Re-identifying register data by survey data: an empirical study

Bender, Stefan; Brand, Ruth; Bacher, Johann

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Statistical Journal of the United Nations ECE 18 (2001) 373–381
IOS Press

Re-identifying register data by survey data: An empirical study[1]

Stefan Bender[a], Ruth Brand[a] and Johann Bacher[b]
[a]*Institute for Employment Research* (*IAB*), *Regensburger Str. 104, 90478 Nürnberg, Germany*
Tel.: +49 911 179 3082; Fax: +49 911 179 3297; E-mail: Stefan.Bender@iab.de
[b]*Department of Sociology at the University of Erlangen-Nuremberg, Findelgasse 7/9, 90402 Nürnberg, Germany*
Tel.: +49 911 5302 680; Fax: +49 911 5302 660; E-mail: bacher@wiso.uni-erlangen.de

Abstract. More and more empirical researchers from universities or research centres would like to use register data collected by statistical agencies or the social security system, because these data can be used for several empirical studies, e.g. the analysis of special groups or quantitative effects of economic policies. Most of the register data required have to be (factually) anonymised before they are disseminated to preserve confidentiality. Therefore re-identification risks for register data are examined by matching a sample of register data with survey data, collected especially for scientific purposes. Three methods were applied: the uniqueness approach, a simple distance estimation and a cluster analysis. The data sets used were two birth cohorts (1964 and 1971) of the German employment statistics (register data) and the German Life History Study. The analysis show that a re-identification of real persons may be possible by a standard-cluster analysis or a simple distance criterion if an intruder has access to additional information. The number of re-identifiable persons is remarkably high although the proportion of re-identifiable persons is less than expected on the basis of the uniqueness-approach.

1. Introduction and description of the data

More and more German empirical researchers from universities or research centres would like to use register data collected by statistical agencies or the social security system. These data can be used for several empirical studies, e.g. the analysis of special groups or quantitative effects of economic policies. A register typically covers nearly all persons or enterprises in a population. The rules for collecting these data are usually defined on a legal base. In Germany most of the big registers are based on laws in which the purposes of the register and the basic rules for data processing are defined. Due to this only some persons are allowed to work with the data (mostly persons who are working in the institutions where the data are collected) and the data have to be (factually) anonymised before they are disseminated to the scientific community.

The concept of factual anonymity means that the data can be allocated to the respondent or party concerned only by needing "an excessive amount of time, expenses and manpower" [9]. Therefore, it is needed to take into account the additional information and the knowledge of statistical or computational methods that a potential attacker might have for de-anonymising the data.

[1]The life history surve was co-financed by the European Social Found. The research was supported by the grant of the PROCOPE French-German coopreration by the DAAD, project no 00311VG.

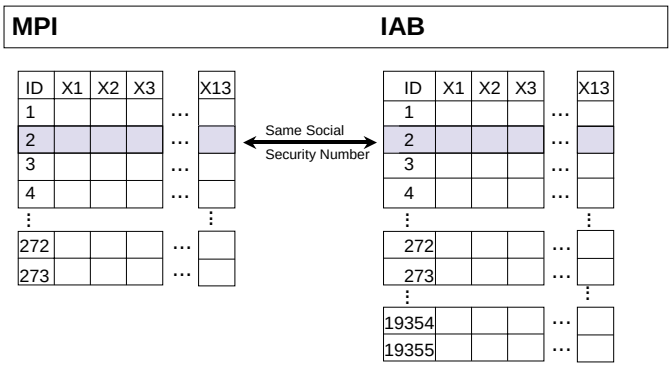

Fig. 1. Relationship between the two data sets.

In this paper re-identification risks for register data are examined by matching a sample of register data with survey data, collected especially for scientific purposes. The register data set is a sample of the German employment statistics. Its basis is the integrated notifying procedure of health insurance, statutory pension scheme and unemployment insurance. The procedure requires that employers report all information of their employees subject to the social security system to the social security agencies. Exact daily information on employment is included in the data and some characteristics (sex, age, employment duration and earnings covered by social insurance contributions) are very accurate, since they mainly serve insurance law purposes. Every person in the data set can be identified by the insurance number given [2].

The second data set is the German Life History Study conducted by the Max Planck Institute for Human Development (MPI-data set), which is a retrospective survey that evaluates quantified life histories – measured in months – for different dimensions (e.g. schooling, apprenticeship, employment, partnership, family, housing). The life course protocols are edited and corrected by using taped recordings. In cooperation with the institute for employment research (IAB), the MPI collected interviews of nearly 3 000 women and men born in 1964 or 1971 in Western Germany. About 80% of the interviewed persons gave their permission to match their responses with the data stored by the social security system, but the insurance number is available only for approximately 800 persons in the data ([7]; for a detailed description of the German Life History Study, see [6]).

For our analysis we take a sub-sample in the following form: we matched the data sets for those persons who were employed in Western Germany in November 1997 by their insurance number. By doing so we have 273 persons for whom we have information in both data sets (real twins). The main aim of this empirical study is the examination of the re-identification risk for samples from register data, especially for samples from the German employment statistics. Therefore a 2%-random sample of the two birth cohorts out of the register data is drawn ($n = 19\,082$) to which the 273 real twins are added (IAB-data set). At the end we have two data sets (Fig. 1):

– MPI-data set, which contains information of the 273 real twins out of the survey, and
– IAB-data set, which contains information of 19 082 persons (for these we have only information as available in the IAB-data) and 273 real twins ($n = 19\,355$).

The data sets used have in common the variables shown in Table 1. A descriptive comparison already shows that some of the data of the social security system differ substantially from those of the survey on the individual level (Table 3). While the basic demographic variables and some discrete variables, based on highly remarkable facts (like number of months worked), are not affected or not strongly affected by

Table 1
List of variables – comparison of the variables between the MPI and IAB-data sets

Variable	Remarks/Categories	MPI (retrospective survey)	IAB (social security data sets, register entries base on notifications by the employer)
Income	Daily earnings	Information for all occupations available	Payment for occupations notified by the social security system
Year of birth (birth-cohort)	Identical in both data sets, persons are born in 1964 or 1970	Checked before starting the interview	Part of the social security number
Sex	Identical	Checked before starting the interview	Part of the social security number
Schooling/Training	A-level (Abitur): yes/no and university/polytechnic degree	Broad range of categories	Highest secondary school qualification known by the employer
Hours of work per week at the main employment	Part time: yes/no	Number of hours	Part-time-work: less than 18 hours per week, 18 hours or more
Occupational Status	Unskilled blue collar, skilled blue collar, craftsmen, white collar	Broad range of categories	Blue-collar-annuity insurance/white collar annuity insurance
Occupation	327 categories	Open specification, manual coded by the researcher	Notified is the occupation at work not the educated profession
Region	Federal state	Open specification, manual coded by the researcher	Added by the local labour office via the firm identifier
Number of months in the working life	Maximum 24 months		
Interruption in the Working Life	Having a gap (over 3 months) in the working life		
Duration of the last employment (Change of establishment)	Changes during the last three years		
Martial status	Married / not married	Broad range of categories	Family status notified by the employer: married / not married
Number of children		Number of children living in the same household	Number of children notified by the employer
Nationality			Nationality or country of origin

the different data generating processes, others like some continuous variables, e.g. income, are not very similar in both data sets. This is explained by the different respondents (employer and employee) and the usual effects in retrospective studies. For example, 93.8% of the respondents in the MPI-survey have a zero as the last digit of the income variable. Most of these zeros appear because the respondents round their income to 0 in the last digit; in the IAB-sample there is no such rounding because of the notification procedure, so only 16.1% have a 0 in the last digit. 83.9% of the MPI-survey have 00 in the last two digits (IAB-sample: 4.4%) and 18.3% have 000 in the last three digits (IAB-sample: 1.1%). There are also remarkable concentrations in the MPI-Survey on the last three digits at 500 and 800 (Table 2).[2]

[2]We would like to thank Ulrich Rendtel and Jens Hanisch to give us the idea to do this comparisons.

Table 2
Comparison of the last digits in income between the MPI-survey and the IAB- data (in %)

Last Digits	MPI-survey	IAB-data
0	93.77	16.12
1	0.37	5.86
2	1.47	10.26
3	0.73	9.16
4	0.73	8.42
5	0.73	11.36
6	0.37	12.09
7	0.37	13.55
8	0.37	5.49
9	1.10	7.69
00	83.88	4.40
50	6.23	1.83
000	18.32	1.10
300	7.33	0.37
500	15.75	0.73
800	11.36	0.73

2. Protection and release of individual data sets in Germany

In Germany, article 16 (6) of the Federal Statistical Law establishes a scientific privilege for the use of official statistics. Microdata that are factually anonymous may be disseminated for scientific purposes. "Factual anonymity means that the data can be allocated to the respondent or party concerned only by employing 'an excessive amount of time, expenses and manpower." [9]. The data covered by the social security system are under the special protection of the Social Code "Sozialgesetzbuch (SGB)". Thus the social insurance bodies are allowed to pass on factual anonymised data to researchers.

The concept of factual anonymity takes into account a rationally thinking intruder, who calculates the costs and benefits of the re-identification of the data. Factual anonymity depends on several conditions and is not further defined by law. It is necessary to estimate the costs and benefits of a re-identification for every data set anew. The results of a cooperative project carried out by the University of Mannheim, the Federal Statistical Office and the Centre for Survey Research and Methodology (ZUMA) in Mannheim [3, 11,12] form the basis of the release of data in Germany. In this project, attempts were made to re-identify persons in the data set under different realistic scenarios. The experiments were based on comprehensive analyses of the costs and potential benefits of a re-identification. For this purpose it was also necessary to inquire into the motives scientists could have to re-identify the data and to determine the additional information available. Rules for data release were established based on the knowledge gained from those different attempts at re-identification. On the results of this project, data sets were anonymised and are now available for researchers working in Germany.

3. Measuring the re-identification-risk

In general the violation of the anonymity of data (disclosure of information) only means that additional information on an observed unit in a given data set is gained. The most important case of disclosure is re-identification. Re-identification of units (persons, households or firms) is only possible under the following conditions:

– The intruder has information on the unit he/she is interested in (additional information).

– The unit he/she is looking for is included in the data set.
– The data set and the source of additional information have some variables in common.
– It is possible to combine the variables in common so that an unequivocal match results.
– The intruder is sure (at least subjectively) that the link is correct.

Nevertheless, it must be kept in mind that disclosure without re-identification is also possible (attribute disclosure, see [8]). This kind of disclosure is not considered in this paper.

The additional information available and the number of identical variables is of crucial importance to the re-identification, as well as the validity of this information. In general the benefits of a re-identification of individual data are rather low. Therefore anyone interested in re-identifying persons or households would only accept low costs.[3] So we are making our estimations on simple algorithms and easily available software (e.g. SPSS). We used several methods to evaluate the re-identification risk of the records in the sample of the social security data. Therefore, it is assumed that the intruder does not know which observations are linked by the social security number. At first the data were inspected visually and the re-identification risk was calculated by the uniqueness approach. Secondly, a simple distance-estimation is used to "re-identify" the 273 persons who are in both data sets, and finally a cluster-algorithm is applied.

The uniqueness approach is one of the key concepts used in statistical disclosure control for micro-data [10]. An individual, who is the sole possessor of a certain combination of values for a given set of key-variables within a population (a population unique), is at particular risk of identification if these key-variables are present in the microdata file. Observations that possess rarely occurring combinations of key-variables are also endangered if the intruder has some additional knowledge that is not considered in the analysis [15].

On the basis of the uniqueness approach three measures were calculated: the proportion of sample uniques, the proportion of population uniques and the re-identification risk defined as [15]:

$$R = 1 - \exp(-fn(U)f_a)$$

where f is the sample fraction, $n(U)$ is the number of unique persons in the population and f_a is the proportion of persons for whom an intruder has additional information. In the following analysis it is assumed that $f_a = 1$; this means that an intruder has additional information for nearly all persons in the population (in our case: the relevant cohort).

First the uniqueness approach was applied to the IAB-employment sample. For making the variables comparable, in the most cases the MPI data have to be aggregated to the categories of the IAB data. The following 14 variables (see Table 1) were used: occupation (327 categories), birth cohort (two categories), daily earnings (7 categories), sex (dichotomous), land of the Federal Republic of Germany (nominal scaled: 11), nationality (dichotomous), schooling/training (nominal: 3), occupational status (nominal: 4), part time (dichotomous), interruption in the working life (dichotomous), number of months in the working life (count variable: maximum 24 months), duration of the last employment (count variable: maximum 24 months), martial status (dichotomous), number of children (count variable). The income variable was classified to seven relatively broad categories, because this variable is strongly affected by the measurement-differences (Fig. 2). Nevertheless the classes assure that most of the 273 observations that are surely in both data sets are identical in nearly all variables (Fig. 3). The number of identical values in both data sets is mostly over 180 (out of the 273), which means the two data

[3] For firms a totally different situation arises [5].

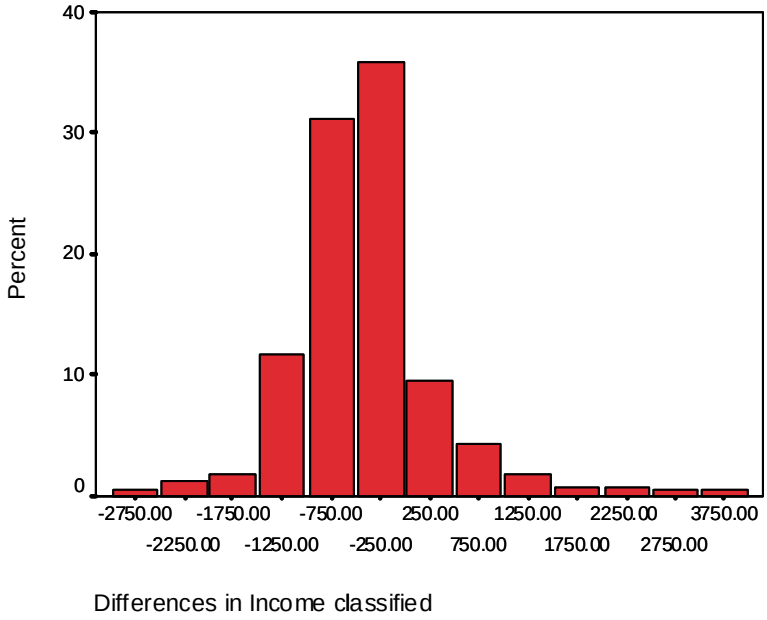

Fig. 2. Differences in Income classified between the MPI and IAB data.

generating processes came mostly to the same result (Table 3). This means that the implicit assumption of error free variables connected with the uniqueness approach is better fulfilled than expected a priori. An exception – as mentioned before – is income, if it is used as a continuous variable.

The results for the total population ($n = 934\ 152$) show that the proportion of unique persons is about 22.5%. The re-identification risk – the probability that at least one person of the MPI-sample will be re-identified – is near one, if it is assumed that an intruder has additional information about most of the people in the population. Taking the sample of the IAB-data as the total population ($n = 19\ 353$) the proportion of unique persons increases to nearly 69% and the re-identification risk for the MPI-sample is one. If it is assumed that the MPI-sample is the data set containing the additional information the re-identification risk for the IAB-sample is about 0.99. Obviously the value increases to one if the whole population would be disseminated. All of the following results have to be measured on these reference values.

Secondly, re-identification experiments were undertaken by comparing the two data sets by a simple distance-criterion [4], assuming that an intruder knows that the 273 observations of the MPI-Data can be found in the IAB-Data (response knowledge). The results show that about 10% of the distances between the real twins are smaller than all other distances. In nearly all other cases more than two distances were smaller than the distance between the original pairs. If a similar analysis is performed, with all observations in the sample of the IAB-employment database, less than a half percent of the real twins have the smallest distance.[4] Calculating a distance criterion can be seen as one step in cluster analysis. Therefore a more sophisticated cluster analysis was used in the third step.

[4]Measuring the re-identification risk in this way stands in line with the work of Paaß/Wauschkuhn [13] who measured the effectiveness of adding random noise to personal data, and Müller et al. [11,12] who tried to "re-identify" persons on the basis of two different real data sets. This method is also useful for measuring the effectiveness of anonymisation by perturbation methods like adding noise and micro-aggregation for business data [4,5].

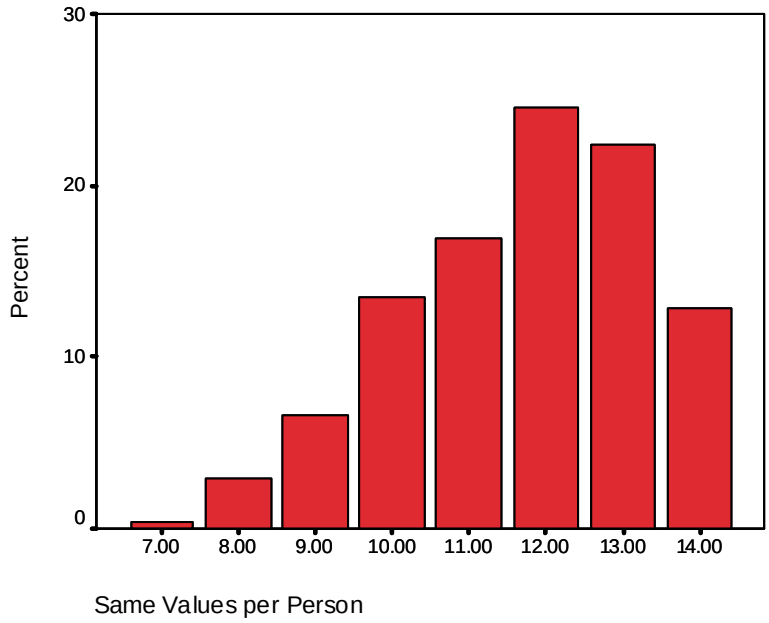

Fig. 3. Number of identical values compared in the both data sets.

Finally, standard cluster methods are applied to the data [1], assuming again that the intruder has response knowledge of the 273 observations in both data sets. We used a k-means cluster algorithm (SPSS-Quick Cluster, [14]: 448–449) with the following 14 variables: occupation (327 categories, treated as interval scaled), birth cohort (dichotomous), daily earnings (interval scaled), sex (dichotomous), land of the Federal Republic of Germany (nominal scaled: 11), nationality (dichotomous), schooling/training (nominal: 3), occupational status (nominal: 4), part time (dichotomous), interruption in the working life (dichotomous), number of months in the working life (count variable maximum 24 months), duration of the last employment (count variable: maximum 24 months), martial status (dichotomous), number of children (count variable). The cluster procedure consists of two steps:

– *Step 1:* Cluster analysis of the MPI data set

  Contrary to the usual application of the clustering procedures in the used programme we tried to have as many clusters as possible in the result. Under ideal conditions, the number of clusters should be equal to the number of persons so that each of the 273 persons builds one cluster and a perfect match is theoretically possible. The result of the first step was that we have 228 clusters for the second step. Each person builds one cluster, 45 cases were eliminated due to missing values (listwise deletion of missing values).[5]

– *Step 2:* Assignment of cases from the IAB data set to the clusters obtained for the MPI data set

  In the k-means cluster algorithm we have deterministic assignments of all cases to the clusters. So we were able to calculate the percentage of correct assignments. The result is that 37.7% of all persons are in the correct cluster, which means we assigned the person of the IAB data set to his real twin of the MPI data set.

[5] SPSS Quick-Cluster offers two possibilities for missing values: pairwise and listwise. The pairwise method is not practicable in this application, because the clusters are used as centers in the next step and centres must be free from missing values.

Table 3
Descriptive Statistics and number of cases with identical attribute values in both data sets

	Means of Variables in MPI	Means of Variables in IAB	Number of Identical values in MPI and IAB
Income	4513.26	4651.91	2
Year of birth	67.38	67.38	273
Sex	1.38	1.38	273
Occupation	630.40	621.95	151
Region	5.79	6.80	262
Nationality –Dummy (one, if German)	0.98	0.98	272
Schooling / Training	1.53	1.56	215
Occupational Status	3.90	3.57	215
Part Time Work	0.10	0.10	261
Interruption in the working life	0.36	0.40	255
Number of months in the working life	19.29	18.01	187
Duration of the last employment	19.06	17.81	189
Martial status	0.43	0.37	238
Number of children	0.55	0.05	183

Note: Some of the means are more or less obvious, but showing the means gives a short impression of the difference between the two data sets.

As a second measure we were looking at the nearest person in the IAB-data set to the cluster centroid, which is in our case identical with one person in the MPI data set. Taking this measure 25 persons are the nearest neighbours to themselves. So 9.2% of all persons in the MPI-data set can be re-identified via this cluster algorithm. The results can be improved by weighting the variables. If occupation is weighted by the factor 10, 39 cases (14.3%) can be re-identified.

4. Conclusion

Summarising the analysis shows that a re-identification may be possible by a standard-cluster analysis or a simple distance criterion if an intruder has very high additional information. For instance, it is assumed that the intruder has detailed information about more than 10 variables and response knowledge for the IAB-employment-sample (cluster analysis with a modified criterion for the optimal number of clusters) or response knowledge for both data sets (simple distance criterion). This confirms the conclusion of the first analysis by a simple uniqueness approach.

Additionally the number of re-identifiable persons is remarkably high although the proportion of re-identifiable persons is less than expected on the basis of the uniqueness-approach. This is caused by measurement errors in the retrospective survey and the special data generating process underlying the data of the social security system. It stresses that measurement errors have different implications in various analysis and that knowledge on the data generating process and the development of special methods can be important for an intruder, e.g. in cluster analysis results can be improved by weighting to their quality and developing tests for judging whether two cases are real twins.

The analysis leads to the conclusion that re-identification risks on real data sets need to be evaluated systematically taking into account the methods empirical researchers usually use and that personal data stemming from registers should be anonymised carefully. If a high level of additional information is available, it could be necessary to use perturbation methods like adding noise or switching categories. In Germany these were not used for personal data in the past.

References

[1] J. Bacher, *Clusteranalyse*, (2nd ed.), Oldenbourg, München, Wien, 1996.
[2] S. Bender, A. Haas and C. Klose, *The IAB-Employment Subsample – Opportunities for Analysis Provided by the Anonymised Subsample*, IZA Discussion Paper No. 117, IZA, Bonn, 2000.
[3] U. Blien, W. Müller and H. Wirth, Identification Risk for Microdata stemming from Official Statistics, *Statistica Neerlandica* **46**(1) (1992), 69–82.
[4] R. Brand, Anonymität von Betriebsdaten, Beiträge zur Arbeitsmarkt- und Berufsforschung (BeitrAB) 237, Bundesanstalt für Arbeit, Nürnberg, 2000.
[5] R. Brand, S. Bender and S. Kohaut, Possibilities for the Creation of a Scientific-Use File for the IAB-Establishment-Panel. Statistical Office of the European Communities (eds): Statistical Data Confidentiality – Proceedings of the Joint Eurostat/UN-ECE Work Session on Statistical Data Confidentiality held in Thessaloniki in March 1999, Eurogramme, 1999, 57–74.
[6] H. Brückner and K.U. Mayer, Lebensverläufe und gesellschaftlicher Wandel. Konzeption, Design und Methodik der Erhebung von Lebensverläufen der Geburtsjahrgänge 1954–1956 und 1959–1961. Teile I-III, Materialien aus der Bildungsforschung Nr. 48, Max-Planck-Institut für Bildungsforschung (MPIfB), Berlin, 1995.
[7] M. Corsten and S. Hillmert, Ausbildungs- und Berufsverläufe der Geburtskohorten 1964 und 1971 in Westdeutschland. Arbeitspapier Nr. 1 des Projektes: Ausbildungs- und Berufsverläufe der Geburtskohorten 1964 und 1971 in Westdeutschland, Berlin (MPIfB), 2001.
[8] G. Duncan and D. Lambert, The Risk of Disclosure for Microdata, *Journal of Business & Economic Statistics* **7** (1989), 207–217.
[9] P. Knoche, Factual anonymity of Microdata from Household and Person-related Surveys – the Release of Microdata for Scientific Purposes, Proceedings of the International Symposium on Statistical Confidentiality, Dublin, Eurostat, 1993, 407–413.
[10] R.J. Mokken, P. Kooiman, J. Pannekoek and L. Willlenborg, Disclosure Risks for Microdata, *Statistica Neerlandica* **46**(1) (1992), 49–67.
[11] W. Müller, U. Blien, P. Knoche and H. Wirth et al., Die faktische Anonymitaet von Mikrodaten, Forum der Bundesstatistik, Metzler-Poeschel, Stuttgart, 1991.
[12] W. Müller, U. Blien and H. Wirth, Identification Risks of Microdata – Evidence from Experimental Studies, *Sociological Methods & Research* **24** (1995), 131–157.
[13] G. Paaß and U. Wauschkuhn, Datenzugang, Datenschutz und Anonymisierung: Analysepotential von anonymisierten Individualdaten, Berichte der Gesellschaft für Mathematik und Datenverarbeitung Nr. 148, Oldenbourg, München, Wien, 1984.
[14] SPSS Inc., SPSS 7.5 Statistical Algorithms, Chicago (http://www.spss.com/tech/stat), 2001.
[15] L. Willenborg and T. de Waal, *Statistical Disclosure Control in Practice*, Springer, New York, 1996.

Mr. Johann Bacher studied sociology and statistics at the University of Linz (Austria), where he received his diploma in 1983. He finished his doctoral thesis on factor analysis in 1986. From 1989–1998 he was research assistant at the Department of Sociology at the University of Linz, from 1986 to 1989 at the University of Erlangen-Nuremberg. Since 1998 he is head of the Chair of Sociology at the University of Erlangen-Nuremberg. His research interests cover methods, data analysis, social inequality and prejudices.

Mr. Stefan Bender studied sociology and statistics at the University of Mannheim, where he received his diploma in 1990. From 1990–1992 he was lecturer at the Department of Sociology, University of Mannheim. Since 1992 he is research assistant at the section "Regional Labour Market Research and Analytical Statistics" of the Institute for Employment Research (IAB), Nuremberg. His research interests cover labour economics, social mobility, anonymisation and data analysis.

Ms. Ruth Brand acquired her graduate in economics in 1992 at the University of Hanover with priorities in econometrics, insurance economics and labour economics and a PhD from the University of Hanover for a thesis entitled Anonymität von Betriebsdaten [anonymisising of business data] in 1999. From 1993–1999 she was a lecturer at the University of Hanover. Since 1999 she has been employed at the federal employment office (BA) at the subdivision of statistics. Her fields in research are data protection, statistical methods and labour dynamics.