STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint ECE/Eurostat Work Session on Statistical Data Confidentiality
(Skopje, The former Yugoslav Republic of Macedonia, 14-16 March 2001)

Working Paper No. 4
English only

Topic I: Application of statistical disclosure control methodology and software in business statistics and social and demographic statistics

RE-IDENTIFYING REGISTER DATA BY SURVEY DATA: AN EMPIRICAL STUDY

Contributed paper

Submitted by the Institute for Employment Research (IAB), Germany[1]

I. INTRODUCTION AND DESCRIPTION OF THE DATA

1. In this paper re-identification risks for register data are examined by matching a sample of register data with survey data. 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 registered by the social security system to the social security agencies. Exact daily information on employment are 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 (Bender et al. 2000).

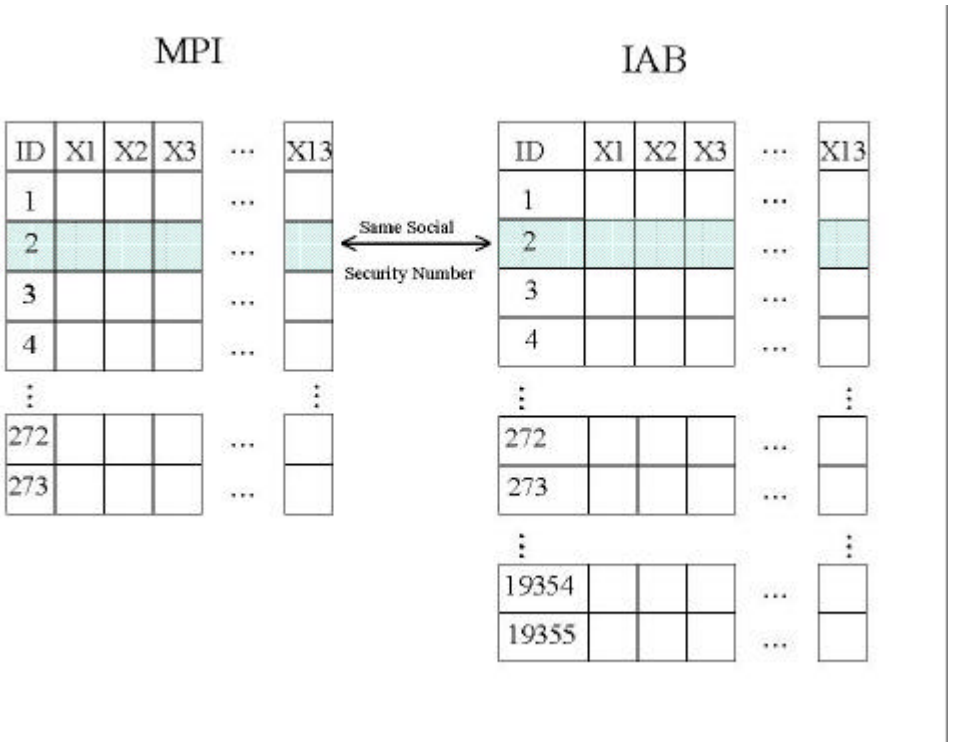
2. The second data set is the German Life History Study conducted by the Max Planck Institute for Human Development (MPI-data set), which is a retrospective survey that evaluates quantified life histories – measured in months - for different dimensions (e.g. schooling, apprenticeship, employment, partnership, family, housing). The life course protocols are edited and corrected by using taped recordings. In cooperation with the IAB the MPI collected interviews of nearly 3,000 women and men born in 1964 or 1971 in Western Germany. About 80% gave their permission to match their responses with the data stored by the social security system, but only for approximately 800 persons, the insurance number is available in the data (for a detailed description of the German Life History Studie, see Brückner, H. / Mayer, K. U. 1995 or for an analysis Mayer, K. U. / Carroll, G. R. 1987).

3. For our analysis we take a subsample in the following form: we matched those persons who were employed in Western Germany in November 1997 by their insurance number. 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 are drawn (n=19.082) to which the 273 real twins are added (IAB-data set). At the end we have two data sets (figure 1):

- MPI-data set, which contains information of the 273 real twins, and
- IAB-data set, which contains information of 19.082 persons, for those we have only information in the IAB-data and 273 real twins (n=19.355).

[1] Prepared by Johann Bacher, Stefan Bender and Ruth Brand.

Figure 1: Relationship between the two data sets



4. 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. 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 the different data generating processes or 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 or process produced data sets.

II. MEASURING THE RE-IDENTIFICATION-RISK

5. Then several methods are used 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.

6. At first the data were inspected visually and the re-identification risk was calculated by the uniqueness approach. Second, a simple distance-estimation is used to "re-identify" the 273 persons who are in both data sets, third a cluster-algorithm is applied.

7. First the uniqueness approach was applied to the IAB-employment sample. The following 12 variables 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 (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), martial status (dichotomous), martial status (count variable). The income variable was classified to seven relatively broad categories, because this variable is strongly affected by the measurement-differences (figure 2). Nevertheless the classes assure that most of the 273 observations that are surely in both data sets are identical in nearly all variables. This means that the implicit assumption of error free variables connected with the uniqueness approach is fulfilled better than expected a priori (table 2). 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 (Willenborg /de Waal 1996), 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 is one. On this reference value all following results have to be measured.

8. Second "re-identification-experiments" were undertaken by comparing the two data sets by a simple distance-criterion (Brand 2000), 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[2].

9. Third standard cluster methods are applied to the data (Bacher 1994), assuming again that the intruder has response knowledge for the 273 observations in both data sets. We used a k-means cluster algorithm (SPSS) with the following 13 variables: birth cohort (dichotomous), daily earnings (interval scaled), sex (dichotomous), land of the federal republic of Germany (nominal scaled: 11), nationality (dichotomous), schooling (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), martial status (dichotomous), number of children (count variable). The cluster procedure consists of two steps:

- Step 1: Cluster analysis of 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 person builds one cluster and a perfect match is theoretically possible. The result of the first step was that we have 225 cluster for the second step.

- 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 36 % of all persons are in the correct cluster, which means we assigned the person of the IAB data set to her real twin of the MPI data set.

10. As a second measure we were looking at the nearest person the IAB-data set to the cluster centroid, which is mostly identical to the values of one person in the MPI data set. Taking this measure only 13 persons are the nearest neighbours to themselves. So only 5.8 % of all persons in the MPI-data set can be re-identified via this cluster algorithm. One reason for this bad result may be that occupation was not included.

III. SUMMARY

11. 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.

12. Nevertheless the number of re-identifiable persons is not very high and the proportion of re-identifiable persons is less than expected on the basis of the uniqueness-approach. One reason for this is the differences in the treatment of the occupation–variable. The number of unique persons in this analysis is quite higher than without using the occupation-variable. However the basic results will not be affected

[2] Measuring the re-identification risk in this way stands in line with the work of Paas / Wauschkuhn (1984) who measured the effectiveness of adding random noise to personal data and Müller et al. (1991, 1995) who tried to „re-identify" persons on the basis of two different real data sets. This design is also useful for measuring the effectiveness of anonymisation by perturbation methods like adding noise and micro-aggregation for business data (Brand et al. 1999, Brand 2000).

by a different way of including this variable: Re-identification risks on real data sets need to be evaluated systematically taking into account those methods empirical researchers usually use.

References

Bacher, J. (1994): Clusteranalyse, München, Wien: Oldenbourg.

Bender, S.; Haas, A. and Klose, C. (2000): The IAB-Employment Subsample - Opportunities for Analysis Provided by the Anonymised Subsample, IZA Discussion Paper No. 117, IZA, Bonn.

Brand, R. (2000): Anonymität von Betriebsdaten, Beiträge zur Arbeitsmarkt- und Berufsforschung (BeitrAB) 237, Nürnberg: Bundesanstalt für Arbeit.

Brand, R.; Bender, S. and Kohaut, S. (1999): 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: p. 57-74.

Brückner, H. and Mayer, K. U. (1995): 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, Berlin: Max-Planck-Institut für Bildungsforschung.

Mayer, K. U. and Carroll, G. R. (1987): Jobs and Classes: Structural Constraints on Career Mobility, In: European Sociological Review, 3, p. 14-38.

Müller, W.; Blien, U.; Knoche, P.; Wirth, H. et al. (1991): Die faktische Anonymitaet von Mikrodaten, Forum der Bundesstatistik, Stuttgart: Metzler-Poeschel.

Müller, W.; Blien, U. and Wirth, H. (1995): Identification Risks of Microdata - Evidence from Experimental Studies; In: Sociological Methods & Research 24, p. 131-157.

Paaß, G. and U. Wauschkuhn (1984): Datenzugang, Datenschutz und Anonymisierung: Analysepotential von anonymisierten Individualdaten, Berichte der Gesellschaft für Mathematik und Datenverarbeitung Nr. 148, München, Wien: Oldenbourg.

Willenborg, L. and T. de Waal (1996): Statistical Disclosure Control in Practice, New York: Springer.

ANNEX

Table 1: List of variables

Variable	Remark	MPI (retrospective Survey)	IAB (social security data sets, register entries base on notifications by the employer)
Income	Daily earnings	All occupations	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		
Sex			
Highest secondary school qualification	Secondary modern school, A-Levels		Highest secondary school qualification known by the employer
Training	University/polytechnic foreman, training on the job		
Hours of work per week at the main employment		Number of hours	Part-time-work: less than 18 hours per week, 18 hours ore more
Hours of work per week at training period.		Number of hours	Not notified for apprentices
Blue collar/White collar at the main employment.			Blue-collar-annuity insurance/white collar annuity insurance
Profession/occupation		Open specification, manual coded by the researcher	Notified is the occupation at work not the educated profession
Region	Federal state		
Work experience	Work experience in month		
Unemployment (Interruption in the working life)	During the last three years: at least three month unemployment		
Duration of the last employment (Change of establishment)	During the last three years: changes during the last three years		
Family status			Family status notified by the employer
Number of children		Number of children living in the same household	Number of children notified by the employer
Nationality			Nationality or country of origin

Table 2: Descriptive Statistics and number of cases with identical attribute values in both data sets

	means MPI	means IAB	means(MPI-IAB)	Number of Identical values in MPI and IAB-data
Income	4513,26	4651,91	-138,65	2
Year of birth	67,38	67,38	0,00	273
Sex	1,38	1,38	0,00	273
Profession/occupation	630,40	621,95	8,45	151
Region	5,79	6,80	-1,01	262
Nationality -Dummy (one, if german)	0,9780	0,9817	-0,00366	272
Highest secondary school qualification	1,5311	1,5641	-0,0330	215
Training	2,55	2,05	0,49	220
Blue collar/White collar at the main employment.	3,90	3,57	0,34	215
Hours of work per week at the main employment	0,1026	0,0952	0,0073	261
Unemployment (Interruption in the working life)	0,3553	0,3993	-0,0440	255
Work experience (in the last 3 years)	19,29	18,01	1,29	187
Duration of the last employment	19,06	17,81	1,26	189
Family status	0,4322	0,3700	0,0623	238
Number of children	0,5531	0,0476	0,5055	183

Figure 2: Differences in Income classified

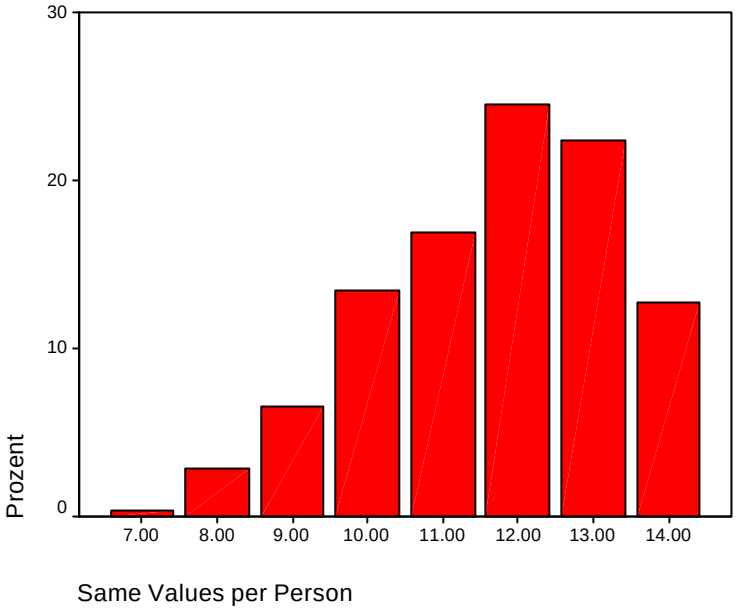


Figure 3: Number of identical values