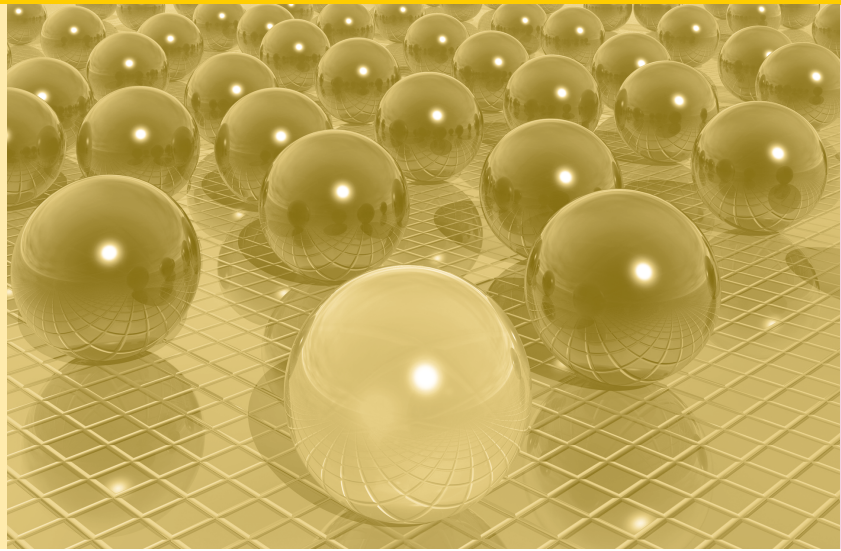


FDZ-Arbeitspapier Nr. 39



Improvement of data access -
The long way to remote data access in Germany

Maurice Brandt, Markus Zwick

2011

Impressum

Herausgeber: Statistische Ämter des Bundes und der Länder
Herstellung: Statistisches Bundesamt - Forschungsdatenzentrum -
Gustav-Stresemann-Ring 11, 65189 Wiesbaden
Telefon 0611 75-4220 • Telefax 0611 75-3915
Internet: <http://www.destatis.de>
E-Mail: forschungsdatenzentrum@destatis.de

Fachliche Informationen zu dieser Veröffentlichung:

Statistisches Bundesamt
Forschungsdatenzentrum

Tel.: 0611 75-4220
Fax.: 0611 75-3915
forschungsdatenzentrum@destatis.de

Informationen zum Datenangebot:

Statistisches Bundesamt
Forschungsdatenzentrum

Tel.: 0611 75-4220
Fax: 0611 72-3915
forschungsdatenzentrum@destatis.de

Forschungsdatenzentrum der
Statistischen Ämter der Länder
– Geschäftsstelle –
Tel.: 0211 9449-2873
Fax: 0211 9449-8087
forschungsdatenzentrum@it.nrw.de

Erscheinungsfolge: unregelmäßig
Erschienen im Juni 2011

Diese Publikation wird kostenlos als **PDF-Datei** zum Download unter www.forschungsdatenzentrum.de angeboten.

© Information und Technik Nordrhein-Westfalen, Düsseldorf, 2011
(im Auftrag der Herausbergemeinschaft)

Fotorechte Umschlag: © artSILENCEcom – Fotolia.com

Vervielfältigung und Verbreitung, nur auszugsweise, mit Quellenangabe gestattet. Alle übrigen Rechte bleiben vorbehalten.

Bei den enthaltenen statistischen Angaben handelt es sich um eigene Arbeitsergebnisse der genannten Autoren im Zusammenhang mit der Nutzung der bereitgestellten Daten der Forschungsdatenzentren. Es handelt sich hierbei ausdrücklich nicht um Ergebnisse der Statistischen Ämter des Bundes und der Länder.

FDZ-Arbeitspapier Nr. 39

Improvement of data access -
The long way to remote data access in Germany

Maurice Brandt, Markus Zwick

2011

Improvement of data access The long way to remote data access in Germany¹

Maurice Brandt, Markus Zwick²

Abstract

For a couple of years, access to microdata in Germany has been possible through different ways. The researcher can use so-called Scientific Use Files, even for business microdata, in his own institution, they can visit the safe centre or can use remote data execution.

User needs are developing towards on-site data access, because most often researchers would like to work with original data.

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" (infiniT) deals with the improvement of remote access. The project aims to find solutions for better remote access in Germany through so-called Structural Data Files and (automatic) output checking procedures.

Keywords: Remote Access, Output Checking, Business Data

Zusammenfassung

Bereits seit einigen Jahren ist der Zugang zu Mikrodaten in Deutschland auf unterschiedlichen Zugangswegen möglich. Wissenschaftler können sogenannte Scientific Use Files, u. a. auch für Wirtschaftsdaten, innerhalb der eigenen Einrichtung nutzen. Weitere Zugangswege bestehen über die Gastwissenschaftlerarbeitsplätze oder die kontrollierte Datenfernverarbeitung. Die Nutzeranfragen entwickeln sich zu einer gesteigerten Verwendung der On-Site-Datenzugänge, da die Mehrzahl der Wissenschaftler die Arbeit mit Originaldaten bevorzugt.

Hieraus resultiert eine besondere Herausforderung für die Mitarbeiter des Forschungsdatenzentrums (FDZ). Sie müssen bei der kontrollierten Datenfernverarbeitung die gängige Analysesoftware beherrschen und die Auswertungen manuell prüfen. Aus diesem Grund entstehen für die Nutzer lange Wartezeiten bis zum Erhalt ihrer Ergebnisse. Vor diesem Hintergrund befasst sich das Projekt "An informational infrastructure for the E-Science Age" (infiniT) mit der Verbesserung des Datenfernzugriffs. Ziel dieses Projektes ist die Entwicklung eines Datenfernzugriffs in Deutschland mit Hilfe sogenannter Datenstrukturfiles und einer automatischen Ergebniskontrollmaßnahme.

Schlüsselwörter: Datenfernzugriff, Ergebniskontrolle, Wirtschaftsdaten

¹ This paper is a reprint of the paper for conference "New Techniques and Technologies for Statistics" (NTTS) 2010.

² Federal Statistical Office Germany, Gustav-Stresemann-Ring 11, 65189 Wiesbaden - maurice.brandt@destatis.de, markus.zwick@destatis.de

1 Introduction

Since the Research Data Centres (RDCs) of the statistical offices in Germany 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. For the on-site use of the data, e.g. 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 outside the statistical offices. 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, the demand for data from economic statistics is increasingly shifting towards original microdata through on-site use, and in particular towards remote data execution. This kind of access is more comfortable for researchers because it does not require traveling to, and staying at the RDCs. Upon request, the researcher receives a so-called structural data file (SDF) consisting of a fully anonymised subsample whose structure is identical to that of the original data and which can be used to write the program for data analysis. The final analysis in the end will be done with the original microdata and not with the SDF. This way of access is labour-intensive for two reasons. First, the programs must be adjusted several times as the current Structural Data Files do not reflect the original data very well 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.

The project is funded by the Federal Ministry of Education and Research and carried out jointly with other partners, namely the Institute for Applied Economic Research (IAW), the statistical offices of the federal states Berlin and Hesse and the Institute for Employment Research (IAB).

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 innovative 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 specially panel data of economic statistics. De facto anonymity in this context means that the costs of reidentifying an observation need to exceed the benefit a potential data intruder can gain from this information. 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 – reduce their information potential that will have been redesigned by such a linking. As a result the demand for on-site use will further increase.

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 NORC³. 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 an “RDC in RDC” solution where data of an 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, internationalization 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 Luxembourg⁴ and the methods of the Dutch⁵ and Danish⁶ 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, MONA⁷ 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, SAM⁸ is a first technical solution. JoSua⁹, too, might be extended to become such an application.

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

- (1) Developing anonymous Structural Data 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 standardized and (completely automated) checking of results.
- (3) Simultaneous consideration of microdata anonymisation and checking of results.

The development of basic strategies for producing anonymised Structural Data 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 Structural Data Files which are necessary for a useable remote access for both sides, the researcher and the staff in the National Statistical Institutes (NSIs). Current Structural Data Files – which are what users of remote data execution receive – allow only syntactic checking. Methods that might be applied to produce such Structural Data 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 Structural Data Files will comply with scientific standards.

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 standardized results becoming available regularly. For the flexible analyses performed in the Research Data Centres, the methods developed so far – even at the international

³ NORC: National Opinion Research Center at the University of Chicago

⁴ See e.g.: Coder, J. and Cigrang, M.: LISSY Remote Access System (2003)

⁵ Hundepool, A. and de Wolf, P.: OnSite@Home: Remote Access at Statistics Netherlands (2005)

⁶ Borchsenius, L.: New Developments in the Danish system for access to microdata . (2005)

⁷ MONA: Microdata ON-line Access, Statistics Sweden

⁸ Heitzig, J.: Wissenschaftsserver zur Auswertung von Mikrodaten (Science servers for microdata analysis) (2006)

⁹ The data centre of the Institute for the Study of Labor in Bonn (IZA) 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 IZA activities because it is no longer necessary to start the programs manually.

level – are far from sufficient. The project's function here is to extend the current state of knowledge, to extend 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 less than three observations 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 Structural Data Files

A first goal of the project is the standardisation of the data in the form of so-called Structural Data 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 back 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 Structural Data Files have to meet.

1. All kind of analyses, which are made by original micro data, should be also modeled by Structural Data Files.
2. The analytical result, made by Structural Data 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 Structural Data Files.
 - 1.(a) The Structural Data File should contain the same variables as the original micro dataset.
 - 1.(b) Structural dependencies and relationships should be preserved.
 - 1.(c) The Structural Data File should achieve the range of metric variables.
 - 1.(d) Categorical, ordinal and nominal values should be preserved.
 - 1.(e) The dimension of the Structural Data 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 Structural Data 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 preserved (e.g. if the company does not operate on trade, it should also been shown in the Structural Data 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 Structural Data 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 Structural Data 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.

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 Structural Data Files in their current form are not very helpful because, in practical work, they cause adjustment problems and require much co-ordination between external researchers and the Research Data Centres. These data are just a part of the data in the projects AFID and COMBIFID, so that the data in this projects are even more complex. For example, such Structural Data 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 Structural Data 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 panel 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 staffs 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 Structural Data 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. Zwick et al. 2009 and Bender et al. 2008). The question of whether the developed data sets are absolutely anonymised Public Use Files (PUFs) 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 Structural Data 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 Structural Data 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 Structural Data 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 Structural Data 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 Structural Data Files. When calculating the analyses, however, the original data are used. The data themselves, e.g. both the Structural Data Files and the original data, remain on the servers in the protected rooms of the statistical offices. In this respect, the Structural Data 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

A remote system 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.

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, e.g. 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 Eurostat (2007).

Frequently, parts of the cells of a specific table are identical to parts of 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 de Waal (1996) 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 could end up in inconsistencies between tables whereas the method is unbiased. Therefore, the additivity 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 examines and compares 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

perturbing 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 in Germany, 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) could 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.

It needs to be taken into account that a hidden data attack with a combination of tables and multivariate analyses is also possible. For example with covert dummy variables it's feasible to unveil information during multiple multivariate analyses. Ronning 2010 ("Remote Access. A World Without Microdata??", previously undisclosed paper) has shown the existence of various scenarios.

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 "An informational infrastructure for the E-Science Age" 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 in Germany.

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. 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 at any 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 aim of the project is, to provide the methodical bases for fully automated remote access. This can be achieved by producing Structural Data Files and the tools needed to produce Structural Data 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.: Improvement of Access to Data from Official Statistics – the case of Germany, UNECE/Eurostat Work Session on Statistical Data Confidentiality, Bilbao (2009).
- Bender, S., Rosemann, M., Zühlke, S. and Zwick, M. (ed.): Betriebs- und Unternehmensdaten im Längsschnitt – Neue Datenangebote und ihre Forschungspotenziale (Longitudinal data on local units and enterprises – New data offers and their research potential). In: AStA – Wirtschafts- und sozialstatistisches Archiv, Volume 2, Number 3, Springer (2008).
- Borchsenius, L.: New developments in the Danish system for access to micro data, Monographs of Official Statistics. Luxembourg (2005).
- Coder, J. and Cigrang, M.: LISSY Remote Access System, working paper No.7 of the Joint ECE/Eurostat work session on statistical data confidentiality. Luxemburg (2003).
- Eurostat: Handbook on Statistical Disclosure Control. CENEX SDC, Version 1.01. Luxembourg (2007).
- Gomatam, S., Karr, A.F., Reiter, J.P., Sanil, A.: Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers. *Statistical Science* 20, pp. 163-177, (2005).
- Heitzig, J.: The “Jackknife” Method: Confidentiality Protection for complex statistical Analyses, UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva (2005).
- Hundepool, A. and de Wolf, P.: OnSite@Home: Remote Access at Statistics Netherlands, UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva (2005).
- Reiter, J.P.: New Approaches to Data Dissemination: A Glimpse into the Future (?). *Change* 17, pp.12-16, (2004).
- Ronning, G.: Remote Access. A World Without Microdata??, presentation on dagstat conference - statistics under one umbrella, Dortmund (2010).
- Sparks, R., Carter, C., Donnelly, J., O’Keefe, C.M., Duncan, J., Keighley, T., McAullay, D.: Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving Analytics. *Comput. Methods Programs Biomed* (2008).
- Willenbourg, L. and de Waal T.: Statistical Disclosure Control in Practice. *Lecture Notes in Statistics* no. 111 Springer Verlag, New York (1996).
- Zayatz, L.: 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 (2007).
- Zühlke, S., Zwick, M., Scharnhorst, S. and Wende, T.: The research data centres of the Federal Statistical Office and the statistical offices of the Länder, *Journal of Applied Social Science Studies* 124 (4), pp. 567-578, (2004).
- Zwick, M., Lenz, R.: Business Micro Data in Germany: Access and Anonymization, *Journal of Applied Social Science Studies (Schmollers Jahrbuch)* 129 (4), pp. 645-653, (2009).

Bisher sind in der Reihe folgende FDZ-Arbeitspapiere erschienen:

Arbeitspapier Nr. 38: Decentralised Access to Confidential Microdata in Europe, M. Brandt/ M. Zwick, Juni 2011

Arbeitspapier Nr. 37: Masking Micro Data with Stochastic Noise, J. Höhne/ J. Höniger, Mai 2011

Arbeitspapier Nr. 36: Enthüllungsrisiko beim Remote Access: Die Schwerpunkteigenschaft der Regressionsgerade, A. Vogel, April 2011

Arbeitspapier Nr. 35: Temporary agency work and firm performance, S. Nielen/ A. Schiersch, April 2011

Arbeitspapier Nr. 34: Harmonisation of statistical confidentiality in the Federal Republic of Germany, M. Brandt/ A. Crößmann/ C. Gürke, März 2011

Arbeitspapier Nr. 33: Remote Access. Eine Welt ohne Mikrodaten ??., G. Ronning/ P. Bleninger/ J. Drechsler/C. Gürke, Januar 2011

Arbeitspapier Nr. 32: Compiling a Harmonized Database from Germany's 1978 to 2003 Sample Surveys of Income and Expenditure., T. Bönke/ C. Schröder/ C. Werdt, Mai 2010

Arbeitspapier Nr. 31: The Research Potential of New Types of Enterprise Data based on Surveys from Official Statistics in Germany., J. Wagner, Oktober 2009

Arbeitspapier Nr. 30: Geschlechterspezifische Einkommensunterschiede bei Selbstständigen im Vergleich zu abhängig Beschäftigten - Ein empirischer Vergleich auf der Grundlage steuerstatistischer Mikrodaten, P. Eilsberger/ M. Zwick, Januar 2008

Arbeitspapier Nr. 29: Reichtum in Niedersachsen und anderen Bundesländern - Ergebnisse der Steuergeschäftsstatistik 2003 für Selbstständige (Freie Berufe und Unternehmer) und abhängig Beschäftigte, P. Böhm/J. Merz, November 2008

Arbeitspapier Nr. 28: Exports and Productivity in the German Business Services Sector. First Evidence from the Turnover Tax Statistics Panel, A. Vogel, Juli 2009

Arbeitspapier Nr. 27: Künstler in den Daten der amtlichen Statistik, C. Haak, August 2008

Arbeitspapier Nr. 26: Union Density and Varieties of Coverage: The Anatomy of Union Wage Effects in Germany, B. Fitzenberger/ K. Kohn/ A. C. Lembcke, August 2008

Arbeitspapier Nr. 25: German engineering firms during the 1990's. How efficient are export champions?, A. Schiersch, Juli 2008

Arbeitspapier Nr. 24: Zum Einkommensreichtum Älterer in Deutschland – Neue Reichtumskennzahlen und Ergebnisse aus der Lohn- und Einkommensteuerstatistik (FAST 2001), P. Böhm/ J. Merz, Februar 2008

Arbeitspapier Nr. 23: Neue Datenangebote in den Forschungsdatenzentren. Betriebs- und Unternehmensdaten im Längsschnitt, M. Brandt/ D. Oberschachtsiek/ R. Pohl, November 2007

Arbeitspapier Nr. 22: Stichprobendaten von Versicherten der gesetzlichen Krankenversicherung- Grundlage und Struktur des Datenmaterials, P. Lugert, Dezember 2007

Arbeitspapier Nr. 21: KombiFiD - Kombinierte Firmendaten für Deutschland, S. Bender/ J. Wagner/ M. Zwick, November 2007

Arbeitspapier Nr. 20: Neue Möglichkeiten zur Nutzung vertraulicher amtlicher Personen- und Firmendaten, U. Kaiser/ J. Wagner, Juni 2007

Arbeitspapier Nr. 18: Die Gehalts- und Lohnstrukturerhebung: Methodik, Datenzugang und Forschungspotential, H.-P. Hafner/ R. Lenz, Mai 2007

Arbeitspapier Nr. 17: Anonymisation of Linked Employer Employee Datasets. Theoretical Thoughts and an Application to the German Structure of Earnings Survey, H.-P. Hafner/ R. Lenz, Dezember 2006

Arbeitspapier Nr. 16: Die europäische Union - Integration von unten oder Eliteprojekt? Eine Sekundäranalyse von Mikrodaten der amtlichen Statistik, R. Nauenburg, November 2006

Arbeitspapier Nr. 15: Keeping in Touch - A Benefit of Public Holidays Using German Time Use diary Data, J. Merz/ L. Osberg, November 2006

Arbeitspapier Nr. 14: Zur Konzeption eines Taxpayer-Panels für Deutschland, D. Vorgrimler/C. Gräßl/ S. Kriete-Dodds, November 2006

Arbeitspapier Nr. 13: Anonymisierte Daten der amtlichen Steuerstatistik, D. Vorgrimler, September 2006

Arbeitspapier Nr. 12: Mikrosimulation in der Betriebswirtschaftlichen Steuerlehre, R. Maiterth, August 2006

Arbeitspapier Nr. 11: Der Anteil der freien Berufe und der Gewerbetreibenden an der Gemeindefinanzierung, M. Zwick, September 2006

Arbeitspapier Nr. 10: Konstruktion und Bewertung eines ökonomischen Einkommens aus der faktisch anonymisierten Lohn- und Einkommensteuerstatistik, T. Bönke/F. Neher/C. Schröder, August 2006

Arbeitspapier Nr. 9: Anonymising business micro data - results of a German project, R. Lenz/ M. Rosemann/D. Vorgrimler/R. Sturm, Juni 2006

Arbeitspapier Nr. 8: Scientific analyses using the Continuing Vocational Training Survey 2000, R. Lenz/ H.-P. Hafner/ D. Schmidt, Juni 2006

Arbeitspapier Nr. 7: A standard for the release of microdata, R. Lenz/ D. Vorgrimler/ M. Scheffler, Juni 2006

Arbeitspapier Nr. 6: Measuring the disclosure protection of micro aggregated business microdata, R. Lenz, Juni 2006

Arbeitspapier Nr. 5: De facto anonymised microdata file on income tax statistics 1998, J. Merz/ D. Vorgrimler/M. Zwick, Oktober 2005

Arbeitspapier Nr. 4: Matching German turnover tax statistics, R. Lenz/ D. Vorgrimler, Juni 2005

Arbeitspapier Nr. 3: The research data centres of the Federal Statistical Office and the statistical offices of the Länder, S. Zühlke/ M. Zwick/ S. Scharnhorst/ T. Wende, März 2005

Arbeitspapier Nr. 2: Eine kommunale Einkommen- und Körperschaftsteuer als Alternative zur deutschen Gewerbesteuer: Eine empirische Analyse für ausgewählte Gemeinden, R. Maiterth/ M. Zwick, April 2005

Arbeitspapier Nr. 1: Ein Vergleich der Ergebnisse von Mikrosimulationen mit denen von Gruppensimulationen auf Basis der Einkommensteuerstatistik, H. Müller, März 2005

Stand: 20.06.2011

