Schmollers Jahrbuch 124 (2004), 315 – 325
Duncker & Humblot, Berlin

The Employment Panel of the German Federal Employment Agency

By Iris Koch and Holger Meinken

1. Introduction

In the last years the demand for detailed information about the German labour market has increased constantly, and refers no longer only to regularly produced statistical material or customized statistical summaries provided by the *Federal Employment Agency (FEA)*[1]. Instead, external labour market researchers prefer an easier access to mostly confidential micro-level data to obtain the possibility of more differentiated data analyses for their own. Therefore, the statistics department of the FES headquarter has established a research project with the intention to prepare and supply anonymized microdata of employment covered by social insurance. The project is located in the emerging discussions between social sciences and official statistics about efforts to improve the informational infrastructure in Germany.These were initiated by the *Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI).* The major aim of this discussion was to develop new solutions and to improve the access to microdata of the statistics offices and the other public data producers in Germany (KVI, 2001).

The microdata of the employment statistics are subject to the social data protection as part of the social legislation. Thus, the original data may not be analyzed by the researchers of external scientific institutions. Only anonymized microdata are allowed to be conveyed. That meant that in the beginning appropriate procedures of anonymization had to be developed and examined, before the microdata could be made available. Meanwhile the work advanced so far that a passing on of the anonymized dataset can be realized by the *Central Archive of Empirical Social Research (University of Cologne)*[2].

The dataset of employment covered by social insurance can be ordered from the Central Archive under the study No. 3887 for purposes of labour market research, employment research or occupational research. The data are developed as a panel dataset. Samples of the quarterly data of the employment

1 The German Federal Employment Agency is called *Bundesagentur für Arbeit (BA).*

2 In German it is called *Zentralarchiv für Empirische Sozialforschung (ZA).*

statistics of the years 1998 – 2002 form the basis of this panel. The single specifications of the employments were anonymized and set up to a *scientific use file*. This dataset includes all substantial compulsorily notifiable individual variables liable to social insurance as well as some additional variables referring to the respective employment establishments. Beginning with the first quarter of 1998, the panel contains at present 18 waves, each containing information of about 600,000 employees (including the »marginal« part-time employees). The FEA's employment panel will be updated annually.

2. Data Sources

The most important data source of the panel are the quarterly individual datasets of the employment statistics of the FES. The quarterly data result from data records supplied by the notification procedure for the social insurance system. In the notification procedure data records of all employments liable to social insurance in Germany are registered, regardless of the duration of an employment spell, e.g. only one day or several years. The notification procedure exists in its fundamentals since 1973 and was changed substantially in 1999 for the last time (Neidert, 1998).

The notification of each individual employment is issued by the employer and generates a single data record. As far as for the employees, the notifying facts are available, the establishments or employers have to deliver their notification of employment covered by social security within certain periods. For example, any newly established employment cases must be notified by the employer to the social insurance agencies. If an employment exists continuously with the same employer, the employer has to issue an annual confirmation at expiration of each calendar year, specified with the indication of the annual gross wage or salary liable to social insurance. If an employment spell ends, the employer has to submit a notice of departure with data of the legal ending of the employment and the gross wage obtained up to then in the current fiscal year. Thus a duration of employment with the gross wage obtained therein is – as a rule – completely verified by annual notifications and the end of the employment can be derived exactly from the notice of departure. The data records are collected by the health insurance agencies, proceeded to the federal or regional pension funds for administering the individual pension accounts and finally sent to the FEA. The FEA uses the records solely for statistical purposes. E.g., these data are evaluated to determine the stock figure of employments at particular quarterly deadlines (31. 3., 30. 6., 30. 9., 31. 12.).

In order to guarantee a high data quality the social insurance agencies use a mandatory test program which validates the data records with receipt of the dates of notification. The data validations are limited essentially to the legally crucial data, e.g. to the insurance number, to the address specifications, to the

indicated period of employment and to the payment liable to social insurance. The socio-demographic data contained in excess of that are only submitted to a brief formal examination.

Microdata proceeded from the employment statistics of the FEA can have substantial advantages over data taken from polls or surveys. Due to the standardized, mandatory notification procedure the administrative data of social insurance do not exhibit the usual shortcomings of poll data. The reliability of the data is very good in the most important dimensions, because each employee receives a copy of each notification that has been delivered from the respective employer. This works as an internal control mechanism in the notification process.

In particular the payment and the period of employment are not afflicted with the typical lack of the poll data to this field of topics such as memory gaps, response errors, item-nonresponse or consciously wrong data. The obligation to report each employment provides a reliable data flow, so problems like unit-nonresponse and sample loss are negligible, too. Especially in the longitudinal perspective the data possess the advantage that no observation unit fails because of temporary absence, change of residence or unit-nonresponse in general. The losses of the available data are only limited to employees who separated temporarily or finally from an employment. On the other hand new persons advance, i.e. those, that take up an employment liable to social insurance for the first time or again after a break.

Due to the mandatory notification procedure for employments covered by social security, this data base covers approx. 75 – 80 % of all employment cases in Germany. Excluded are only employees not covered by social security such as freelancers, one-person-businesses, civil servants, and family workers. Thus the employment data of the FEA represent a unique base for investigating labour market and occupation. At the same time it is possible to make a linkage between these compulsory notification data and the microdata of other administrative areas of the FEA.

The most important advantage of the data resulting from the notification procedure of employment liable to social insurance may be seen in the fact that it yields a total collection of data. That is why the samples of the employment data, e.g. the FEA's employment panel, can be verified in their similarity with the population, which in this case consists of all employees liable to social insurance. Due to the knowledge of the frequency distribution of variables in the population, it can easily be examined for any samples whether their distributions deviate significantly from those of the population. As a disadvantage of the employment data we have to mention the small extent of variables and the small degree of differentiation within single variables. This can be regarded as a specific characteristic of secondary statistical data, which are developed on the basis of very large populations.

It should be noted that there is a delay of the data input of the notification procedure with negative effects on current evaluations. Most current evaluations are based on a waiting period of six months related to the respective quarterly deadline. Although there are only some 90 % of the employment cases included within this time span, this figure supplies quite reliable results using some approximative calculations. However, a small inaccuracy remains. If conversions of the notification procedure happen or additional verification of data is necessary, then longer waiting periods have to be chosen.

Starting in 1999, some remarkable innovations in the notification procedure have taken place. That is the introduction of the obligation to register »marginal« part-time employments, the different reporting of partial or early retirement, the reorganization of the »dual system« of vocational education and training, and the occupations connected with the introduction of the nursing care insurance. In the notification procedure these changes refer to the extension of the circumstances triggering a notification and the change of the variables which have to be reported.

The data records of the employment notifications build the base to compute the stock of employees liable to social insurance and of the »marginal« part-time employees at the four quarterly deadlines of a year. In order to get this information, the latest notification data record indicating or confirming current employment have to be drawn for each employee by an extensive inquiry pattern. It has to be checked whether the employee in question was occupied liable to social insurance at the quarterly deadline. In other words, the actual data records of notification thus generate a cross-sectional information. The resulting dataset is called quarterly file. These quarterly files are the starting points of the samples of the employment panel. To extend the information content of the panel some variables of the respective employment establishments are added. By means of its identification number the branch of industrial activity and the regional classification of the establishments can be identified for each employment case from the FEA's establishment file. Additionally, the establishment sizes and different ratios of employees for each quarterly deadline are computed from the stocks of the quarterly files.

3. Data Description

3.1 Sampling Procedure

The employment panel is drawn as a random sample based on a birthday selection. The social insurance number of each employee that is used for all employment notifications as a biunique identification contains the date of birth of the respective employee. With first notification of an employment liable to social insurance an employee is given his social insurance number

which remains unchanged in the further working life. For the sample only those employees are selected who have the same day of birth as one of seven predefined days of the year. Since the predefined days are fixed, and the social security numbers of the employees do not change, it is quite simple to select the right employees at each quarterly deadline respective of each quarterly file. In this manner we obtain a sample that contains 1,92 % of the employees of the population.

If there are employments with no substantial changes over time, the particular employee is selected by the date of birth for each passing quarter. Employees who change jobs vary with regard to the kind of employment and the new establishment. However, they remain in the sample durably, because they can be identified at each quarterly deadline due to their unique social insurance number. If a person has no employment liable to social insurance at a particular quarterly deadline, no data record of this person is included in that wave. Only if a new employment spell at a following quarterly deadline occurs, the employee is included again in the dataset with the information about this employment.

The applied sample procedure offers a number of advantages for the sample composition. At first we can point out the similarity in the structure between sample and population, both in the cross-sectional and in the longitudinal perspective. The loss of units caused by panel mortality becomes automatically balanced. Older employees who retire from active work are replaced by young persons who take up their first employment and exhibit one of the selected birthdates. Frequency counts by birth cohorts are always proportional to those in the population. Thus the sample provides excellent possibilities of executing cohort analysis. In addition, the sample procedure ensures the illustration of existing seasonal or cyclical effects, because always a constant ratio of all employees is selected. Contrary to the German microcensus with its area sample, the sample of the employment panel has the advantage that e.g. the regional mobility of employees can be analyzed and does not lead to sample losses.

Since the samples of the single waves are proportional to the population at the respective quarterly deadline, their sizes vary with the employment figures from quarter to quarter. Table 1 gives an overview of the sample size and population for the last wave. Given an entire number of about 32 million employees, the samples include approx. 600,000 units. For the first five waves the panel reaches only a sample size of approx. 500,000 employees, because at these times the »marginal« part-time employees were not yet contained in the quarterly file (cf. Bundesanstalt für Arbeit, 2003).

In order to examine the quality of the sample, the univariate frequency distributions of all variables in the dataset were compared with those of the population. As a result the sample provides a very good representation of the population.

Table 1

Sample Size of the Employment Panel

Quarter data 30. 6. 2002	West Germany		East Germany*		Germany	
	Sample	Population	Sample	Population	Sample	Population
Employees covered by social insurance	426,055	22,182,502	104,065	5,388,645	530,120	27,571,147
»Marginal« part-time employees	68,942	3,599,798	10,768	569,368	79,710	4,169,166
Total	494,997	25,782,300	114,833	5,958,031	609,830	31,740,313

* incl. Berlin.

3.2 Variables

The dataset covers 52 variables. Some of them can be considered as individual variables of the employees, and the remaining are variables of the respective establishments (Table 2).

Table 2

Variable List of the Employment Panel

Employee Characteristics	Establishment Characteristics
• Identification number • Date of the quarterly deadline • Number of the panel wave • Reason of notification • Sex • Age in years • Nationality (23 categories) • Educational level (7 categories) • Occupational status / Working hours (9 categories) • Pension insurance agency (blue-collar / white-collar worker) • Remuneration liable to social insurance • Group of employment (10 categories) • Group of contributions (6 categories) • Occupation (some 300 categories) • Job-change (5 categories)	• Size of Establishment (8 categories) • Economic branch (48 categories) • Region (East / West Germany) • Ratio of women • Ratio of employees in respective age • groups (11 groups) • Ratio of German nationality • Ratio of part-time employees • Ratio of employees respective • educational levels (6 levels) • Ratio of trainees • Ratio of blue-collar workers • Ratio of skilled workers • Ratio of white-collar workers • Ratio of pension insurance (blue-collar / white-collar)

3.3 Anonymization Procedure

For more than a decade, several research projects have been carried out to improve the access to official microdata by creating disclosure control and anonymizing data. As a consequence the microdata could be made easier to use in the form of *scientific use files*. For the first time anonymization procedures were developed for the German microcensuses (Müller et al., 1991), and then converted in a pilot project (Köhler et al., 2000). After that, further anonymized datasets were developed which essentially based on the same or similar anonymization procedures.

In general the procedures for the anonymization of microdata used so far in the official statistics, can be characterized as follows (Köhler, 1999):

- Modifying the data (adding random errors to values, producing artificial data records),
- Collapsing response categories (aggregating categories with small frequencies or extreme values),
- Suppressing units or variables (sample drawing, removing of identifiers),
- System-free sorting of the data.

The employment panel is built without any modifying of the original data, because the implications of analyzing these modified data are not quite clear up to now. Without transparency about the specific anonymization procedures and its consequences for estimating statistical parameters, this kind of anonymization must be assessed rather critically (Lechner / Pohlmeier, 2003). The anonymization steps used for the panel follow those procedures which were already developed for the microcensuses. That means in detail:

(1) Drawing a random sample of the population with a selection probability of 1,92 %,

(2) Replacing the social insurance number by a random-generated, system-free identification number. Suppress variables of personal identification and particularly sensitive variables, e.g. date of birth or establishment number,

(3) System-free sorting by the new generated identification number,

(4) Collapsing of response categories:
   - Reducing regional information to West and East Germany.
   - Nationalities with less than 50,000 persons in the population are aggregated with other nationalities.
   - If necessary, the categories of all other variables are aggregated in such a way that the frequency counts in the population cover at least approx. 5,000 persons. The only exceptions are residual categories already aggregated, whose informational content is little specific. This collapsing essentially concerns the variables “occupation” and “industry”.

- The distribution of the variable "age" was aggregated at the left and the right end of the scale (younger than 14, older than 70).

Despite of the reduction of information caused by the anonymization process a lot of potential remains for scientific investigation. Also the integrity of the data is protected to a large extent. Serious restrictions have to be faced only for the regional specification. Of course, the reduction to East and West Germany represents an important shortcoming for researchers, in particular, if they are interested in regional analyses. However, a considerable risk of disclosure unique establishments would remain for delivering deeply structured regional specification, together with the information about the industry classification. Thus, it is not responsable to give more detailed information on the variables "economical branch" and "regional placement of the establishment" on the same time. Only for one of those two variables a detailed classification is possible. As an alternative to the selected procedure the anonymization of all other establishment data would be applicable. However, this would lead again to substantial informational restrictions in the data. On one hand establishment data are to be anonymized more extensively than employee data, and on the other hand the solutions in developing disclosure control for establishment data are so far less satisfactory than necessary (Brand, 1999).

Researchers who want to analyze the two variables "branch (industry)" and "region" in a more detailed way get another chance to do so. The solution of the described dilemma of reduced analytical possibilities because of suppressed variables or collapsed categories should be the installation of a special *research counter office.* In this framework then e.g. differentiated regional analyses also can be accomplished. In principle, all research projects which need more detailed classifications of variables than available in the anonymized employment panel should use the capability of the counter office. At first, external researchers are required to send their analysis programs (command files) by e-mail to the counter office. After an examination of statistical confidentiality these programs are executed (at FEA) by FEA staff who proceed the results (output files) back to the researchers. The very successful IAB establishment panel is a proof of the practicability of this procedure (Koelling, 2001).

3.4 Differences between FEA Employment Panel and IAB Employment Subsample

Apart from the FEA's employment panel, the anonymized data of the IAB employment subsample and the regional employment subsample are already available (Bender et al., 2000; Haas, 2001). Although these datasets are based to a large extent on the same primary data source, namely the employment notifications, nevertheless there are some substantial differences. In the case

of the employment panel the data are processed as single cross-sections (with quarterly deadlines) which are joined to the panel. On the other hand the employment subsample consists of event history data that contain the entire employment processes from the beginning up to the end of employment spells of individual employees. The periods of employment are measured in days. Therefore the employment subsample is particularly suitable for analyzing occupational careers or working lives. Even if the employment panel also contains the individual status changes of the basic employment processes, however, it exists only in a rougher time slot pattern (quarterly). The panel has the advantages that in the first place the data represent the official quarterly data very well, and so, it's directly usable for deadline-oriented analysis. In the second place it also offers possibilities for longitudinal research designs like time series analysis, panel analysis, and cohort analysis.

We have to point out that the cross-sectional data of the employment panel are not easily restorable from the event history data of the IAB employment subsample. The reason is that there are different data transformations in both cases. The quarterly files of the employment statistic, on which the panel is based as a sample, contain special transformation procedures. Therefore notifications not registered within the six-months waiting period at the FEA become balanced. If researchers plan to make cross-sectional-referred computations, for example comparisons between quarters or years, using the panel avoids a complicated transfer of event history data into cross-sectional data, and besides, the panel facilitates forecasts on the official data.

Additionally, it is also more difficult to derivate the cross-sections from the event history data of the employment subsample, because these data are anonymized (i.e. shifted) in the profile. The employment panel so far contains non-anonymized profiles, because the single disclosable employment career has gaps of three months, and thus it exhibits a process-related inaccuracy. Therefore arising changes of occupation or employment are only recorded quarterly.

The scope of time in the two datasets is quite different. While the IAB employment subsample covers a very long time period (1975 – 1997), the employment panel refers so far to a short time interval (1998 – 2002). Thus, the panel concentrates on the current time period. This supply of more current data is possible because the quarterly files are provided with waiting periods of six months. On the other hand, considerably longer waiting periods (at least 18 months) are necessary to build the IAB employment subsample, in order to be able to include late arriving employment notifications, too.

With regard to the contents, the IAB employment subsample offers the advantage that it also includes data about periods of unemployment (e.g. unemployment benefit or assistance). These data are to be added to the employment panel in the next version.

In summary, the differences of employment panel and employment subsample exist in unique data transformations and the different analysis designs. With the employment subsample the emphasis is on event history analysis, with the employment panel it is on analysis of deadline-oriented cross-sections and, if necessary, changes between deadlines (trends). Depending upon the research design, both datasets have their pros and cons.

3.5 Data Access

Due to the data protection law for social insurance data, some conditions must be considered when delivering the employment panel. Firstly, the panel so far may not be delivered into other countries than Germany. Secondly, scientific institutions which would like to order the panel may use the data only for the purpose of labour market or occupational research. Thirdly (and finally), the time of use is limited to the respective research project. For determining the formal conditions a user contract is obligatory. The employment panel may be only used as *scientific use file* in the framework of research projects, and not in connection with (university) courses.

The waves of the employment panel are released as a package of raw data files (ASCII). In uncompressed form the 18 waves need approx. 2.2 GB disk memory. The fee of data supply is 75 . The delivery contains a detailed data description (codebook) including basic frequency counts, as well as SAS set-up files and test data.

4. Prospect

With the FEA's employment panel, a further milestone was reached in enlarging the access to administrative microdata. This enables us to deliver a further prepared sample for research on labour market and occupation. The next steps concerning this new installed panel dataset should attain some methodical improvements and some extensions in its contents.

One aim is a new assignment of the gross payment liable to social security. The assignment of the payment has to be performed in a retrospective way. This leads to more exact information.

At present the data are assigned to the concerning current waves. Unfortunately, the retrospective assignment can only be reached with considerable expenditure. The main reason is that the notification procedure delivers the information of the payment with a time lag.

Another task will be the integration of further data sources. Gaps that are found in the waves of the panel caused by individual unemployment are to be filled with information about the period of unemployment, the receipt of un-

employment benefit or the participation in employment-creation measures. These additional features will extend again the usefulness of the panel for other research aims. Finally, it is still to be pointed out that in the future the employment panel is taken care of by the research data center of the FEA. Annual updates of the dataset should be made available there.

References

Bender, S. / *Haas,* A. / *Klose,* C. (2000): The IAB Employment Subsample 1975 – 1995, Schmollers Jahrbuch – Journal of Applied Social Science Studies, Vol. 120 (4), 649 – 662.

Brand, R. (1999): Anonymität von Betriebsdaten, Beiträge zur Arbeitsmarkt- und Berufsforschung, Vol. 237, Nürnberg.

Bundesanstalt für Arbeit (ed.) (2003): BA-Beschäftigtenpanel, 1. Quartal 1998 – 2. Quartal 2002, Codebuch, Nürnberg.

Haas, A. (2001): Die IAB-Regionalstichprobe 1975 – 1997, ZA-Information, Vol. 48, 128 – 139.

Köhler, S. (1999): Anonymisierung von Mikrodaten in der Bundesstatistik und ihre Nutzung – Ein Überblick, in: Statistisches Bundesamt (ed.), Methoden zur Sicherung der statistischen Geheimhaltung, Schriftenreihe Forum der Bundesstatistik, Vol. 31, Stuttgart.

Köhler, S. / *Schimpl-Neimanns,* C. / *Schwarz,* N. (2000): Pilotprojekt zur Erleichterung der Nutzungsmöglichkeiten von faktisch anonymisierten Mikrodaten, Wirtschaft und Statistik, Vol. 1 / 2000, 30 – 37.

Kölling, A. (2001): Ein "Schalter" für die Forschung, IAB-Werkstattbericht No. 9, Nürnberg.

KVI (2001): Towards an Improved Statistical Infrastructure – Summary Report of the Commission set up by the Federal Ministry of Education and Research (Germany) to Improve the Statistical Infrastructure in Cooperation with the Scientific Community and Official Statistics, Schmollers Jahrbuch – Journal of Applied Social Science Studies, Vol. 121 (3), 443 – 468.

Lechner, S. / *Pohlmeier,* W. (2003): Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten, Vortrag für die Tagung "Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" des Statistischen Bundesamtes und des IAW, 20. / 21. März 2003.

Müller, W. / *Blien,* U. / *Knoche,* P. / *Wirth,* H. (1991): Die faktische Anonymität von Mikrodaten, in: Statistisches Bundesamt (ed.), Schriftenreihe Forum der Bundesstatistik, Vol. 19, Stuttgart.

Neidert, A. (1998): Neues Meldeverfahren zur Sozialversicherung ab 1999, Deutsche Rentenversicherung, Vol. 5 / 1998, 315 – 330.