

**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**  
(Thessaloniki, Greece, 8-10 March 1999)

Working Paper No. 14  
English only

Topic (i): new applications of disclosure control methods

**POSSIBILITIES OF THE CREATION OF A SCIENTIFIC-USE-FILE  
FOR THE IAB-ESTABLISHMENT PANEL**

Prepared by Ruth Brand (University of Hannover, Germany)  
Stefan Bender and Susanne Kohaut (Institute for Employment Research (IAB), Germany)

**Contributed paper**

**I. INTRODUCTION**

1. Recently the demand side of the labour market and thereby firm data has gained some importance in applied social research, because millions of jobs are lacking in the OECD countries. The longitudinal observations of single firms over time help us to get an empirical microfoundation of macroeconomic employment analysis and policies. Better insight into firms' demand for labour and into the causes and consequences of different employment strategies of firms could not only improve labour market research but also labour market policies. Therefore more and more researchers ask for access to firm data.

2. In Germany, the legal requirements of data security do not allow the disclosure or release of any microdata stemming from official statistics without modifications. Disclosure control rules for microdata release exist for individual and household data only. The aim of the paper is to investigate into the possibilities of releasing business data on the basis of the IAB-Establishment-Panel. Therefore at first a short description of the data set is given. Second the current practice of microdata releases in Germany is described. Third the additional information for business data is roughly examined. In section four some estimations for the proportions of endangered population units are presented. In section five the possibilities of the creation of a scientific-use file are discussed. Because of the high proportions of endangered units it seems necessary to use data modifying techniques, that preserve several relevant statistics, i.e. the covariance-matrix. Therefore, a suitable approach seems to be the masking algorithm proposed by Sullivan (1989). The results of an experiment carried out with this approach are presented in this section, too.

**II. THE IAB-ESTABLISHMENT-PANEL**

3. As mentioned before, more and more researchers ask for access to firm data existing in institutions like the Institute for Employment Research at the Federal Employment Services. Register data as well as survey data on firm or establishment level are collected there. In Germany, only few firm or establishment surveys are carried out. One of the most comprehensive establishment surveys is the IAB-Establishment-Panel (Bellmann 1997). The IAB-Establishment-Panel surveys the same establishments every

year taken from all branches of industry and different size categories. In western Germany the survey has been conducted on a regular basis since 1993. In 1996 a representative sample of establishments in eastern Germany was surveyed for the first time.

4. The IAB-Establishment-Panel is a random sample from the IAB Establishment Database according to the principle of optimum stratification. The stratification cells are defined by ten classes for the size of the establishment and by 16 economic sectors. This selection process means that the selection probability of an establishment increases with its size.

5. Data on about 4000 establishments in western Germany and about 5000 in eastern Germany are collected each year. The density of coverage in eastern Germany is much higher than in western Germany, which means that it is also feasible to conduct studies at state level for eastern Germany.

6. Except for a few additional questions for eastern Germany, the questionnaires are identical for all firms. It includes questions on the development of employment, business policy and business development, the level of technology used. Altogether each wave of the IAB-Establishment-Panel covers more than 100 variables. Like other firm panels this one consists of a wide range of continuous and discrete variables like sales, sectors or pattern of employment.

### III. PROTECTION AND RELEASE OF INDIVIDUAL DATA SETS IN GERMANY

7. In Germany due to article 16 (6) of the Federal Statistic Law, there is a scientific privilege for the use of official statistics. It was one consequence of the population census judgement in 1979. For scientific purposes the absolute anonymization of microdata is given up for the concept of factual anonymity. "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'" (Knoche 1993).

8. The term factual anonymity is not further defined by law. To clarify the conditions under which the factual anonymity of released microdata can be attained, a co-operative project was carried out by the university of Mannheim, the Federal Statistic Office and the Centre for Surveys, Methods and Analyses (ZUMA) in Mannheim (Müller et al. 1991, Blien et al. 1993, Müller et al. 1995).

9. It was attempted to re-identify persons in the data set under realistic conditions with the help of so called scenarios. For this purpose it was necessary to inquire into the motives scientists could have to re-identify the data and the additional information available had to be determined. Rules for data release were established through different attempts of re-identification. These measures now serve as a basis for an improved practice on the part of the statistical bureaus in releasing microdata (Alba et al. 1994). So far the following factual anonymized scientific-use files are released by the Federal Statistic Office:

- microcensus for the years 1989, 1991, 1993 and 1995
- sample survey of income and expenditure 1993
- european household survey 1994 - 1996 (data for Germany).

For 1999 the release of the time use survey 1991/92 is planned. (Köhler 1999).

10. There are also other anonymized data sets for individuals like the IAB employment statistics sample. The statistical disclosure rules for that data set were derived from the

project described above. The anonymization rules were extended to longitudinal aspects (Bender et al. 1996). The different data sets mentioned earlier “constitute the backbone of the German social statistics” (Alba et al. 1994: 66). Although there are statistical disclosure rules for individual data sets and many samples may be used by the scientific community, many problems are still to be solved. So far the data sets are only released for well defined research projects, so that continuous research is hardly possible (Müller/Wirth 1996).

#### IV. THE ANONYMIZATION OF FIRM DATA: GENERAL REMARKS

11. In general the German Bundestag thought the anonymization of firm data not suitable when the concept of factual anonymity was first introduced in 1979. Nonetheless, the release of firm data sets is common practice in other countries like France, Great Britain, Canada and Finland (Köhler 1999).

12. In Germany, there are different legal regulations for different institutions. For the Institute for Employment Research (IAB) the law concerning the social security system is applied. The law concerning the social security system (SGB) makes no distinction between persons and enterprises respectively and establishments (SGB I, §35). Therefore it is necessary to ensure the factual anonymity of business data, too. Moreover many similarities between firm data and individual data can be found. Very often the information on the firm is identical to the information on the owner. This is always the case for sole ownerships regardless whether a joint-stock corporation or a one plant enterprise is concerned. It must also be kept in mind that information on firms is always provided by a natural person that might as well be identified by the information given. In summary, it is necessary and helpful to treat firm data equally to individual data.

13. The additional information available and the number of identical variables are of crucial importance to the re-identification of a firm, as well as the validity of this information. In comparison to the additional information available for individuals the additional information on firms can be used much easier and much more information is at hand. There are many databanks containing general information on enterprises like the following (cf. eg. GBI 1998):

- AZ Bertelsmannprofile deutscher Unternehmen
- Hoppenstedt Profile deutscher Unternehmen
- Creditreform Firmenprofile

14. These databases cover at least information on the firm's name and address, legal form, date of foundation, number of employees and sales. That shows that the number of variables in the additional information is quite high on firm level compared to individuals. Moreover, the information on firms is probably much more precise than for persons, because of the publication requirements for joint-stock companies and other large enterprises. The firms' representatives answer survey questions knowing that certain facts about the enterprise are publicly known. Therefore, the independence of the key variables cannot necessarily be assumed as in the case of individual data (cf. Müller/ Wirth 1996).

15. Usually firm surveys are stratified samples with respect to industry and firm size (e.g. IAB-Establishment-Panel). In these surveys the selection probability of the large firm sizes is much higher than for individuals. For instance, a household has a probability to be included in the microcensus of 1 %, whereas a firm with 1000 to 5000 employees has an inclusion probability of about 90 %. If an intruder is interested in an establishment that is included with a high probability, he faces nearly the same situation as an intruder who knows that the unit he searches for is in the survey. Experiments of Müller et al. (1989) show that the “participation knowledge is one of the most important preconditions” (Müller et al. 1995, Müller/Wirth 1996) for re-identification.

16. Looking at individual data, the benefits of a re-identification are rather low, therefore someone interested in re-identifying persons or households would only accept low costs. Comparing costs and benefits of a re-identification of a firm this could be totally different, because firm data sets usually contain information on future variables like investments or innovations planned. The information could be of enormous importance to competitors, so that the costs accepted for re-identification are much higher. For this reason it is most important that the scientific research interests do not equal the interests of (potential) competitors.

17. It could be argued that data security can be maintained through the fact that the responding unit is - as in the case of the IAB-Establishment-Panel - the local unit or workplace not the company, whereas additional information is available on enterprise or company level only. But this security factor is undermined because it is known whether the unit interviewed is a single plant enterprise or part of a company. For instance, more than 70 % of the firms in the IAB sample are independent single plant enterprises.

18. For reasons discussed above it seems plausible to assume that the risk of re-identification is much higher for business than for individual data sets. To quantify the proportions of endangered individuals several estimations are presented in the next section.

## V. PROPORTIONS OF ENDANGERED POPULATION UNITS

19. One of the key concepts used in statistical disclosure control for microdata is the uniqueness concept (e.g. Mokken et al. 1992). An individual, which 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. If it is taken into account that the intruder could have some information, which is not considered to be identifying, observations, which possess rarely occurring combinations of values of key variables are endangered, too (Willenborg/de Waal 1996).

20. In order to quantify the re-identification risk, we estimated the proportion of population uniques  $P(U_p)$  and the proportion of population individuals who have rarely occurring key-variable combinations  $P(S_p)$ . Therefore we define a key variable combination (key-value) as rare, when it occurs less than four times in the population. Additionally we estimated the corresponding probabilities conditioning on the sample uniques. These estimates can be used as an indicator for the increase of the re-identification risk, when the attacker restricts himself to the sample uniques (cf. Skinner et al. 1990).

21. The following key-variables were used:

sales (8 categories)

number of employees (establishment-size) (8 categories)

percentage of sales generated by export (6 or 21 categories)

business sector (3 categories)

legal form (3 categories)

firm-age (5 or 8 categories)

existence of a collective agreement (3 categories)

22. Except for the variable 'existence of a collective agreement' all of these key-variables are available in all databases discussed in section 4. But it seems to be plausible that this information can be easily achieved by an intruder. So this variable is an example for a possibly identifying variable, that is not in the stock-market or product-market orientated databases discussed above.

23. To categorize the variables, the following classifications were used: for the variables sales, export, sector, legal form it is assumed that the released data set contains only little information, because of the obvious re-identification potential. Therefore the establishment size was categorized into classes, which are equivalent to the strata. For the sales variable a coarse categorization was chosen, too: less than 1 Mio. DM, 1 - 5, 5 - 10, 10 - 20, 20 - 50, 50 - 100, 100 - 150, 150 - 250, more than 250 Mio. DM. For the sector it is assumed, that it is released on the on digit-level, so that there are three remaining industries: Mining/Energy and basic industry, capital goods, consumer goods industry. Furthermore the potentially highly identifying variable legal form was reduced to three main categories: partnership, limited company, other.

24. Moreover it was suggested that the export-variable, the firm age and the existence of a collective agreement are of scientific interest. Therefore different 'releasing options' for these variables should be tested by the experiments.

25. The export-rate was categorized in two ways. First 6 categories were constructed: no export, less than 20% of sales, among 20 and 40%, 40 - 60, 60 - 80, more than 80%. Second a more detailed releasing-option, in which the values of the variable are rounded to the nearest "5%-point" was modelled: 0%,5,10,...95,100%.

26. The firm age variable was surveyed in a two-step question. First all firms were questioned whether the firm was founded before 1960. Only younger establishments were asked in a second step since when they exist. For these establishments we modelled the key variable in two ways. For a first categorisation it was assumed that the intruder has only knowledge about the decade. The resulting key variable has 5 categories. Secondly it was assumed, that the intruder knows in which half of the decade the establishment was founded. This leads to a categorised key-variable with 8 categories. The last key-variable, existence of a collective agreement was assumed to be known in the original categories: yes, no, orientated at a related collective agreement.

27. The experiments are based on of the fourth wave of the IAB-Establishment-Panel 1996. The sample is restricted to west German establishments from mining/energy and the manufacturing sector, because for several services industries the variables sales and export-rate were not surveyed. Therefore the basic sample was reduced to less than 2000 firms. Additionally there is a high percentage of item non-responses to some variables. This is especially true for sales. More than 35 % of the firms in the sub-sample did not answer this question. Therefore we received a net sample with about 1100 observations.

28. Table 1 shows the proportions of sample uniques in the strata of establishments with less than 1000 employees. In all strata the proportion of sample uniques is very high. Except for the size class 5 - 9 employees the proportion of sample uniques increases, when it is assumed that firm age and export are known in more detail (Key 2). Especially in higher firm classes we observe significant differences. This finding is primarily founded in the higher skewness of the distributions of the key-variables in the higher firm classes.

Unfortunately the amount of the differences in the proportions between the firm-classes cannot be interpreted due to the big differences in the net sample sizes (n)<sup>1</sup>.

**Table 1: Proportion of sample uniques (in %)**

| <i>firm-size</i> | 1 - 4 | 5 - 9 | 10 - 19 | 20 - 49 | 50 - 99 | 100 - 199 | 200 - 499 | 500 - 999 |
|------------------|-------|-------|---------|---------|---------|-----------|-----------|-----------|
| Key 1            | 57    | 56    | 66      | 62      | 77      | 65        | 50        | 44        |
| Key 2            | 61    | 55    | 68      | 65      | 79      | 84        | 73        | 61        |
| n                | 46    | 80    | 73      | 125     | 99      | 129       | 209       | 135       |

Source: IAB-Establishment-Panel 1996

29. The estimations of the population proportions were performed using the assumption that in each stratum the frequencies of the key-values in the population and in the sample follow approximately a logarithmic distribution<sup>2</sup>. The logarithmic distribution is a limiting form of the negative binomial distribution, which is usually used to model the distribution of the frequencies of the key-values (Skinner/Holmes 1993)<sup>3</sup>.

30. Table 2 shows that the estimated percentage of population uniques  $P(U_p)$  and the proportion of establishments with rare key-values, that is key values that occur less than 4 times in the population, increase with the establishment size. For establishments with 50 or more employees the estimated proportions are relatively high. Furthermore, the results indicate a strong increase in the proportion of population uniques with increasing information on establishments with more than 100 employees. The same tendencies are shown by  $P(S_p)$  while the absolute values of the estimated proportions are higher. Very high values for  $P(S_p)$  occur in the establishment classes with 50 or more employees.

31. If the intruder restricts himself to the sample uniques, the estimated proportions of endangered observations ( $P(U_p|U_s)$ ,  $P(S_p|U_s)$ ) are necessarily higher than in the unconditioned case. In general this increase is stronger, if the information level is relatively low (Key 1). This kind of search seems reasonable for an intruder, who has a low level of additional information.

32. Additionally experiments were carried out, in which the variable "existence of a collective agreement" was not included. For large and medium sized establishments the results indicate only a small reduction in the estimated proportions while for establishments with less than 100 employees a clear reduction can be observed.

<sup>1</sup> For the impact of the sample size on the number of sample uniques see e.g. Elliot/Skinner/Dale 1998.

<sup>2</sup> The theoretical deviations can be found in the appendix.

<sup>3</sup> The basic distributional assumption is that the distribution of the frequencies of the key values can be adequately described by a poisson mixture. Several poisson-mixtures were tested. The results indicate that for moderate key-sizes, the logarithmic distribution is a simple and well performing distribution model (Indicators  $C^2$ -test, graphical tests analogous to Hoaglin 1986)

**Table 2: Parameter estimates and estimated proportions (in %)**

| Key 1           | 1 - 4  | 5 - 9  | 10 - 19 | 20 - 49 | 50 - 99 | 100 - 199 | 200 - 499 | 500 - 999 |
|-----------------|--------|--------|---------|---------|---------|-----------|-----------|-----------|
| $\Theta_s$      | 0.4077 | 0.4746 | 0.3583  | 0.4203  | 0.2403  | 0.3640    | 0.5236    | 0.5783    |
| $P(U_s)$        | 59.23  | 52.54  | 64.17   | 57.97   | 75.94   | 63.60     | 47.64     | 42.17     |
| $P(U_p)$        | 0.23   | 0.29   | 0.41    | 0.65    | 2.87    | 3.48      | 4.36      | 6.84      |
| $P(S_p)$        | 0.68   | 0.86   | 1.22    | 1.95    | 8.37    | 10.07     | 12.53     | 19.14     |
| $P(U_p U_s)$    | 0.40   | 0.51   | 0.62    | 1.06    | 3.74    | 5.34      | 8.77      | 15.39     |
| $P(S_p U_s)$    | 2.40   | 3.04   | 3.71    | 6.27    | 21.33   | 29.74     | 46.41     | 73.57     |
| Key 2           |        |        |         |         |         |           |           |           |
| $\Theta_s$      | 0.3761 | 0.4746 | 0.3375  | 0.4090  | 0.2064  | 0.1611    | 0.2765    | 0.3940    |
| $P(U_s)$        | 62.39  | 52.54  | 66.25   | 59.10   | 79.36   | 83.89     | 72.35     | 60.60     |
| $P(U_p) P(U_s)$ | 0.23   | 0.29   | 0.43    | 0.68    | 3.48    | 9.69      | 11.60     | 13.41     |
| $P(S_p)$        | 0.69   | 0.86   | 1.30    | 2.03    | 10.08   | 26.35     | 30.92     | 35.07     |
| $P(U_p U_s)$    | 0.38   | 0.52   | 0.63    | 1.05    | 4.41    | 11.47     | 15.95     | 21.81     |
| $P(S_p U_s)$    | 2.27   | 3.11   | 3.77    | 6.20    | 24.96   | 56.68     | 76.48     | 95.45     |

Source: IAB-Establishment-Panel 1996

33. Because of the relatively small sample sizes it could be suggested that the estimates are not very reliable. Analogous estimates were performed for the first wave of the Hanover Panel. The Hanover Panel is a similar survey conducted for establishments of the manufacturing sector in Lower Saxony (Brand et al. 1996). The results are similar to the estimated parameters and the sample proportions. The estimated proportions for the populations are much higher because of the restriction to a particular region.

34. Summarizing, the estimated proportions of endangered observations are high even if the assumed information level is relatively low. In reality the information level of the intruder will be higher because the intruder will probably possess more and very detailed additional information. Therefore, the results can be regarded as lower bounds for the true proportions. On the other hand the underlying assumption that all key variables are known on the establishment level is restrictive, because only the variables "existence of a collective agreement" and "legal form" are necessarily identical on the establishment and the enterprise level. For all other key-variables this identity is fulfilled only, if the establishment is a one plant enterprise. For establishments that are subsidiaries of a company or are parent firms this identity is not necessarily valid.

## VI. METHODS OF DISCLOSURE CONTROL

35. However, these results give a strong indication, that the usually used statistical disclosure avoidance techniques are not sufficient. Therefore, it is necessary to use a data modifying technique. The aim is to conduct a data set that is useful for empirical economic and social research. Methods regularly applied in this field of research should be applicable. This is only possible if the means and the variance covariance matrix of the scientific-use file and the original data differ by random only. In order to analyse the variables the univariate distribution of the sample and the original data set should be approximately the same.

36. In the literature, the generation of a scientific-use file for business data is accomplished by means of micro aggregation methods (e.g. Corsini et al. 1998, Mateo-Sanz/Domingo-Ferrer 1998). However most of the micro aggregation methods change the variances and the correlation structure. This problem seems to be of special importance to data sets with small correlations, which is a characteristic of the IAB-Establishment-Panel, too.

37. Another possible approach is the masking by means of transformation and addition of random noise. Two important methods of this group are those from Sullivan (1989) and Kim (1986), (cf. Kim/Winkler 1995). These methods are suited for the generation of a scientific-use file, because the empirical moments are approximated and the expected values and the variance-covariance-matrix of the original data can be estimated consistently by the masked data. A special advantage of the method proposed by Sullivan (1989) is that for both discrete and metric variables the univariate distributions are sustained.

38. The masking procedure conducted by Sullivan (1989) consists of several steps<sup>4</sup>. At first the observed variables are transformed into nearly normal distributed variables. Therefore, the sample cumulated distribution function is constructed and these values are transformed to normal variables. Second normal error is added to the transformed variables. To avoid distortions of the means and the covariance matrix the error has to have zero mean and a variance-covariance-matrix that equals the variance-covariance-matrix of the transformed data set. Third the masked values are transformed back to the original scale.

39. With this approach, some studies were conducted for the different strata of the whole data set. In all experiments the differences in means and variances are very small and the univariate distributions of the masked variables exhibit only minor differences. But unfortunately we observed clear differences in the covariance for several variables in many experiments. The main reason for this is that for several variables the transformations to normality were not successful, because of the original distributions of the variables cannot be transformed into normality by usual transformations. This is due to the fact that several variables have extreme distributions, this means that the modus lies at an extreme, or follow a distribution with a very high skewness. Therefore experiments with a reduced number of variables were performed (Bellmann et al. 1998).

40. The results of these experiments are promising. The differences in the covariance-matrix are in general relatively small. Therefore only minor differences in the results of the tested regression analysis<sup>5</sup> occurred. Tables 3 and 4 show the results for a regression conducted for the size class 50 - 99 employees. The masking was performed with an amount of noise that is much higher than the amount of noise tested by Sullivan (1989) (relative amount of noise: 2). Because of this the algorithm conducted by Sullivan was modified. In Sullivan's original program (Sullivan 1989) a subroutine is implemented which equals the correlations between the original and the masked data. This subroutine does not work, if the amount of noise is bigger than one. Therefore this 'correction' of the correlations was not used.

---

<sup>4</sup> For a more detailed description see Fuller (1993).

<sup>5</sup> The model chosen for the regression analysis is an estimation for the wages in Western Germany for the mining/energy and manufacturing sectors. The dependent variable is the firm's wage per employee in June 1996 ( $\ln(\text{wage})$ ). Different determinants like the proportion of women, number of employees, overtime work, legal form, technical equipment and work on weekends were included as independent variables to determine their influence on the wages.

41. The values of the estimated coefficients and the estimated standard errors differ only for the dummy variables constructed from the legal form clearly. For all other variables the estimates are reproduced very well. The same is true for the goodness of fit measures. Therefore the general conclusions are not seriously affected.

**Table 3: Regression results original data**

|                |         |                     |       |
|----------------|---------|---------------------|-------|
| Valid cases:   | 691     | Dependent variable: | Y     |
| Missing cases: | 0       | Deletion method:    | None  |
| Total SS:      | 108.620 | Degrees of freedom: | 682   |
| R-squared:     | 0.200   | Rbar-squared:       | 0.190 |
| Residual SS:   | 86.909  | Std error of est:   | 0.357 |
| F(8,682):      | 21.296  | Probability of F:   | 0.000 |

  

| Variable          | Estimate  | Standard Error | t-value   | Prob > t | Standardized Estimate | Cor with Dep Var |
|-------------------|-----------|----------------|-----------|----------|-----------------------|------------------|
| CONSTANT          | 7.154450  | 0.281178       | 25.444553 | 0.000    | ---                   | ---              |
| Prop. of woman    | -0.225207 | 0.049852       | -4.517549 | 0.000    | -0.171058             | -0.166280        |
| ln(employees)     | 0.184617  | 0.065845       | 2.803792  | 0.005    | 0.096346              | 0.091921         |
| dummy - variables |           |                |           |          |                       |                  |
| western Germany   | 0.290406  | 0.028262       | 10.275424 | 0.000    | 0.361818              | 0.345165         |
| overtime work     | 0.079769  | 0.029924       | 2.665685  | 0.008    | 0.098541              | 0.106231         |
| limited company   | 0.155248  | 0.042834       | 3.624441  | 0.000    | 0.184115              | 0.075382         |
| other legal form  | 0.244770  | 0.051272       | 4.773942  | 0.000    | 0.250106              | 0.004355         |
| satisfactory      |           |                |           |          |                       |                  |
| technical level   | -0.008728 | 0.029325       | -0.297634 | 0.766    | -0.010351             | -0.018537        |
| work on weekend   | -0.075510 | 0.030084       | -2.509985 | 0.012    | -0.090940             | -0.132638        |

Source: Bellmann et al. 1998

**Table 4: Regression results masked data (masking procedure based on previously transformed data)**

|                |         |                     |       |
|----------------|---------|---------------------|-------|
| Valid cases:   | 691     | Dependent variable: | Y     |
| Missing cases: | 0       | Deletion method:    | None  |
| Total SS:      | 109.903 | Degrees of freedom: | 682   |
| R-squared:     | 0.155   | Rbar-squared:       | 0.145 |
| Residual SS:   | 92.870  | Std error of est:   | 0.369 |
| F(8,682):      | 15.635  | Probability of F:   | 0.000 |

  

| Variable         | Estimate  | Standard Error | t-value   | Prob > t | Standardized Estimate | Cor with Dep Var |
|------------------|-----------|----------------|-----------|----------|-----------------------|------------------|
| CONSTANT         | 7.211193  | 0.293703       | 24.552693 | 0.000    | ---                   | ---              |
| prop. of woman   | -0.138962 | 0.050673       | -2.742342 | 0.006    | -0.109181             | -0.114952        |
| ln(employees)    | 0.187154  | 0.068161       | 2.745755  | 0.006    | 0.097160              | 0.069459         |
| dummy-variables  |           |                |           |          |                       |                  |
| western Germany  | 0.274963  | 0.028997       | 9.482385  | 0.000    | 0.340571              | 0.332141         |
| overtime work    | 0.080277  | 0.030492       | 2.632738  | 0.009    | 0.098589              | 0.097860         |
| limited company. | 0.074937  | 0.042294       | 1.771807  | 0.077    | 0.088351              | 0.060530         |
| other legal form | 0.087294  | 0.049268       | 1.771827  | 0.077    | 0.088675              | -0.037660        |
| satisfactory     |           |                |           |          |                       |                  |
| technical level  | -0.032029 | 0.030346       | -1.055459 | 0.292    | -0.037762             | -0.054602        |
| work on weekend  | -0.073766 | 0.030538       | -2.415549 | 0.016    | -0.088319             | -0.100145        |

Source: Bellmann et al. 1998

42. In order to test whether the masking is effective a distance-criterion is used. For 22 establishments, the distance between the masked record and the original data is smaller than all other distances between a masked and the interesting record in the original file. Considering this criterion, only 3% of the records are not sufficiently masked. Additionally matching probabilities are estimated, based on the simplifying assumption that all data are normally distributed (cf. Sullivan 1989). The results show that only very few records are at a high risk of re-identification. Except for those records who are not sufficiently masked, the estimated matching probabilities are very low. This is reasoned in the high and not systematic differences between the original and the masked values.

43. Similar results are achieved for all other masked data sets, based on small- or medium-sized firms. The proportion of observations that cannot be sufficiently masked is always very low. Thus the approaches tested in our study can be regarded as suitable. Tests based on the strata, which incorporate large firms (more than 500 employees) show the problems corresponding to this approach. Because of the large differences between the observed values a sufficient mask can only be achieved if the added error is very high. This leads to serious distortions of the covariance matrix, when the necessary transformations do not perform very well.

## **VII. CONCLUSIONS**

44. In Germany, the legal requirements of data security/ protection are very strict, so that the disclosure of the original microdata is not possible. For several reasons it must be assumed that business and individual data must be treated differently regarding data release or disclosure. In comparison to individual data, the additional information available on firm level is comprehensive and precise and the benefits of a re-identification of firm data are much higher, especially for competitors. Moreover, the probability to be included in a sample is very high for many establishments and thereby the risk to be re-identified. Proportions of endangered units are estimated on the basis of the uniqueness concept. All in all, the results show that not only large- and medium-sized establishments are at a high risk of re-identification.

45. Therefore it is necessary to use data modifying disclosure limitation techniques. A helpful scientific-use file has to fulfil the requirement that methods regularly applied in this field of research should be applicable. This is only possible if the means and the variance-covariance-matrix of the scientific-use file and the original data differ by random only. To analyse the variables, the univariate distribution of the sample and the original data set should be approximately the same. The approach of Sullivan (1989) seems to be suitable for this purpose.

46. The results conducted by an experimental mask are promising. Regression estimates performed with the masked data set show no significant differences to the estimates based on the original data. First estimations of the effectiveness of this approach show a clear reduction of endangered observations.

47. However practical problems and unanswered questions have to be solved by further investigations. For instance, the application runs into problems if the transformation of original variables into normally distributed variables does not succeed. The results of Winkler (1998) show that the re-identification risk increases significantly if record linkage algorithms are used. Therefore, it seems necessary to perform experiments with such algorithms and a complete test of the effectiveness of a masking algorithm is still to be conducted.

## References:

- Alba, R., Müller, W. and B. Schimpl-Neimanns (1994): Secondary Analysis of Official Microdata, in: Borg, I., Mohler, P. (Eds.): Trends and Perspectives in Empirical Social Research, Berlin: de Gruyter.
- Bellmann, L. (1997): Das Betriebspanel des IAB, in: Hujer, R. (ed.): Wirtschafts und sozialwissenschaftliche Panelstudien, Sonderhefte zum Allgemeinen Statistischen Archiv, Heft 30, Göttingen: Vandenhoeck und Ruprecht.
- Bellmann, L., Bender, S., Bielski, H., Brand, R., Hinz, T., Kohaut, S. and E. Wolf (1998): Die Anonymisierung des IAB-Betriebspanels, Rechtslage und praktische Verfahrensvorschläge, Projektbericht, Institute of Employment, Nürnberg.
- Bender, S., Hilzendegen, J., Rohwer, G. and H. Rudolph (1996): Die IAB-Beschäftigtenstichprobe 1975-1990. Eine praktische Einführung. Beiträge zur Arbeitsmarkt- und Berufsforschung 197, Nürnberg.
- Blien, U., Müller, W. and H. Wirth (1993): Identification Risk for Microdata stemming from Official Statistics, *Statistica Neerlandica*, 46, 1, 69-82.
- Böhning, D. (1983): Maximum likelihood estimation of the logarithmic series distribution, *Statistische Hefte* 24, S.121--140.
- Brand, R., Carstensen, V., Gerlach, K. and T. Klodt (1996): The Hanover Paenl, Das Hannoveraner Firmenpanel -- Diskussionspapierer -- discussion paper No. 2, Hannover 1996.
- Corsini, V., Franconi, L., Pagliuca, D. and G. Seri (1998): An Application of Microaggregation to Italian Business Surveys, Paper presented at the Conference of Statistical Data Protection '98, Lisbon, Portugal.
- Elliot, M.J., Skinner, C.J. and A. Dale (1998): Special Uniques, Random Uniques and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on disclosure Risk., Paper presented at the Conference of Statistical Data Protection '98, Lisbon, Portugal.
- Fuller, W.A. (1993): Masking Procedures Microdata Disclosure Limitation, *Journal of Official Statistics*, Vol. 9, 383 - 406.
- Hoaglin, D.C. (1980): A Poisson Plot, *The American Statistician*, 34, 146 - 149.
- Kim, J.J. (1986): A Method for Limiting Disclosure in Microdata based on Random Noise and Transformation, in: American Statistical Association: Proceedings of the Section on Survey Research Methods, 303 - 308.
- Kim, J.J. and W.E. Winkler (1995): Masking Microdata Files, American Statistical Association: Proceedings of the Section on Survey Research Methods, 114 - 119.

- Knoche, P. (1993): Factual Anonymity of Microdata from Household and Person-related Surveys - The release of Microdata Files for Scientific Purposes, in: Proceedings of the International Symposium on Statistical Confidentiality, Dublin: Eurostat, 407 - 413.
- Köhler, S. (1999): Anonymisierung von Mikrodaten in der Bundesstatistik und ihre Nutzung - Ein Überblick -, in: Statistisches Bundesamt (ed.): Methoden zur Sicherung der statistischen Geheimhaltung, Band 31 der Schriftenreihe Forum der Bundesstatistik, forthcoming.
- Mateo-Sanz, J.M. and J. Domingo-Ferrer (1998): A Method for Data-Oriented Multivariate Microaggregation, Paper presented at the Conference of Statistical Data Protection '98, Lisbon, Portugal.
- Mokken, R.J., Kooiman, P., Pannekoek, J. and L. Willenborg (1992): Disclosure Risks for Mikrodata, *Statistica Neerlandica*, 46, 1, 49-67.
- Müller, W., Blien, U., Knoche P., Wirth, H., u.a. (1991): Die Faktische Anonymität von Mikrodaten, in: Statistisches Bundesamt (ed.): Schriftenreihe Forum der Bundesstatistik, Band 19, Stuttgart: Metzler-Poeschel.
- Müller, W., Blien, U., and H. Wirth (1995): Identification Risks of Microdata. Evidence from experimental studies, *Sociological Methods & Research*, 24, 2, 131-157.
- Müller, W. and H. Wirth (1996): Confidentiality and Disclosure of Microdata Sets obtained from Statistical Surveys, Paper presented at the International Symposium 'Exploring a New Frontier of Statistical Data Analysis with Micro Data Sets', October, 14 - 15, 1996, Fukuoka, Japan.
- Skinner, C.J., C. Marsh, S. Openshaw und C. Wymer (1990): Disclosure Avoidance for Census Microdata in Great Britain, U.S. Department of Commerce, Bureau of the Census, Proceedings of the Annual Research Conference, Washington DC.
- Skinner, C.J. and D.J. Holmes (1993), Modelling Population Uniques, in: Proceedings of the International Symposium on Statistical Confidentiality, Dublin: Eurostat, 175-199.
- Sullivan, G.R. (1989): The Use of Added Error to Avoid Disclosure in Microdata Releases, unpublished PhD-Thesis, Iowa State University.
- Willenborg, L. and T. de Waal (1996): Statistical Disclosure Control in Practice, in: Bickel et al. (ed.): *Lecture Notes in Statistics* 111, New York: Springer.
- Winkler, W.E: (1998): Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata, Paper presented at the Conference of Statistical Data Protection '98, Lisbon, Portugal.

## Appendix:

Let  $F_i$  be the number of population units with key value  $i$ ,  $i = 1, 2, \dots, C$ , and  $f_i$  the number of observations in the sample with key value  $i$ ,  $i = 1, \dots, C^+$ ,  $C^+ \leq C$ . A usually used model for the distribution of the  $F_i$  and the  $f_i$  is the Poisson-gamma model. If it is assumed that the parameters of the Poisson distributions are proportional the (known) inclusion probabilities  $p$ ,  $p = \text{const.}$ , the marginal distribution of the mixture is a negative binomial distribution (e.g. Skinner/Holmes 1993):

$$P(f_i = k) = P_k(\Theta) = \binom{a+k-1}{k} \Theta^k (1-\Theta)^a, \quad k = 0, 1, 2, \dots,$$

with parameter  $\Theta = \frac{pb}{1+pb}$ .

The probability generating function for the  $F_i$  is analogously achieved with  $p = 1$ .

It is well known that a limiting model of the zero truncated negative binomial distribution is the logarithmic distribution. The probability generating function of the logarithmic distribution with parameter  $\Theta$ ,  $0 < \Theta < 1$ , is defined as:

$$P(f_i = k) = P_k(\Theta) = \frac{1}{-\ln(1-\Theta)} \cdot \frac{\Theta^k}{k}, \quad k = 1, 2, \dots$$

The parameters can be estimated by maximum likelihood. These estimates are consistent and asymptotically efficient (Böhning 1983).

To estimate the proportions of individuals who are in size classes with  $k$  units the following approximation can be used (cf. Skinner/Holmes 1993):

$$P(k) \approx \frac{kP_k(\Theta)}{\sum_{k=1}^{\infty} kP_k(\Theta)} = \Theta^{k-1}(1-\Theta)$$

Therefore the proportion of sample uniques ( $P(U_s)$ ), the proportion of population uniques ( $P(U_p)$ ) and the proportion of population units that possess a key-value occurring less than  $d$  times in the population ( $P(S_p)$ ) can be estimated by:

$$\hat{P}(U_p) = 1 - \hat{\Theta}_p, \quad \hat{P}(U_s) = 1 - \hat{\Theta}_s, \quad \hat{P}(S_p) = P(k_p < d) = \sum_{k=1}^{d-1} \hat{\Theta}_p^{k-1} (1 - \hat{\Theta}_p),$$

where  $\hat{\Theta}_s = \frac{p\hat{b}}{1+p\hat{b}}$ ,  $\hat{\Theta}_p = \frac{\hat{b}}{1+\hat{b}}$ .

As corresponding conditional probabilities we obtain:

$$\hat{P}(U_p|U_s) = \frac{n \cdot \hat{P}(U_p)}{n(U_s)} \quad ,$$

$$\hat{P}(S_p|U_s) = \hat{P}(k_p < d|U_s) = \frac{N}{n(U_s)} \cdot \sum_{k=1}^{d-1} \hat{\Theta}_p^{k-1} (1 - \hat{\Theta}_p) \cdot P(U_s|k_p) \quad ,$$

where  $n$  stands for the sample size,  $N$  for the size of the population and  $n(U_s)$  for the number of sample uniques.  $P(U_s|k_p)$  describes the conditional probability of a sample unique stemming from a cell of size  $k$  in the population. If it is assumed that the sample is drawn by random sampling without replacement this probability is defined as:

$$P(U_s|k_p) = k \cdot \frac{\binom{N-k}{n-1}}{\binom{N}{n}} \quad .$$