situation of data use in Sweden. In Germany, SAM[6] is a first technical solution. JoSua[7], too, might be extended to become such an application.

[1] NORC: National Opinion Research Center at the University of Chicago
[2] See e.g.: Coder, John and Cigrang, Marc (2003): LISSY Remote Access System
[3] Hundepool, Anco and de Wolf, Peter-Paul (2005): OnSite@Home: Remote Access at Statistics Netherlands
[4] Borchsenius, Lars (2005): New Developments in the Danish system for access to microdata.
[5] MONA: Microdata ON-line Access, Statistics Sweden
[6] Heitzig, Jobst (2006): Wissenschaftsserver zur Auswertung von Mikrodaten (Science servers for microdata analysis)
[7] The data centre of the Institute for the Study of Labor in Bonn (IdZA) has developed an application allowing external researchers to start microdata analyses via the internet. That application (JoSuA) is, first, user-friendly because researchers can monitor the status of their orders from their workstations and, second, it facilitates IdZA activities because it is no longer necessary to start the programs manually.

The research project seeks to establish the bases for the following three methodical steps:

(1) Developing anonymous data structure files which can be used to specify analysis models. Therefore they must be suitable for semantic analysis and allow developing analysis programs that are error-free in terms of syntax.

(2) Developing and assessing methods of standardised and (completely automated) checking of results.

(3) Simultaneous consideration of microdata anonymisation and checking of results.

The development of basic strategies for producing anonymised data structure files which allow checking a program run for syntactic and semantic errors is the project's purpose. Therefore the project group has established criteria on data structure files which are necessary for a useable remote access for both sides, the researcher and the NSI staff. Current data structure files – which are what users of remote data execution get – allow only syntactic checking. Methods that might be applied to produce such data structure files are in particular the data perturbation methods of multiplicative stochastic noise, multidimensional microaggregation and multiple imputation. First investigations of the implementation of these methods are initiated. The project team uses a concrete scientific objective regarding labour market development to ensure, that the data structure files will comply with scientific standards.

By now checking the results post-tabular is always time-consuming and labour-intensive. Results of remote data execution even intermediate results and of activities performed at safe centres are checked for confidentiality before being released. Such checks are extremely difficult for complex tables and large estimation outputs. Automated procedures have been developed only in some cases for standardised results becoming available regularly. For the flexible analyses performed in the research data centres, the methods developed so far – even at the international level – are far from sufficient. The project's function here is to extend the current state of knowledge and to widen the issue and to perform a systematic comparison between data-based and result-based safeguarding of the protection of the carriers of variables with regard to the analysis potential. It should be a goal to develop methods allowing the user to decide, before running the analysis, whether it should be performed with anonymised data and without restrictions for results or on the basis of the original data set and with restricted release of results. At this point, mixed forms could be envisaged. Before automatic output checking can be realised all sorts of results have to be categorised in safe and unsafe. Results received by commands like list, describe and frequency, for example, are classified as unsafe. As well are magnitude tables with n less than three characteristics per cell. On the other hand regressions,

variances and standard deviation can be classified as safe. However specific combinations of analyses will be discussed.

3.1 Producing data structure files

A first goal of the project is the standardisation of the data in the form of so called data structure files. Such anonymised data sets, which have the same structure as the original data sets, are sent to researchers who made a request for use, so that they can develop their program codes for analysis and send them to the relevant research data centre. That program code is then applied by the RDC staff to the original data and the output is returned to the researchers after checking for data security and confidentiality.

The project has set two objectives which the data structure files have to meet.

1. All analyses, which are made by original micro data, should also be available by data structure files.
2. The analytical result, made by data structure files, should be interpretable such that the same syntax specification would be chosen as if the output based on original micro data.

The criteria are classified in requirements with respect to the first and the second objective.

1. All analyses, which are made by original micro data, should also be available by data structure files.
   - 1.(a) The data structure file should contain the same variables as the original micro dataset.
   - 1.(b) Structural dependencies and relationships should be preserved.
   - 1.(c) The data structure file should achieve the range of metric variables.
   - 1.(d) Categorical, ordinal and nominal values should be preserved.
   - 1.(e) The dimension of the data structure file should be comparable to the original micro data.
   - 1.(f) The frequencies of discrete variables should be approximately preserved.

2. The analytical result, made by data structure files, should be interpretable such that the same syntax specification would be chosen as if the output based on original micro data.

2.(a) The range of metric variables should be preserved approximately.

2.(b) The logical data structure should be preserved (e.g. the total sale is the sum of the partial sales)

2.(c) 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).

2.(d) Descriptive statistics (e.g. mean, median, etc.) should be preserved approximately.

2.(e) Correlations, especially the sign of significant correlation, should be preserved.

2.(f) Regression coefficients based on data structure files should retain the same signs as in case of regressions based on the original micro data

a. same parameter significance or insignificance

b. no change of sign in case of significant coefficients.

So far, many of the data structure files consist of a sample of the original material, which has been subjected to additional anonymisation measures, or of values generated at random within the value range of the data set. Although the variables are maintained in both approaches, their attributes and the dependence structures (filter, variance-covariance matrix) with regard to other variables are completely destroyed. Although researchers can check whether their programs are executable, they do not get any information on whether the actual issue has adequately been implemented. Therefore, in many cases, the researchers' analysis programs cannot be taken in an identical form for the subsequent application to the original data. Often adjustments must be performed by the researchers and the RDC staff.

For more complex data such as the Linked Employer-Employee Data Sets (structure of earnings survey (1995, 2001, 2006), Linked Employer-Employee Data Set of the Institute for Employment Research – IAB), the data structure files in their current form are not very helpful because, in practical work, they cause enormous adjustment problems and require much co-ordination between external researchers and the research data centres. For example, such data structure files do not allow performing consistency checks of whether the analysis program developed by the researcher is correct and can be applied without errors to the original data. To allow suitable adjustment of the programs to the remote data execution procedure, empirically working scientists must be able to test not only univariate calculations but also, and to an increasing extent, programs referring to multivariate issues. Evaluating multivariate analyses with the existing data structure files is difficult

because the covariances are not maintained here and because researchers, too, must repeatedly adjust the models with regard to the original data. This causes problems especially where the full number of observations is needed (e.g. in analyses with several waves). In addition, not all logical restrictions are always correctly represented. This means that the developed program codes must be sent several times between the researchers and the staff of the relevant institution before the desired result is available. But it also means that all analyses performed must be checked by the staff to ensure that data protection is not violated when providing the results.

Another advantage of data structure files that are error-free in terms of semantics and syntax is that the researchers can determine more exactly the number and extent of their tables containing the results and that they can adjust the evaluations on their own workstation until the desired tables are produced. This reduces the efforts required to check tables for confidentiality which possibly are not included in the publication.

Data perturbation methods of anonymisation have already been developed or been adjusted to the requirements of economic statistics of the German statistical offices or of the Federal Employment Agency (cf. Ronning et al. 2005 and Bender et al. 2008). The question of whether the developed data sets are absolutely anonymised public use files or de facto anonymised is currently in discussion. A public use file has the advantage that it can be sent to the researcher already before, or during the process of making the request.

For de facto anonymised data structure files, the request for use must first of all be decided upon and a contract on data use must be signed before the data material can be sent to the researchers. Where a public use file sufficiently reflects the data structure, it should be preferred to a de facto anonymised data set because the analysis program will be applied to the original data anyway, irrespective of whether the data structure file has been fully or de facto anonymised. Whether the data structure can be sufficiently represented by absolutely anonymous data sets or only with de facto anonymous data sets is under investigation of the project group.

Before a true remote access solution can be implemented in the form of a functioning technical infrastructure, it is indispensable to use data structure files because this provides the researchers with some flexibility when working on the analysis programs and because those programs can be developed independent of time and location.

For a functioning remote access, too, the data structure files – once developed – continue to be necessary. As viewing the original data on the user's screen would

represent a transmission of original data (which is problematic in terms of data protection), it might be useful for remote access to present on the researcher's screen the view of the microdata from the data structure files. When calculating the analyses, however, the original data are used. The data themselves, i.e. both the data structure files and the original data, remain on the servers in the protected rooms of the statistical offices. In this respect, the data structure files are both an interim solution until real remote access has been developed and a final product which can be implemented as an important component of a hardware-based access solution.

3.2 Result-based confidentiality

Remote access is an optimal solution to provide the scientific community with access to confidential data. It allows external researchers to perform analyses on their own computer via a remote server and the results are displayed in real time. Data protection receives in any case particular attention. For that reason a solution of how a researcher can be controlled in case of data protection being violated has to be found or basically how such violation of data protection can definitely be prevented. The safest way is to check the results for potential data protection risks before they are transmitted in real time.

So far, the results of remote data execution and of activities performed in safe centres are manually checked for confidentiality before they are released. Such checking is highly difficult for complex tables and large estimation outputs and it is very time consuming and labour-intensive.

To ensure confidentiality of tables, parts of the cells are generally suppressed in case of business data. As the tabular data disseminated by the statistical offices are linearly linked with each other through subtotals and marginals, i.e. they are additive, additional cells must be suppressed ("secondary suppression"), to protect the primarily secret cells against disclosure through subtraction. Establishing suitable secondary suppression – which would minimise the loss of information caused by the suppression – is a complex linear problem of optimisation. An overview of the common standard methods is given, for example, in Giessing, 1999.

Frequently, parts of the cells of a specific table are identical to cells of another table. In such cases, the selection of secondary suppressions must be co-ordinated across tables to avoid a situation where users can disclose suppressed cells by performing comparisons between such overlapping tables.

A further disclosure control method is rounding cell values to integer multiples of a fixed base (see Willenborg and Waal for details). The objective is to add acceptable uncertainty about the real values of each cell.

A distinction is drawn between random rounding and controlled rounding. Random rounding means that each cell value is rounded independently from other cells. An advantage of this method is that the marginal totals are rounded separately from the inner cells of the table. Nevertheless random rounding may be result in inconsistencies between tables whereas the method is unbiased. Therefore the additively of the table is not preserved which is in turn a serious limitation for research.

Unlike random rounding, controlled rounding yields additive rounded tables. Due to the additive advantage of the inner cells and totals the resulting table is biased. Thus the researcher does not know exactly how the table was rounded.

Each of the disclosure control methods modify the original tables and therefore result in information loss. The project group examine and compare the results of both methods by measuring the information loss via different measurement procedures.

Checking the results of tables produced in safe centres or through remote data execution is necessary, among other reasons, to ensure that the results finally published by the users cannot disclose the secondary suppressions from the statistical offices' own publications. This requires co-ordinating secondary suppression between standard publications and users' tabular data. There are no suitable automated methods available for that purpose. As handling of user tables is subordinate to handling standard publications here, considerable impairment of the quality of results due to massive secondary suppression is expected.

For the above, and other reasons, data perturbation methods have increasingly been proposed in the literature over the last few years as an alternative or in addition to the suppression methods. Most of the methods proposed perturbate the data at the aggregate level. However, there are also methods perturbating the microdata. As maintaining the quality of results in specific tables is the main focus also for those methods to be applied to microdata, such methods are considered as confidentiality methods for tables rather than anonymisation methods for microdata.

Data protection regarding table output has been a topical issue for a long time already at the statistical offices of the German Federation and federal states, whereas no systematic study has been performed yet on the problem of data protection for estimation output and non-linear analyses. The studies by Gomatam, Karr, Reiter and Sanil (2005) may be used as a basis here.

Heitzig (2005) provided an approach to result-based confidentiality. It transfers the idea of the jackknife approach to estimating standard errors to the confidentiality of microdata. The method's basic idea is to replace a value of the underlying original data by a random value (from a distribution with sufficient variance). The analyses will then be performed successively with all modified data sets and the interval of the values of the results is published.

There are other approaches to result-based confidentiality. The US Census Bureau, for example, follows an approach in which automated checking of results is not obtained by presenting those results in less detail but by applying restrictions of use which are controlled by the system (Zayatz, 2007). This refers, first, to the data made available to users for analysis (e.g. showing microdata in less detail, such as by combining categories of specific variables) and, second, to a limitation of the studies that can be done with the system to a limited catalogue of analyses.

Other approaches, known from the literature, deal specifically with the problem of confidentiality of regression residuals (Reiter, 2004 and Sparks et al. (2008). While Reiter proposes, among other things, to provide users with synthetically generated residuals instead of the original residuals, the approach by Sparks et al. is based on the production of box plots for the residuals.

4 Summary and outlook

The project described here forms an important bridge between the developments in improving data access channels for the scientific community over the last few years and the concepts planned already today for the future by the research data centres. It is a major milestone on the way towards real remote access.

With the current development of demand for microdata in Germany, and especially the development of access for on-site use, manual remote data execution is getting more and more difficult anyway for the capacity reasons mentioned above. Due to the increasing demand for various statistics, timely provision of scientific use files from highly different surveys is nearly impossible, too.

In the long term, real remote access seems to be the only feasible solution both nationally and internationally; all the more so as a method, once developed, can rapidly be transferred to other surveys and could allow "just in time" delivery of data. The technical developments have reached a phase where online access is possible from anywhere or will be possible soon with the relevant range.

Real remote access allows researchers to process the data independent of time and location and has the advantage that the data remain in the protected rooms (and on the protected servers) of the statistical offices. Also, that kind of data access increases the networking among researchers and the scientific transparency because any researcher may access the data and replicate results any time.

What is more, care should be taken especially in the e-science age that the development of the informational infrastructure is not left behind by technological development. The possibilities offered by the technical infrastructure are far from exhausted and have further potential for development in the future. Nevertheless legal issues, too, must be settled at this point.

The purpose of the project is, first, to provide the methodical bases for fully automated remote access. Second, it will reduce the burden on the staff of the research data centres of the statistical offices of the Federation and the federal states already in the course of the project. This can be achieved by producing data structure files and the tools needed to produce data structure files regarding any statistics for controlled remote access as well as the guidelines and tools for categorising and automated confidentiality checks. The project will be able to take account of the methods developed in other countries and to benefit in international working groups from the experience acquired. It also benefits from the methodical projects performed in the last few years in the area of anonymisation of business microdata.

References

Bender, S., Himmelreicher, R., Zühlke, S., and Zwick, M. (2009) *Improvement of Access to Data from Official Statistics – the case of Germany*. In: RatSWD Working Paper 118, (05.11.2009), http://www.ratswd.de/download/RatSWD_WP_2009/RatSWD_WP_118.pdf

Bender, S., Rosemann, M., Zühlke, S., and Zwick, M. (ed.) (2008) *Longitudinal data on local units and enterprises – New data offers and their research potential*. In: AStA – Wirtschafts- und sozialstatistisches Archiv, Vol. 2, No. 3, Springer

Borchsenius, Lars (2005) *New developments in the Danish system for access to micro data*. In: Monographs of Official Statistics, Luxembourg

Coder, John and Cigrang, Marc (2003) *LISSY Remote Access System*, working paper No.7 of the Joint ECE/Eurostat work session on statistical data confidentiality, Luxemburg

Gießing, S. (1999) *Statistical confidentiality in tables*. In: Statistisches Bundesamt (ed.): Methods of ensuring statistical confidentiality, Wiesbaden, pp. 6-26

Gomatam, S., Karr, A.F., Reiter, J.P., Sanil, A. (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

Heitzig, J. (2005) *The "Jackknife" Method: Confidentiality Protection for complex statistical Analyses*. UNECE/Eurostat Work Session on Statistical Data Confidentiality,Geneva

Hundepool, Anco and de Wolf, Peter-Paul (2005) *OnSite@Home: Remote Access at Statistics Netherlands*. UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva

Reiter, J.P., (2004) *New Approaches to Data Dissemination: A Glimpse into the Future (?)*. In: Chance, Vol. 17, No. 3, pp. 12-16

Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M. and Vorgrimler, D. (2005) *Manual on anonymisation of microdata of economic statistics*. In: Statistik und Wissenschaft, Vol. 4/2005. Statistisches Bundesamt

Sparks, R., Carter, C., Donnelly, J., O'Keefe, C.M., Duncan, J., Keighley, T., McAullay, D. (2008) *Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving Analytics*. In: Computer Methods and Programs in Biomedicine, Vol. 91, No. 3, pp. 208-222

Willenbourg, L. and de Waal T. (1996) *Statistical Disclosure Control in Practice*. In: Lecture Notes in Statistics no 111, Springer Verlag. New York

Zayatz, L., (2007) New *Implementations of Noise for Tabular Magnitude Data, Synthetic Tabular Frequency and Microdata, and a Remote Microdata Analysis System*. Proceedings of the Joint UNECE/Eurostat work session on statistical data confidentiality, Manchester, United Kingdom, 17-19 December 2007

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

Zwick, M. (2006), *Research data centres – Benefits and costs of an informational infrastructure for science, politics and data producers*. In: Wirtschaft und Statistik 12, pp. 1233 – 1240