WP.10
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (i): Microdata

ANONYMISATION OF LINKED EMPLOYER EMPLOYEE DATASETS USING THE EXAMPLE OF THE GERMAN STRUCTURE OF EARNINGS SURVEY

Supporting Paper

Prepared by Hans-Peter Hafner (Statistical Office of Hesse, Germany) and
Rainer Lenz (University of Applied Sciences Mainz, Germany)

Anonymisation of Linked Employer Employee Datasets using the example of the German Structure of Earnings Survey.

Hans–Peter Hafner[*], Rainer Lenz[**]

[*] Research Data Centre of the Statistical Offices of the Länder, Statistical Office of Hesse, Rheinstr. 35/37, 65185 Wiesbaden, Germany
(hhafner@statistik-hessen.de)

[**] Department I: architecture, civil engineering and geoinformatics, University of Applied Sciences Mainz, Holzstr. 36, 55116 Mainz, Germany
(rainer.lenz@fh-mainz.de)

Abstract. The anonymisation of linked employer employee datasets constitutes a special problem for data producers. Concerning the employees there is generally less risk of reidentification, but their information can be used to identify the employer. We present a strategy that permits to measure dependencies between employer and employee data, to evaluate whether these dependencies have an impact on the reidentification risk of the employer and, if necessary, to anonymise the data of the employees in such a manner that the reidentification of the employer is very complicated. Finally, we embed this strategy in the generation process for a scientific use file of the German Structure of Earnings Survey 2001.

1 Introduction

Linked employer employee datasets (LEED) enable labour market researchers to split observed effects in one fraction caused by the employer and one fraction dependent on the employee. Since the middle of the 1990ies the number of analyses in this field has escalated. Abowd and Kramarz (1999) provide an overview on projects executed during the 90ies and on datasets from 17 countries that were available at that time.

LEED for Germany that are currently available to the scientific community are the Linked Employer Employee Data of the Institute for Employment Research (LIAB) and the Structure of Earnings Survey of the Federal Statistical Office and the statistical offices of the Länder (federal states). A description of the LIAB and selected studies conducted with it can be found in Alda et al. (2005); an overview on the Structure of Earnings Survey and related studies is provided by

Hafner and Lenz (2007).
As interesting as LEED are to the science community, it is very complicated for the data producers to generate anonymised scientific use files from such sources so that researchers can work with them in their institutes.

In this paper we propose a procedure, which besides classical information reducing methods, applies only selective one-dimensional microaggregation to especially sensitive variables. We show that thereby the reidentification risk associated with the data is reduced.
Chapter 2 summarises the thoughts that have to be considered for the anonymisation of the employer data. Chapter 3 deals with the information about the employees. Here we present a method that guarantees that the variables of the employees do not increase the disclosure risk of the employer. Following an overview on the methodology and the attributes of the Structure of Earnings Survey (Chapter 4) we apply our procedure to the German dataset of this survey of the year 2001.

2 Anonymisation of the Employer Data

In order to reidentify an enterprise, a data intruder needs additional knowledge, for example from commercial databases. This additional knowledge must have attributes in common with the target file (key variables). For Germany, the greatest electronically available resource is the so called Markus database. Details can be found at http://www.creditreform.de/.
The risk of reidentification can now be determined by means of matching experiments between the target file and the additional knowledge. The aim of the data intruder is to decide whether or not the pair $(a, b) \in A \times B$ of records belongs to the same employer. In a non-technical way, the concept of matching may be introduced as a way of bringing together pieces of information in pairs from two records taken from different data sources. For this purpose, a reasonable concept of similarity is necessary. Roughly spoken, the greatest possible similarity between two records turns into identity if the considered records correspond with regard to all key variables. In the case of small deviations of the key variables, two objects are felt to be strongly related, so that the matching result essentially depends on the concept of similarity. For technical details see Lenz (2006).
In recent years, regarding German business statistics, often the value 0.5 was accepted as the upper bound for the reidentification risk, provided that this value is reached only for a few parts of the file, for example for large companies in low frequented branches of economic activity, see Ronning et al. (2005). When evaluating the risk one has to consider that the calculation presumes that a data intruder has some knowledge of participation about an enterprise. Therefore the risk in sample surveys is reduced by a factor corresponding to the sample fraction of the stratum.

In practice, the risk is minimised by combining especially vulnerable classes of categorial variables with others and by applying data perturbing procedures like microaggregation to numerical variables.

3 Anonymisation of the Employee Data

In most cases the risk of reidentification for employees is negligible since there is no systematic additional knowledge. Furthermore we restrict our thoughts to sample surveys. Hence a data intruder has no participation knowledge about a person, s.t. the information about the employees is sufficiently anonymised when taken alone. However, it might be possible to draw conclusions from it about the enterprises so that the anonymisation made can be reversed in parts by a data intruder. To formalise these thoughts we need some notations.

Let A be an employer file whose attributes are the random variables $S_1, \ldots, S_n$, and B an employee file with attributes $T_1, \ldots, T_m$. $C = (A, B)$ is a *linked employer employee file* if it is possible to assign the employees covered in B to the employers covered in A.

An attribute Y of an employee file that is independent of an attribute X of the employer file would not yield further insights to a data intruder. But in practice absolute independence is rather the exception. Hence we need measures to calculate the degree of dependence between two variables. In doing so we have to differentiate by the scale of the attributes:

- Both attributes are metric. Then Pearson's correlation coefficient measures the degree of dependency.

- Both attributes are ordinal. Then one can use Spearman's rank correlation coefficient. The same holds if one attribute is ordinal and the other metric.

- Both attributes are nominal. Then Cramer's V is an adequate measure, which can also be applied if one attribute is ordinal and the other nominal.

- One attribute is nominal, the other metric. In this situation one can use the measure of association η. In contrast to the other measures, η is not symmetric. That means the metric attribute is the dependent variable while the nominal attribute is the independent one. Let $X = (x_1, \ldots, x_l)$ be the nominal attribute, n_i the number of observations in category i, $\overline{y}$ the mean of Y and $\overline{y}_i$ the mean of Y in category i. Then we define

$$\eta = \sqrt{\frac{\sum_{i=1}^{l} \sum_{j=1}^{n_i} (y_{ij} - \overline{y})^2 - \sum_{i=1}^{l} \sum_{j=1}^{n_i} (y_{ij} - \overline{y_i})^2}{\sum_{i=1}^{l} \sum_{j=1}^{n_i} (y_{ij} - \overline{y})^2}} \tag{1}$$

In empirical research it is common to assume a strong association between the two variables whenever the measure exceeds a value of 0.3. As regards the application to datasets of official statistics, the determination of the bound will depend on the need for data protection. There must be more protection if the data are very sensitive or if the benefit of a reidentification seems very high. This benefit is influenced by the age of the data and by the availability of the information through other sources.

To test which combinations of variables of the linked employer employee file $C = (A, B)$ are especially vulnerable, we compute the measures of association for all key variables of A and all variables of B.

For the rest of this chapter, let S be an attribute of the employer file A and T an attribute of the employee file B so that the value of the measure of association between S and T is above the predefined bound. To simplify matters, we suppose that S and T are both categorical, which can always be achieved by grouping values. Let s be the number of categories of S, t the number of categories of T. Furthermore, A^* should be the anonymised version of A, and S^* the attribute that originated from S in A^*.

At this point it seems necessary to meet some assumptions about the behaviour of a data intruder. These are of pure theoretic nature since the existence of a data intruder is a hypothetical construct.

Below we describe in three steps how a data intruder might proceed in order to find out more about an employer by using the association between S and T.

Step 1: Since the anonymisation of A took place without using data modifying methods, the intruder knows (because of the description of the anonymisation) for every value x^* of S^* the set $X = \{x_1, \ldots, x_k\}$ containing the corresponding original value x.

Step 2: The intruder manages to get the marginal distributions of T for every category of S. Maybe he finds them in a publication of the statistical office or he asks for calculation via remote data access.

Step 3: The intruder compares every employer's distribution of T with the marginal distributions of T for $x_1, \ldots, x_k$ and he chooses that x_i which presents the smallest differences in respect to the distribution of T.

The adequate statistical procedure for step 3 is to perform a test of goodness of fit. In the case of discrete variables the χ^2 test is the most common tool. Using this tool, the observed sample is analysed as to whether it can be a random sample of a specific distribution by comparing the observed and the expected frequencies. Let $e_i, i = 1, \ldots, t$, be the expected frequency for category i of T and f_i the observed frequency. Then the well known test statistic χ^2 is obtained by

$$\chi^2 = \sum_{i=1}^{t} (f_i - e_i)^2 / e_i \tag{2}$$

The null hypothesis indicates that the observed sample originates from the assumed distribution. It is rejected if the value of χ^2 is greater than the quantile for the chosen level of significance.
The distribution function of (5) fits asymptotically the distribution function of the χ^2 function with $t-1$ degrees of freedom. The rule of thumb mostly mentioned for the application of the χ^2 test is that for at least 80 percent of the categories the expected frequencies should be 5 or more and that the expected frequencies of the other categories should be at least 1. Koehler and Larntz (1980) show that the appoximation is suitable even for smaller expected frequencies provided that the square of the number of observations at least equals the number of categories multiplied by 10.
To improve the goodness of the prediction, one can combine categories whose fraction is very low in all marginal distributions that have to be tested. A modification of the χ^2 test that can be applied to small samples and small expected frequencies has been developed by Haldane. He does not compare the test statistic with the asymptotic χ^2 distribution, but with the exact distribution that holds under the null hypothesis, for instance see Bortz et al. (1990).
Let us now return to the data intruder. We assume that he has conducted his tests and calculated the statistics. Before he makes his decision, he has to determine a level of significance. We suppose that he can live with a probability of error of 20 percent; that is he chooses $\alpha = 0.2$. If the null hypothesis is accepted for exactly one value the decision is clear. The case that the null hypothesis is accepted for more than one value should not occour in practice; but it will happen very often that none of the alternatives is accepted since the χ^2 test almost always rejects the null hypothesis if the sample size is large. In this case one can decide on the basis of the corrected contingency coefficient C_{corr}.

The χ^2 test and the contingency coefficient provide global measures for the match of distributions. If no definite decision can be made on this basis it is additionally recommended that the components of contingency are compared. The component of contingency c_i for the category $i \in \{1, \ldots, t\}$ is defined as

$$c_i = \begin{cases} +\sqrt{\frac{(f_i-e_i)^2/e_i}{n+((f_i-e_i)^2/e_i)}} & \text{if } f_i - e_i \geq 0 \\ \\ -\sqrt{\frac{(f_i-e_i)^2/e_i}{n+((f_i-e_i)^2/e_i)}} & \text{if } f_i - e_i < 0. \end{cases}$$

A component of contingency with value 0 corresponds to a perfect match of the observed sample with the testing distribution with respect to this category; a negative (positive) component indicates a lower (higher) fraction of the category in the sample. The data intruder will now look at the components of contingency of the categories of T which are typical for the alternative x_i that should be tested. We call a category d of T *characterising* for x_i if the following conditions are satisfied:

1. The fraction of d in the marginal distribution of T given x_i exceeds a value f.
2. The fraction in category x_i exceeds a bound g in the distribution of d over the categories of S.
3. There is at least one alternative $x_j \neq x_i$ such that $|P(s|x_i) - P(s|x_j)| > h$ where h is a specified bound.

The last condition excludes that the fractions of the category are nearly equal for all alternatives. Such a category would not contribute to the decision-making.
Regarding the selection of f, g and h, one has to look at the distributions of the attributes involved; therefore, no universally valid statements are possible. After the selection of the characterising categories, the data intruder looks at the corresponding components of contingency. He may either sum up all these components separately for every alternative or sum up only the positive components in each case. An argument for summing up only the positive components is that not in every case all characterising categories are represented equally well. A value appearing above the average of some characterising categories is often a better indicator as we will see in our application in chapter 5. If the global contingency coefficient and the analysis of the single components yield the same result, the decision comes down to this alternative. In all other cases there is no sufficient confidence.
Finally, we have to evaluate the thoughts outlined above from the point of view of the data producers. Let $r_1, \ldots, r_s$ be the risks of reidentification for the categories $1, \ldots, s$ of S in the original employer file A and let r^* be the risk of reidentification for the anonymised category x^*. Furthermore, let p be the fraction of employers whose original value x_i can be derived from x^* with the help of the attribute T and the method described above. For these employers the risk of reidentification is as high as if no anonymisation with respect to S had taken place. Hence, the risk of reidentification for an employer of category x_i adds up to

$$pr_i + (1-p)r^*. \tag{3}$$

If (7) exceeds 0.5, the categories of T that contributed most to the disclosure of x_i (that means, the categories with the highest fractions of positive components of contingency) have to be combined with other categories. This subsumption can be carried out for the complete dataset or, alternatively, only for those employers with value x^* for S.

4 The Structure of Earnings Survey

Based on an EC regulation of 1999, the survey is held in all EU countries every four years, so that the data produced are comparable all over Europe. As most countries conducted the latest survey for 2002, the next one will be performed for 2006.

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).The SES is a two-stage sample survey. In the first sampling stage, a stratified random sample is drawn from the local units. At the second stage, the employees to be included from the selected local units are determined through the personal identification number shown on the staff lists. For 2001, a total of a good 22,000 local units supplied data on over 845,000 employees.

There are separate questionnaires for data on the local unit and one each (or several for larger local units) for white-collar and blue-collar employees. Further information on the methodology and variables of the 2001 SES is contained in Frank-Bosch (2003) and in the metadata provided on the web site of the research data centres of the statistical offices of the Federation and the Länder (http://dok.fdz-metadaten.de/6/62/621/621110/erheb/200100/).

5 Anonymisation of the German Structure of Earnings Survey 2001

Since spring 2005 the research data centres of the Federal Statistical Office and the statistical offices of the Länder have conducted a project with the aim to generate a scientific use file of the German structure of earnings survey taken in 2001. The project has been concluded in autumn 2006 with the publication of the file. Scientists participated in an advisory capacity in the conception of the anonymised dataset to ensure that the result will be of interest to a broad circle of users.

At first, the key question was which regional units should be displayed. Two alternatives with five and eight regions consisting of adjacente federal states were tested. Depending on the model used, some 30 to 40 economic sectors were displayed. Furthermore, the number of employees of a company was microaggregated if a company had at least a thousand employees or if it was among the three largest companies of the economic sector in the region. Each group for microaggregation consisted of at least three companies.

The key variables which the employer dataset had in common with commercial databases were the region, the economic sector, the number of employees of the enterprise and the influence of the public sector. The last-mentioned attribute can be compared with *partner - agency, state, administration* in the Markus database. Using these attributes, matching experiments as described in chapter 2 were conducted to calculate the risks of reidentification for the several alternatives. It transpired that the risks for some economic sectors were too high when eight regions were displayed. Thus we opted for five regions and we lowered the threshold from which the number of employees was microaggregated to 500. It turned out that the attribute *participation of the public sector* was not critical with respect to reidentification; hence we could display the original value.

Now we will describe the anonymisation of the employee data more precisely following the scheme we developed in chapter 3. As an example we take the two combined economic sectors of the drapery / clothing trade and the leather industry. Of these sectors, 429 local units and 14,826 employees are contained in the survey.
At first we must examine which attributes of the employees have a strong association with the economic sector. Table 1 shows that the measure of association indicates a strong dependence on the economic sector only for the occupation class. Since all other values are far below 0.3 we conclude that there are no further variables with a strong relation to the economic sector. Hence we can carry out two χ^2 tests between the economic sector and the occupation class. First we test the observed distribution of the occupation classes against the distribution of the drapery / clothing trade, and then against the one of the leather industry. For that purpose we combine some occupation classes which contain only few

Table 1: Measures of association between the economic sector and the attributes of the employees

Sex	0.044
Wage Tax Class	0.045
Allowance for Children	0.043
Position in Job	0.119
Education	0.071
Type of Contract of Employment	0.022
Occupation Class (2-digit)	0.659
Paid Working Hours Total	0.003
Gross Earnings in Accounting Period	0.041
Extra Pay for Shift Work	0.077
Extra Pay for Night Work	0.113
Income Tax	0.035
Pension and Unemployment Insurance	0.048
Health and Care Insurance	0.055
Gross Annual Earnings	0.035
Supplementary Grants in the Reporting Year	0.051
Net Annual Earnings	0.033
Holiday Entitlement	0.003
Net Earnings in Accounting Period	0.043

employees so that we obtain 18 classes. According to Koehler and Larntz (1980) , the χ^2 distribution is a good approximation if the number of observations is at least $\sqrt{10 * 18} = 13.42$. Choosing $\alpha = 0.2$, one of the two null hypotheses is accepted in only 15 cases. In 12 of theses cases the prediction is correct, in the other cases the number of observations is smaller than 14. As expected, the χ^2 value taken alone is not a good predictor. For this reason, the next step consists of the calculation of the contingency coefficients. On the basis of these coefficients, the prediction is correct in 394 of 429 cases. Moreover, there is only a small difference in the fraction of correct predictions between the drapery / clothing trade and the leather industry. While for the drapery / clothing trade 92.6 percent of the assignments are correct, this applies to only 90.4 percent of the local units of the leather industry.

Before we start to calculate the components of contingency, we have to decide which occupation classes are characterising for the economic sectors under review. Tables 2 and 3 show that textile fabricators and textile producers doubtlessly are characterising occupation classes for the drapery / clothing trade, and leather producers, leather and coat fabricators for the leather industry. Furthermore, 73.8 % of all workers with spinning occupations and 76.8 % of the textile refiners are employed in the drapery / clothing industry, so that these occupations can also be regarded as characterising for this economic sector. For all other occupations listed in the tables below, the fraction of corresponding workers amounts to less than 5 % in the two economic sectors. In accordance with condition 2. for characterising categories, we leave these occupation classes out of consideration.

Table 2: Fractions of the most frequent occupation classes: Drapery / Clothing Trade

Textile Fabricators	19.3 %
Office Workers	14.6 %
Textile Producers	9.3 %
Product Inspectors, Shipping Finalisers	6.3 %
Technicians	5.5 %
Product Traders	5.5 %
Storekeepers, Warehousemen, Transport Workers	4.4 %
Spinning Occupations	4.4 %
Textile Refiners	3.0 %

If we add up only the positive coefficients we can make a prediction for 326 of the 429 local units. Out of 238 predictions for units of the drapery / clothing trade 225 are correct (94.5 %), out of 88 predictions for units of the leather industry 86 are correct (97.3 %). If we sum up all coefficients, we reach a correct prediction for only 56.6 % of the units of the drapery / clothing trade, while the fraction of correct predictions in the leather industry is 88.5 %. These figures suggest that the results are getting worse by using negative coefficients. Thus it

Table 3: Fractions of the most frequent occupation classes: Leather Industry

Leather Producers, Leather and Coat Fabricators	47.4 %
Office Workers	16.9 %
Product Traders	4.8 %
Technicians	3.9 %
Entrepeneurs, Organisers, Accountants	3.1 %
Plastics Fabricators	3.1 %
Storekeepers, Warehousemen, Transport Workers	3.0 %

might be better to use only the positive coefficients and to be content with fewer assignments. In exchange the risk of a misclassification is small.

If we combine the results of the analysis of the contingency coefficient and of the separate components and assign an economic sector to a local unit only if both predictions correspond, then there is a very small risk for the data intruder. He can make predictions for 82 local units of the leather industry and all are correct; for the drapery / clothing trade 181 of 190 (95.3 %) possible predictions are correct. Thus he can choose whether he wants to assign more units with a higher risk or fewer units with a lower risk.

We suppose the data intruder to be risk averse and assigns only units for which both analyses yield the same result. Then he can correctly assign 82 of the 104 units (78.9 %) of the leather industry.

As mentioned in section 2, matching experiments are applied in order to estimate the reidentification risk for employers of specific industries. Regarding the leather industry, the resulting risk is 0.61. If one joins the two sectors leather industry and drapery / clothing trade, the risk is reduced to 0.12. Hence, with (7) the overall risk for local units of the leather industry is estimated by 0.789 * 0.61 + 0.211 * 0.12 = 0.506. Since this value exceeds 0.5 we have to combine the characteristic occupation classes of the sectors of the drapery / clothing trade and leather industry.

The application shows that our methods can be applied to linked employer employee datasets in order to increase the data protection. However, further experience is needed to improve and to standardise the suggested methods so that they will be easier applicable and less time-consuming.

References

Abowd, J. M. and Kramarz, F. (1999) "The Analysis of Labor Markets Using Matched Employer-Employee Data", In: Ashenfelter, O., Card, D. (eds.): *Handbook of Labor Economics*, Amsterdam, vol. **3**, 2629–2733.

Alda, H., Bender, S. and Gartner, H. (2005) "The linked employer-employee dataset of the IAB (LIAB)", *IAB Discussion Paper* **06/2005**, Institute for

Employment Research, Nuremberg, Germany

Bortz, J., Lienert, G. A. and Boehnke, K. (1990) “Verteilungsfreie Methoden in der Biostatistik”, Berlin, Heidelberg: *Springer*

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

Hafner, H.-P. and Lenz, R. (2007) “Die Gehalts- und Lohnstrukturerhebung. Methodik, Datenzugang und Forschungspotential”, *discussion paper* **18**, Research data centres of the statistical offices of the federal and the states

Koehler, K. and Larntz, K. (1980) “An empirical investigation of goodness-of-fit statistics for sparse multinomials”, *JASA* **370 (75)**, 336–344

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

Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M. and Vorgrimler, D. (2005) “Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten”, *Statistik und Wissenschaft*, volume **4**, Statistisches Bundesamt