Schmollers Jahrbuch 129 (2009), 645 – 653
Duncker & Humblot, Berlin

Business Microdata in Germany: Linkage and Anonymisation

By Rainer Lenz and Markus Zwick

1. Introduction

Scientists increasingly express the desire to use official statistics microdata for their own empirical economic and social research. In Germany, the road prescribed by the legislator is that microdata should be converted into a so called "factually" anonymised form, before they are made available to scientists. Accordingly, data items are regarded as sufficiently anonymised, if the expenditure needed for a possible reallocation is unreasonably high.

Fortunately, the work done by statistical offices in cooperation with scientists during 2002 – 2005, which was sponsored by the Federal Ministry of Education and Research (BMBF), on the project "Factual Anonymisation of Business Microdata" has shown that factual anonymisation of cross-sectional business statistics microdata can, as a rule, be achieved by using special information-reducing and data perturbing methods, see Lenz et al. (2006b) and Ronning et al. (2005). Evidence of that kind has still to be provided in respect of data items that are linked longitudinally. That is why, during 2006 – 2008, the Research Data Centre of the Federal Statistical Office, in cooperation with the Research Data Centre of the Statistical Offices of Germany's Federal States, the Institute for Applied Economic Research (IAW), Tübingen, the Institute for Labour Market and Vocational Research (IAB), Nuremberg and the University of Applied Sciences Mainz, has been conducting a BMBF-sponsored project "Business Statistics Panel Data and Factual Anonymisation". Some of the anonymised data files described below have also evolved from cooperation projects between the Land Statistical Office of Hesse and the Federal Statistical Office of Germany.

The required degree of confidentiality mainly depends on the way of data access the user decides for. The various ways of data use in general are not mutually exclusive; rather, an appropriate combination of approaches permits an adaption of anonymisation measures to the specific requirements of the data users. For instance, a scientist may in a first stage adapt his program codes to so called Campus-Files (data for teaching purposes, which have previously been absolutely anonymised and possess the same structure as the ori-

ginal data) and in a second stage he may apply his adapted programs to the original data via remote access.

In this paper we give an overview on various surveys of German business microdata associated with the way of their potential access; namely "on-site", different variants of "off-site" (like Scientific-Use-File or the previously mentioned Campus-File) and "remote data access". For detailed description of the different approaches to use German official data see Zühlke et al. (2004).

2. Business Microdata

In this section we describe several surveys, which have been made available for scientists so far. These data are namely the German Structure of Costs Survey (SCS), the German Monthly reports on local units in the manufacturing industry, the German Retail Trade Statistics (RTS), the German Turnover Tax Statistics (TTS), and the German Structure of Earnings Survey (SES). The data were generated in the context of several cooperation projects between German official statistics and empirically oriented research institutes.

2.1 German Structure of Costs Survey

The German structure of costs survey (SCS), limited to the manufacturing industry, is a projectable sample and includes a maximum of 18000 enterprises with 20 or more employees. All enterprises with 500 or more employees or those in economic sectors with a low frequency are included. That is, a potential data intruder has knowledge about the participation of large enterprises in the survey. The survey covers 33 numerical variables (among which are Total turnover, Research and Development and the Number of employees) and two categorical variables, namely the Branch of economic activity (abbreviated: NACE), and the Type of administrative district.

These data are available as Campus-File and Scientific-Use-File concerning the SCS 1999. Since the beginning of the year 2007 the longitudinal linked data 1999 – 2002 can be used on-site, and via remote data access. In the beginning of the year 2009, these data will also be available as Scientific-Use-File for off-site use. Further information on the SCS and the corresponding anonymisation strategies can be found in Lenz et al. (2006b).

Currently in progress is the anonymisation of the prescribed SCS-data for off-site use. The same holds for the Monthly reports on local units in the manufacturing industry 1995 – 2004. These data are – as the previously mentioned SCS data – generated during the project 'Business Statistical Panel Data and Factual Anonymisation' and can by now be used via on-site and remote data access.

2.2 German Retail Trade Statistic

The German Retail Trade Statistics (RTS) is a projectable sample containing about 23500 enterprises. In each branch of economic activity, the dominant enterprises have been included into the survey. The RTS consists of 33 numerical and 3 categorical variables. The results of this annual survey yield important information to economic-political problems concerning the structure, profitability and productivity of enterprises of this sector.

These data are available for the year 1999 as Scientific-Use-File or can be used on-site and via remote data access. Further information on the RTS and the corresponding anonymisation strategies can be found in Lenz et al. (2005).

2.3 German Turnover Tax Statistics

Turnover tax statistics (TTS) are based on an evaluation of monthly and quarterly advance turnover tax returns to be provided by entrepreneurs whose turnover in the year 2000 exceeds € 16,617 and whose tax amounts to over € 511 per annum. Also excluded are enterprises with activities which are generally non-taxable or where no tax burden accrues (e.g. established medical doctors and dentists without laboratory, public authorities). Nearly all economic branches are presented in the survey. For instance, the evaluation of the year 2000 contains almost 3 million records. The Federal Statistical Office of Germany published the following selected survey characteristics in tables: Deliveries and other performances (= taxable and non-taxable turnover), Branch of economic activity, Legal form, Bases of turnover tax (deliveries and other performances, intra-community acquisitions, input tax by tax rates, etc.).

These data are available as Scientific-Use-File for the TTS 2000. Since the beginning of the year 2007 the longitudinal linked data 2000 – 2004 can be used on-site and via remote data access. Further information on the TTS and the corresponding anonymisation strategies can be found in Vorgrimler et al. (2005) and Lenz et al. (2005).

2.4 German Structure of Earnings Survey

The German Structure of Earnings Survey (SES) is carried out by the Statistical Offices of the Federation and the Länder. For instance, for 2001, a total of a good 22,000 local units supplied data on over 845,000 employees. The group of reporting units comprises local units of the industry and selected parts of the service sector. The survey covers all employees who are subject to social insurance contributions and receive a remuneration in the month of report (October of the year of survey), including apprentices, trainees and those

in partial retirement. For the local units or for the enterprise to which the local units belong information is available on the economic activity according to WZ93 (German classification derived from NACE Rev. 1), the influence of the public sector on the business management, the number of employees of the enterprise, the number of employees of the local unit. Among the data covered for the employees are socio-demographic variables like sex and month and year of birth, data on job and qualification, information on working hours and earnings and the performance group of the employee which is an indicator for the complexity of the job.

These data are available as Campus-File and Scientific-Use-File concerning the SES 2001. Moreover, the data can be used on-site and via remote data access regarding the SES 1995 and SES 2001. Further information on the SES and the corresponding anonymisation strategies can be found in Frank-Bosch (2003) and Hafner / Lenz (2006). Currently, the anonymisation of the SES 2006 is in progress.

2.5 Second European Continuing Vocational Training Survey 2000

The German data of the Second European Continuing Vocational Training Survey 2000 (CVTS2) for 1999 as the reference year collected data from about 3200 enterprises with more than 10 employees in the economic sectors C-K and O of the NACE rev.1 on their employees' participation in continuing vocational training measures in 1999. The data contain information on the various forms offered in terms of continuing vocational training, about participants, hours of instruction and the cost involved (in relation to tuition classes) as well as qualitative data about the conceptual approach to continuing vocational training and the importance that the respective enterprise attaches to such training. We were, in particular, successful in our efforts to make sure that the data, which had been anonymised, remained suitable for a scientific treatment of relevant questions concerning economic sectors and employee size classes. Further information on the data and on the Scientific-Use-File can be found

These data are available as Campus-File and Scientific-Use-File. Moreover, the data can be used on-site and via remote data access. Further information on the CVTS2 and the corresponding anonymisation strategies can be found in Lenz et al. (2006a). Recently, the anonymisation of the CVTS3, which has been carried out in 2006 and refers to the year 2005, has been made available as Scientific-Use-File.

3. Measuring the Anonymity of Business Microdata

In order to evaluate the degree of anonymity of previously anonymised microdata, it was necessary to develop a technique for simulating data-intrusion scenarios a potentially attacking data intruder might apply. One important constellation is the so-called database cross match scenario. In a database cross match scenario, an attacking data intruder tries to assign as many external database units as possible (extra knowledge) unambiguously to units of an anonymised target database in order to extent the external database by target database information.

In a first phase, the database cross match scenario was mathematically modelled as a multicriterial assignment problem, which was then converted, by way of suitable parameterisation, into an assignment problem with one objective function to be minimised. Then, the main concern was to choose the best-fitting coefficients of this objective function. Whereas in the past a distance measure, generated across all matching variables of the two data sources (key variables and overlaps), proved to be well suited for the examination of cross-sectional data, see Lenz et al. (2006b), it turned out that the examination of panel data requires the use of additional, more elaborated measures. As the information on variables, which is in case of panel data available to a potential data intruder, has been collected in several waves, it seems obvious that this more complex structure should be reflected in the coefficients of the linear program as well. With that goal in mind we have implemented and tested several promising approaches. A more detailed description of these approaches can be found in Lenz (2008).

3.1 Conventional Distance Based Approach

For every numerical key variable v_i and every pair of records (a,b) of the cartesian product of the two data sources, the standardised square deviation is calculated. Afterwards, these component deviations are summed up. It may be advisable in some cases to assign additional weights to the various deviations on variable level. However, a weakness of that measure becomes apparent in cases where the definition of some key variable slightly differs between the two data sources, for example, if a variable such as “number of employees” relates to the number of all employees in absolute terms in one data set, whereas that number is converted into full-time workers in the other data set.

3.2 Correlation Based Approach

Let $v_l^e, \ldots, v_k^e$ and $v_l^t, \ldots, v_k^t$ ordinal key variables of the external and target data, respectively. We define $\boldsymbol{v}^e$ and $\boldsymbol{v}^t$ as variables from which have been

drawn k realisations and calculate the empirical correlation $corr(\mathbf{v}^e; \mathbf{v}^t)$ using Spearman's coefficient. As less this coefficient deviates from 1 as more likely the record pair (a,b) belongs to the same individual. Note that this coefficient can be applied either in case of numerical (and hence also ordinal) variables or in case of categorical variables, whose range forms a well-ordered set.

3.3 Distribution Based Approach

In a panel data situation we can take it for granted that an attacking data intruder will have information over several years for every key variable, for example, total turnover of an enterprise from 1999 to 2002. In general, we can assume the existence of a bias between the two sources of data in these variables. In order to counteract this problem, we consider the annual changes of a key variable and treat them like a frequency distribution of a discrete variable. Hence, we can apply statistical methods in order to measure the "similarity" of the frequency distributions on either side, external and target data.

3.4 Collinearity Approach

A data intruder might have information on two key variables over a period of n years in both sources of data, for example, "total turnover" $(u_1, \ldots, u_n)$ and "number of employees" $(b_1, \ldots, b_n)$ of an enterprise. If we interpret the pairs of values (u_i, b_i) as realisations of two random variables, those units that belong together in the different data sources can be expected to reveal empirical correlation cofficients that are 'similar'. It should, however, be considered that what is measured by correlation is just the linear interrelation of two variables. In special cases the two estimated correlation coefficients can diverge from each other very clearly, even if the variables are linked by a direct functional relationship.

3.5 Combination of Approaches

Because of the mentioned weaknesses of the various measures described above they are combined in a suitable way. Here we distinguish between two types of combination, *hybrid* and *composite* matching, see Lenz (2008). Once the coeffcients $d(a_i, b_j)$ are calculated, one can solve the linear assignment problem using classical established methods such as the simplex method. For larger data blocks (typically generated when dealing with tax statistics) it is recommendable for reasons of efficiency that approximation heuristics should be used. Fortunately, the usage of appropriate heuristics yields results near the optimum solution of the assignment problem, see Lenz (2003).

4. Aims and Scope

The aim of our investigations is to improve data access for researchers interested in cross-sectional and longitudinal business microdata. Recently, the focus was extended to longitudinal (particularly panel) data because of their high analytical potential and their limited availability so far. The availability of longitudinal business microdata for research purposes is improved by providing these data via remote data access, on-site use at the research data centres and Scientific-Use-Files which can be used at the researchers' own workplaces. An essential objective is to verify the extent to which business statistical panel data can be anonymised without loosing too much information. Detailed descriptions can be found in Ronning et al. (2005) and Rosemann (2006).

Currently, three new projects are carried out in order to improve the researcher's access to cross-sectional and longitudinal business microdata. Since in 2006 the Federal Statistic Act has been modified, it is now possible by law to merge various German business statistics. The German Business Register constitutes the core of this idea. On the basis of a direct identifier, the so-called enterprise number, the first project AFiD (translated: Official German Enterprise data) has the goal to combine several official business statistics in order to obtain an integrated microdata file. The second project KombiFiD (translated: Combined German Enterprise data) has the goal to merge official business statistics with those collected by other German data producers. The Federal Statistical Office, the Federal Employment Agency, the university of Lueneburg, the university of applied sciences Mainz and the German Central Bank are cooperation partners of this second project. Since for this proposal there is no legal basis in Germany, the only way to realise this project was to ask the respondents for their submission. The resulting sample for the combined survey contains about 100.000 enterprises.

The third project infinitE (translated: an informational infrastructure for the E-Science age) has been started recently. Its objective is to further automate the controlled remote data processing. Today, controlled remote data processing is a rather time-consuming procedure because, first of all, the programme syntax has to be checked for possible disclosure strategies and the data output must subsequently be checked for cases where data have to be kept secret. Those work steps still have to be done manually. Although first automated procedures are available now for such checks, it is not possible yet even with those approaches to fully automate controlled remote data processing.

The previously described and further data of German official statistics can be requested using the application forms available at the web site http://www.forschungsdatenzentrum.de/en/ of the research data centres of the statistical offices of the Federation and the Länder. Further information on the methodology and variables of the data is also contained in the corresponding metadata provided on http://www.forschungsdatenzentren.de/bestand/.

This work has been supported by the Federal Ministry for Education and Research (BMBF).

References

Frank-Bosch, B. (2003): Verdienststrukturen in Deutschland: Methode und Ergebnisse der Gehalts- und Lohnstrukturerhebung 2001 (in German), Wirtschaft und Statistik 12, 1137 – 1151.

Hafner, H.-P. / *Lenz,* R. (2006): Anonymisation of Linked Employer Employee Datasets, Proceedings of the Privacy in Statistical Databases Conference PSD 2006, Rome, December 13 – 15.

Lenz, R. (2003): Disclosure of confidential information by means of multi-objective optimisation. Proceedings of the Comparative Analysis of Enterprise Data Conference (CAED), London (CD-ROM publication, see http: // www.statistics.gov.uk/ events/caed/abstracts / lenz.asp).

Lenz, R. (2006): Measuring the disclosure protection of micro aggregated business microdata – An analysis taking the example of German Structure of Costs Survey, Journal of Official Statistics 22 (4), 681 – 710.

Lenz, R. (2008): Risk Assessment Methodology for Longitudinal Business Micro Data, Journal of the German Statistical Society, Wirtschafts- und Sozialstatistisches Archiv 2, 241 – 257.

Lenz, R. / *Hafner,* H.-P. / *Schmidt,* D. (2006a): Scientific analyses using the Continuing Vocational Training Survey 2000 (CVTS 2), Proceedings of the European Conference on Quality and Methodology in Survey Statistics (Q2006), April 24 – 26, Cardiff.

Lenz, R. / *Rosemann,* M. / *Vorgrimler,* D. / *Sturm,* R. (2006b): Anonymising business micro data – results of a German project, Schmollers Jahrbuch – Journal of Applied Social Science Studies 126 (4), 635 – 651.

Lenz, R. / *Vorgrimler,* D. / *Scheffler,* M. (2005): A standard for the release of microdata. Monographs of Official Statistics – Research in Official Statistics, 2006 edition, 197 – 206. (Invited paper, UN / ECE work session on statistical data confidentiality, November 9 – 11, Geneva).

Ronning, G. / *Sturm,* R. / *Höhne,* J. / *Lenz,* R. / *Rosemann,* M. / *Scheffler,* M. / *Vorgrimler,* D. (2005): Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten (in German), Statistik und Wissenschaft 4, Wiesbaden.

Rosemann, M. (2006): Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten (in German), PhD-Thesis, University of Tübingen.

Smeeding, T. M. / *Jesuit,* D. K. / *Alkemade,* P. (2002): The LIS / LES Project Databank: Introduction and Overview, Schmollers Jahrbuch – Journal of Applied Social Science Studies 122 (3), 497 – 517.

Statistical Office of Sweden (2003): Access to Microdata in the Nordic Countries. Report.

Vorgrimler, D. / *Dittrich,* S. / *Lenz,* R. / *Rosemann,* M. (2005): Wissenschaftliche Analysen anhand der Umsatzsteuerstatistik (in German), Wirtschaftswissenschaftliches Studium 10, 327 – 332.

Zühlke, S. / *Zwick,* M. / *Scharnhorst,* S. / *Wende,* T. (2004): The research data centres of the Federal Statistical Office and the statistical offices of the Länder, Schmollers Jahrbuch – Journal of Applied Social Science Studies 124 (4), 567 – 578.

Zwick, M. (2001): Individual tax statistics data and their evaluation possibilities for the scientific communit, Schmollers Jahrbuch – Journal of Applied Social Science Studies 121 (4), 639 – 648.

Zwick, M. (2007): CAMPUS Files – Free Public Use Files for Teaching Purposes, Schmollers Jahrbuch – Journal of Applied Social Science Studies 127 (4), 655 – 668.