Défis méthodiques de la réalisation de l'accés aux données économiques allemandes par la téléinformatique automatisée

Rainer Lenz

DÉFIS MÉTHODIQUES LORS DE LA RÉALISATION DE L'ACCÈS AUX DONNÉES ÉCONOMIQUES ALLEMANDES PAR LA TÉLÉINFORMATIQUE AUTOMATISÉE

Rainer Lenz

Université des Sciences Appliquées de Mayence
Departement I, Holzstraße 36, 55116 Mayence, Allemagne

Abrégé:
Pour établir la téléinformatique automatisée, il faut partir des préliminaires méthodiques et techniques. L'étude suivante montre le chemin que les Allemands prennent pour atteindre ce but tout en se concentrant sur les défis méthodiques qui se posent avec l'accès à des enquêtes confidentielles du patronat (observées pour une année ou pour une période de plusieurs années). Un pirate informatique tente, à l'aide d'informations complémentaires extérieures, de réidentifier des ensembles de données d'entreprises dans un fichier protégé. Les défis consistent d'une part à produire des fichiers rendus anonymes, transparents et comprehensibles pour les scientifiques qui exercent leur métier chez eux ou à l'institut, des fichiers dont les structures ressemblent à ceux qui sont authentiques et originaux. Ces fichiers dits fichiers de structure sont produits à l'aide des procédés spéciaux garantissant l'anonymat (p. ex. microagrégation, recouvrement stochastique ou imputation multiple). Et d'autre part, les défis consistent à développer des procédés concernant le contrôle standardisé et automatisé de résultats d'analyse.

Mots clés: Enseignement de la statistique, Données économiques et sociales

Abstract:
A number of methodological and technical preconditions have to be met in order to provide (automated) remote data access. This paper outlines the German approach to achieving this goal with a strong focus on the scientific and methodological challenges, in particular, regarding the access to cross-sectional and longitudinal economic statistics. The challenges comprise the generation of anonymised data with a structure similar to that of the real data, which are made available to data users. The data are provided in the form of what are called data structure files, which are produced by using specific variants of microaggregation, multiplicative stochastic noise and multiple imputation. Other challenges include the development of standardised output checking procedures for tables and estimates and the combination of macro and microdata confidentiality methods.

Key words: Statistical data confidentiality, remote data access, business microdata

1 Introduction

The producers of data relating to surveys of economic statistics in Germany have observed a fundamental change in the demand for their products. In early 2000, providing the scientific community with so-called scientific use files (SUFs) – which researchers can use at their own workplaces outside the statistical offices - was considered a way, if not the "royal road", towards giving empirical social and economic research adequate access to official microdata in Germany. Such SUFs have been available for what is called the off-site use of selected and strongly demanded statistics. As regards SUFs of economic statistics, however, these data stocks have not been very

well received. Reasons to be mentioned in this respect are the new data perturbating anonymisation methods which are not yet familiar to researchers, too long waiting periods and the excessive effort required to compile the SUFs. Due to the necessary and partly drastic interference in the information structure of the data in the course of anonymisation and, in particular, the reservations still existing with respect to the data perturbating anonymisation methods used, SUFs of microdata of economic statistics for off-site use have by far not achieved the same prominence as, for instance, SUFs used for off-site purposes in the area of individual or household-related statistics. As has become apparent in the research data centres of the statistical offices of the Federation and the Länder, controlled remote data execution and safe centres have become the most frequently used forms of accessing microdata of economic statistics. For this reason, the current long-term objective is, on the one hand, to improve access through the controlled remote data execution procedure (still manual) in a way to provide researchers with direct insight into so-called data structure files (anonymised form of microdata). The other goal is to facilitate both automated checks of the outcome produced, such as tables or individual values of estimating functions, and hence an early data transmission to researchers.

This paper introduces a project on "An informational infrastructure for the e-science age" which was launched recently and constitutes an essential milestone towards remote data access in Germany. The institutions involved in the project promoted by the Federal Ministry of Education and Research (BMBF) include the research data centre of the Federal Employment Agency in its Institute for Employment Research, the Mainz University of Applied Sciences, the Institute for Applied Economic Research, and the research data centres of the statistical offices of the Federation and the Länder.

Under ideal conditions, an empirical scientist would have authorised round-the-clock access to official statistical data from any computer terminal. The results would be delivered to the researchers in real time upon immediate and fully automated confidentiality checks. To ensure such a form of data access for independent scientific research, however, a number of methodological, technical and legal issues must be settled. The focus of this project is on tackling the methodological challenges of remote data access. During the three-year period of the project, solutions are to be developed for a direct (one-to-one) application of the analysis programs of scientists by compiling what are called data structure files. The latter are used to apply, without any further interference or adjustments by the staff in the research data centres, the analysis programs to the data of official statistics. Furthermore, solutions must be developed with respect to output confidentiality. This refers to confidentiality of both outputs in tables and outputs of estimation. While taking into account the legal situation, too, a technical solution can be implemented only after the methodological aspects of remote data access have been tackled. The project uses a sustainable approach in this respect, which means that it will be possible to implement the results developed, like data structure files and output confidentiality concepts, in the context of future technical solutions. In the following chapters, the paper will outline the overall objective of the project, the anonymisation methods to be used in generating so-called data structure files and the approach to be applied in measuring the protective effect of the methods used.

2 Scientific objectives of the project

Manifold technical, legal and methodological issues (the latter are a focus of this project) must be settled before a pure remote access application can be completely implemented. Although first applications are available - Lissy in Luxembourg, Coder et al. (2003), and the methods of the Dutch (Hundepool and De Wolf, 2005) and the Danish (Borchsenius, 2005) national statistical institutes -, none of them provide fully automated access routines and some of them – e.g. Lissy – are restricted

to specific applications. In Germany, SAM (Heitzig, 2006) is a first technical solution. JoSua[1], too, could possibly be advanced to become an application of this kind. In particular, 'safe communication' through lines would be an issue in this context. However, IT aspects are not an object of this research project. The latter is rather to provide the basis for the following three methodological approaches:

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

(2) developing and assessing standardised and completely automated output checking procedures.

(3) simultaneous consideration of microdata anonymisation and output checks.

In the following we concentrate on the methods of anonymisation used to develop data structure files and on the investigations assuring protection of those data and hence confidentiality regarding the underlying individuals.

3 Development of data structure files

A first goal of the project is the standardisation of data in the form of so-called data structure files. The anonymised data sets, which have the same structure as the original data sets, are sent to the researcher after he/she has submitted a request for use. As a next step, the researcher develops an analysis program code and sends it to the competent research data centre (RDC). Members of the RDC staff then apply the program code to the original data and, upon data security and confidentiality checks, send the output back to the researcher. So far, the data structure files often 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 structure (filter, variance-covariance matrix) regarding other variables are completely destroyed. Hence a researcher can check whether his/her program is executable, though he/she does not obtain any information on whether the actual issue has been adequately implemented. For this reason, the analysis programs of scientists can often not be used in an unchanged form for the subsequent application to the original data. Instead, additional adjustments have to be made by the scientists and the RDC staff.

Basic strategies are to be developed for the production of anonymised data structure files which will allow checking a program run for syntactic and semantic errors. To produce such data structure files, data perturbation methods like multiplicative stochastic noise, multidimensional microaggregation and multiple imputation are particularly suitable. Against a modified background of goals, the results of the projects on "De facto anonymisation of economic microdata" and "Economic statistics panel data and de facto anonymisation" can be built on in this context. Both projects set the methodological basis for producing de facto anonymised data sets for enterprises and local units. In particular, data perturbating anonymisation methods were developed or adjusted to the requirements of economic statistics of the German statistical offices and the Federal Employment Agency (cf. Ronning et al., 2005).

[1] 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. On the one hand, the JoSuA application is user-friendly because researchers can monitor the status of their orders from their workstations and, on the other, it makes IdZA activities easier because the programs must no longer be started manually.

3.1 Production of synthetic data sets

A way of providing data structure files of a significantly better quality is to produce synthetic data sets based on the idea of a multiple imputation of missing values. The decisive advantage of this method is the universality of its approach. Any restrictions and filter structures can be taken into account in the production of the relevant sets. In addition, the approach can be applied to continuous variables in the same way as to categorical variables. Since an attempt is made to maintain the variance-covariance matrix of the complete data set, the results a scientist will get in using the data structure files will usually differ from the original data-based results to a minor extent only. Due to its high flexibility and also applicability to very complex and linked panel data sets, this innovative approach has been increasingly used at the international level in the past few years. Commissioned by the U.S. Census Bureau, a group of researchers headed by John Abowd (Cornell University) has for more than five years worked on creating, based on synthetic data, a public use file of the Survey of Income and Program Participation. Comprehensive studies (Abowd et al., 2006) have confirmed both the high quality and security of the data. Besides researchers in the USA, scientists in Canada, New Zealand (Graham et al., 2008) and Australia are engaged in the production of synthetic data sets. The previous project on "Economic statistics panel data and de facto anonymisation" showed, among other things, that the above approach can lead to very good results with regard to German business data, too. The results of univariate and multivariate analyses of the synthetic data are almost identical with those of analyses of the original data (Drechsler et al., 2007). Recently, the European Union commissioned an expert assessment of the applicability of synthetic data to EU statistics. In addition to two researchers from Spain and Italy, applicants of this project have also been involved in the assessment activities. The proposal to publish synthetic data sets that were generated on the basis of multiple imputation for scientific purposes was first raised by Rubin (1993) and further elaborated in Raghunathan, Reiter and Rubin (2003). The experience gained in producing synthetic data sets from materials of official statistics is described, for instance, in Kennickell (1999). The latter paper outlines the application of the approach to a French linked employer-employee data set. In late 2007, the first synthetic data set was made accessible to the public in the USA (Survey of Income and Program Participation, cf. Abowd et al., 2006). Synthetic data for panel data sets, which are to be made accessible to the general public, are currently being developed in the USA (Longitudinal Business Database, Longitudinal Employer-Household Dynamics Survey, American Communities Survey). The basic principle is to produce several synthetic data sets, which will be singly analysed. The actual result of analysis will be obtained by applying simple combination rules. In principle, fully synthetic and partially synthetic data sets can be distinguished. As regards fully synthetic data sets, all units of the universe which are not part of the survey sample are treated as missing values. As for those units, information has to be derived from the universe (e.g. from the business register or the statistics of employees) and then included in the imputation model. The missing values are imputed from the posterior distribution given the observed values. Various samples of the imputed values are then released for the scientific community.

3.2 Multiplicative stochastic noise

One main challenge regarding additive noise with constant variance is that on one hand small values are strongly perturbed and on the other large values are weakly perturbed. For instance, in a business microdata set the large enterprises -- which are much easier to re-identify than the smaller ones -- remain still high at risk after noise addition. A possible way out is given by the multiplicative noise approach explained as follows. Let X be the matrix of the original data and W the matrix of continuous perturbation variables with expectation 1 and variance $\sigma_{w^2} > 0$. The corresponding anonymised data X^a is then obtained as $(X^a)_{ij} := w_{ij} X_{ij}$ for each pair (i, j). The following approach has been suggested by Höhne (2004). In a first step, for each record it is randomly decided whether its values are increased or decreased, each with 0.5-probability. This is done using the main factors $1-f$ and $1+f$. In order to avoid that all values of some record are perturbed with the same noise,

these main factors are themselves perturbed with some additive noise s (where $s < f/2$). Particularly if the original data follow a strongly skewed distribution, the deviations using this method may strongly depend on the configuration of the noise factors for some few, but large values. That is, despite consistency, means and sums might be unsatisfactorily reproduced. For this reason, (Höhne, 2008) suggests a slight modification of the method.

The experience gained in the context of the project on "Economic statistics panel data and de facto anonymisation" can be built on in using the relevant methods for the production of data structure files. As regards the application to integrated data material, however, some methodological advancements are still required as high-quality data structure files cannot be produced by a simple combination of sampling and the subsequent application of stochastic noise. Only if the existing procedures are extended/adjusted, the characteristics of the overall stock of original data can be maintained in a sub-stock of the data. If data structure files must be produced as fully anonymous public use files, the selected parameters will have to ensure a by far stronger distortion than has been the case with scientific use files. In particular, the methods including additive noise with mixed distributions (see Roque, G.M., 2000 and Yancey, W.E., 2002) are anonymisation methods that can be used to meet the increased protection requirements by appropriate parameter selection. However, detailed tests are required in this respect in order to determine the specific type of additive noise and the parameter constellations that are best suited to fulfil the relevant requirements.

3.3 Uni- and multivariate microaggregation

The rationale behind microaggregation (see Domingo-Ferrer, J. and Mateo-Sanz, J.M. 2002) is that confidentiality rules in use allow publication of microdata sets if records correspond to groups of k or more individuals, where no individual dominates (*i.e.* contributes too much to) the group and k is a threshold value. Strict application of such confidentiality rules leads to replacing individual values with values computed on small aggregates (microaggregates) prior to publication. This is the basic principle of microaggregation. To obtain microaggregates in a microdata set with n records, these are combined to form g groups of size at least k. For each variable, the average value over each group is computed and is used to replace each of the original averaged values. Groups are formed using a criterion of maximal similarity. Once the procedure has been completed, the resulting (modified) records can be published.

The advantage of microaggregation methods is that they produce anonymous values in the form of a linear combination of real values. Thus they automatically make sure that content-related linear dependencies are maintained in variables that are treated together. In the context of the projects on "Anonymisation of economic microdata" and "Economic statistics panel data and de facto anonymisation", the multidimensional microaggregation procedures convinced through their high protective effect (particularly in including categorical variables). With respect to the analysis quality achieved, however, they were greatly inferior to other methods (also the one-dimensional microaggregation procedures). Since data structure files have to fulfil other requirements than scientific use files, however, microaggregation procedures can serve as an approach to meeting the relevant requirements. Compared, in particular, to multiple imputation, the advantage of this approach would be that the data user would have to be provided with only one data structure file.

4 Ensuring data protection in relation to data structure files

In order to measure the degree of data protection, it is planned to develop software tools for simulation of real data attacks. For this reason, in a first step some commercial database has to be prepared for those matching simulations.

4.1 Generating a database for data intrusion simulation

Due to the wide range of data variables and the dynamic additional knowledge linked to it (in some areas, the additional knowledge is broader or differs from the knowledge in other areas), research is required in order to determine the potential additional knowledge of a potential data intruder. Based on this research, an "intrusion database" is to be set up. First, a big effort is required to match the addresses of confidential data and suitable commercial databases.

4.2 Developing and implementing standards for risk assessment

Plans have been made to further develop the existing procedures for the simulation of data intrusion regarding cross-sectional and longitudinal data so that they will meet the requirements of two complex data structures. However, there have been almost no research results or systematic studies to assess the data security in synthetic data. In this respect, this project breaks new ground. As regards fully synthetic data sets, the risk of reidentification seems to be rather small because all values are artificial and each synthetic data set may include other units. There is a danger with partially synthetic data sets that, after a possible disclosure of individual local units or enterprises, the still unmodified or slightly modified part of variables can be useful for a potential data intruder.

As distinguished from the record linkage solutions which, in the case of data perturbation methods, can be applied to cross-sectional data (see Lenz, 2006) and longitudinal data (see Lenz, 2008), there is no way of checking allocations in the case of synthetic data due to the lack of direct identifiers such as name, address or business identity number. Hence another strategy has to be pursued. The data intruder must restock the target variables that are missing in his/her "intrusion database" in a way similar to the one used in (multiple) imputation approaches. Under ideal conditions, each individual value would be used only once in order to obtain a heterogeneous data stock. In multiple use cases, a so-called penal term could be incremented in the context of the allocation procedure in order to minimise the number of multiple uses. Upon allocation, the data intruder has to make use of what are called uncertainty measures. They may, for instance, comprise an estimation of intervals for the original values – the smaller the intervals, the smaller will be the uncertainty of the data intruder. Due to the big storage space they require, the efficiency of the simulation programs must be the focus of attention in developing standards for risk assessment for anonymisation approaches in general, and multiple imputation, in particular. For test purposes, the programs to be developed are to be initially applied to two data sets, namely the so-called industrial enterprise panel and the official turnover tax statistics.

5 Prospects

The project outlined in this paper serves as an important link between the developments that have been made in the past few years to improve the ways of data access for the scientific community and the concepts the research data centres are currently preparing for the future. Hence the project constitutes an essential milestone towards remote data access. In the long run, this way of data access seems to be the only practicable solution at both the national and international level, all the more since a method, once developed, can be applied to other surveys without delay and can hence ensure a just-in-time provision of data. The technical developments have reached a level which allows online access from anywhere and will allow online access to an adequate range of data soon. Remote access allows scientists to process data in a flexible manner which is independent of time and location. Also, it has the advantage that the real data remain in the protected rooms (and on the protected servers) of official statistics. Furthermore, this form of data access increases both networking among researchers and scientific transparency because any scientist may access the data and replicate the results at any time.

References

[1] Coder, John and Marc Cigrang (2003): LISSY Remote Access System*, Proceedings of the Joint Eurostat UN/ECE Worksession on Statistical data confidentiality,* Luxembourg, April 2003.
[2] Hundepool, Anco and Peter-Paul de Wolf (2005): *OnSite@Home: Remote Access at Statistics Netherlands,* Monographs of Official Statistics, Luxembourg.
[3] Borchsenius, Lars (2005): New Developments in the Danish system for access to microdata, *Proceedings of the Joint Eurostat UN/ECE Worksession on Statistical data confidentiality,* Geneva, November 2005.
[4] Heitzig, Jobst (2006): Wissenschaftsserver zur Auswertung von Mikrodaten, Federal Statistical Office of Germany, unpublished paper.
[5] Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M. und Vorgrimler, D. (2005): *Handbuch zur Anonymisierung wirtschaftsstatischer Mikrodaten*, Statistik und Wissenschaft, Bd. 4/2005. Statistisches Bundesamt.
[7] Abowd, J.M., Stinson, M., Benedetto, G. (2006): *Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project,* (www.sipp.census.gov/sipp/SSAfinal.pdf).
[8] Graham, P., Young J. and Penny, R. (Eds) (2008): Methods for constructing synthetic data, *Official Statistics Research Series*, Vol 3 [online] Wellington, New Zealand, Statistics New Zealand [in press].
[9] Drechsler, J., Bender, S., Rässler, S. (2007): Comparing Fully and Partially Synthetic Data Sets. for Statistical Disclosure Control in the German IAB Establishment Panel, *Proceedings of the Joint Eurostat UN/ECE Worksession on Statistical data confidentiality,* Manchester, December 2007.
[10] Rubin, D. B. (1993): Discussion: Statistical Disclosure Limitation, *Journal of Official Statistics*, 462-468.
[11] Raghunathan, T., Reiter, J. And Rubin, D. (2003): Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* 19 (1), Sweden, 1-16.
[12] Kennickell, A. B. (1999): Multiple Imputation and Disclosure Control: The Case of the 1995 Survey of Consumer Finances. In: *Record Linkage Techniques*, Washington D. C., 248-267.
[14] Reiter, J. P. (2003): Inference for partial synthetic, public use microdata sets, *Survey Methodology*, 181-189.
[15] Höhne, J. (2003) Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten. *Forum der Bundesstatistik*, vol. 42, Federal Statistical Office Germany,Wiesbaden.
[16] Höhne, J. (2008) Anonymisierungsverfahren für Paneldaten, *Journal of the German Statistical Society* (Wirtschafts- und Sozialstatistisches Archiv), vol. 2 (3), 259-276.
[17] Roque, G. M. (2000): Masking Microdata Files with Mixtures of Multivariate Normal Distributions, Dissertation, University of California Riverside.
[18] Yancey, W. E. (2002): Working Papers for Mixture Model Additive Noise for Microdata Masking, *Research Report Series* (Statistics #2002-03), Statistical Research Division U.S. Bureau of the Census, Washington D.C. 20233.
[19] Domingo-Ferrer, J. and Mateo-Sanz, J.M. (2002). Practical Data-oriented Microaggregation for Statistical Disclosure Control. *IEEE Transactions on Knowledge and Data Engineering*, 39, 189–201.
[20] Lenz, R. (2006): Measuring the disclosure protection of micro aggregated business microdata - an analysis taking as an example the German Structure of Costs Survey, *Journal of Official Statistics* 22 (4), Sweden, 681-710.
[21] Lenz, R. (2008): Risk Assessment Methodology for Longitudinal Business Micro Data, *Journal of the German Statistical Society* (Wirtschafts- und Sozialstatistisches Archiv), vol. 2 (3), 241-258.