Institute of Employment Research, Nuremberg/ Germany

Technical report: Linked-Employer-Employee-Data (LEE-Data) from the German Federal Agency of Employment

Table 1 summarizes several basic characteristics of German LEE-Data versions. Row one gives an overview for aggregated individual data on plant level, the rows two and three describe links of individual data with the IAB-Establishment-Panel. The second row is a (yet) not applicable maximum, row three shows characteristics of the present available data.

Table **1: characteristics of German LEE-Data versions**

all information are based on workers involved in social insurance (about eighty percent of total workforce)			
	administrative data	***combined with the IAB-Establishment-Panel***	
		„full" version*	***yearly version***
Country	**Germany**	**Germany**	**Germany**
plant link variable present on worker file	national plant id	national plant id	
additional link records required	no	yes	
worker and plant universe coincide	yes	no	
frequency of link	daily precise spells	yearly	
data source	national register ("Beschäftigten-historik")	national register (individual data) questionaire (plant data)	
id has checksum	no	yes	
level of observation	reporting unit (plant)	reporting unit (plant)	
source of industry	national register	national register, self-reported plant report	
firmsize is reported/computed/both	computed	both	

* maximum not yet applicable

(continued next page)

** version applicable

	administrative data	*in combination with the IAB-Establishment-Panel*	
		*„full“ version**	*yearly version***
source of firmsize	self-reported plant report for individuals aggregated on plant level	self-reported plant report for individuals aggregated on plant level, questionaire	
merge var firmsize	same file	same file/ IAB-Establishment-Panel (both)	
age of plant always known	since 1975	~ 80 percent	
freq of plant obs. per year	daily (365)	daily (individual)/ 1 (plants)	1 (30th of June)
percent of plant universe observed	100 percent of all plants with at least one employee connected to social insurance	like administrative data, but selection probability depends on plant size (see table 2) (5 percent)	
source of date of birth (DOB)	national register	national register	
merge var DOB	same file	same file	
source of gender	national register	national register	
merge var gender	same file	same file	
source of education	national register	national register	
merge var education	same file	same file	
source of occupation	national register	national register	
merge var occupation	same file	same file	
unit on employment history	person	person	person tied to plant
start of job always observed	yes	yes	no
plant association of worker known	yes	yes	no
freq of person obs per year	daily precise spells (365)	daily precise spells (365)	one (spell valid at 30th of June)
percent of workers universe observed	100	~ 60 Percent	

* maximum solution not yet applicable

** version for analyses

In practise often it is not comfortable or even possible to work with all of the data described in rows one and two (long computing times, available processor and disk space resources). The reason is the data collection process.

The Institute of Employment Research (IAB) is part of the Federal Agency of Employment, the Bundesagentur für Arbeit, which has the duty by law to create a work history for all persons which are related to the social security system. With a time lag of two years the IAB receives all this individual data and combines them with data from the unemployment register. In principle it is possible not only to observe individual times of employment, but also of unemployment on a daily precise base. Every change in the status of an employee (like finishing apprenticeship training, switching from full- to part time and opposite) is reported by the employer to the social security institutions. If nothing changes, the report has to be made annually. There are legal sanctions for misreporting. The wage variable contains the *daily average* gross wage (including benefits) for such a period, which is at maximum 365 (366) days (or shorter if something relevant is changing). The individual data are organized in spells.

The individual records for any employee connected to social insurance added by information about times of unemployment result in an annually database of over 50 Million spells. The working history for employees was the first time introduced in 1975. In present, data from 2001 are available. There are 26 years of observations resulting in approximately 1.4 Billion spells. It is not easy to handle such a data collection. Researchers have to accept restrictions by adjusting the balance between a sufficient number of observations, computing times and the size of the basic and their working datasets. Based on the original (large) dataset there are several alternatives to get to smaller sample sizes.

A 100 percent dataset of *plants* for instance is created in the IAB by aggregating all the individual data (quarterly, annually). There is an administrative plant identifier in each individual record (table 1). Aggregating all individual data on plant level creates a 100 percent sample of all plants in Germany with at least one employee related to the social security system. A linkable *firm* register is not collected from the Federal Agency of Employment. Linking plants and firms is more or less possible but not covered by the German law at the moment. As a consequence it is yet in general not possible to identify firms in data from the German Federal Agency of Employment.

However, plant data are easy to create by aggregating several individual characteristics on this level. Common variables and therefore plant characteristics are f.e. firm size (number of

employees related to social security), the average wage of a plant, the average wage for different skill or education groups, the sector (5-digit since 1998, can be uniquely recoded in NACE or something similar (f.e. STAN)), and the type of region. Other variables are proportions of different age, skill, education, gender, and nationality groups. Doing this yearly creates the plant register of Germany, in German terms "Betriebsdatei", being a synonym for "Plant Dataset" (PDS). Of course it is possible to link back this data to the individual records resulting in a Linked-Employee-Dataset with limited information of plant characteristics. This is described in column one from table 1.

There are several techniques available to get to easier to handle datasets for specific investigations like random samples of individuals or plants. The most powerful tool for plant level data is the link to the IAB-Establishment-Panel, which is based on a stratified random sample drawn from the PDS.

While the contents of the annual questionnaire are the base for a major part of the available plant characteristic canon, the link to the individual data reflects several characteristics of the IAB-Establishment-Panel. In a technical perspective, an important implication is the over sampling of larger plants. Without weighting there are approximately five percent of all plants from the PDS covered (varying from year to year) and about 60 percent of the total workforce in Germany.

Table 2 shows the distribution of plants by size in the first wave 1993. There is nothing changing in general for the disproportional sample size selection probability. As mentioned before, since 2000 approximately 16.000 plants join the IAB-Establishment-Survey, but this has only a very small effect on the selection probabilities and the response rates in table 3[1].

[1] With the exception of the largest establishments. Panel attrition is difficult to compensate. The (not weighted) number of employees in the individual data library becomes with ongoing time smaller. Some of the largest plants are early panel attrition. It might not only affect the number of individual observations, but also wage statistics.

Table 2: Number of establishments replying, probability of selection and response rates according to size of establishment

Number of workers	Number of establishments surveyed	Selection probability	Number of replying establishments	Response rate
1 - 4	1,072	.0011	625	.67
5-9	431	.0015	250	.64
10-19	466	.0030	299	.71
20-49	862	.0089	542	.70
50-99	535	.0153	350	.72
100-199	543	.0304	376	.77
200-499	923	.0862	615	.74
500-999	479	.1504	304	.71
1.000 – 4.999	1,497	.8765	924	.72
5.000 +	115	.9127	71	.73
Total	**6,923**	**.0043**	**4,356**	**.71**

Source: IAB-Establishment-Panel 1993, Kölling, 2000, 294

The full weighting[2] matrix as the inverse of the selection probability depends not only size, but also on 20 (before 1999: 16) industries. In the following only size implications will be discussed. The tables in the main text contain weighted and not weighted values. There are differences between the two kinds of computation. Both have their own advantages and disadvantages. Values not weighted reflect more the typical larger plant, weighted values the typical plant in the German economy. Weighting the data underestimates the dispersion of several statistics on individual and plant level.

Doing the link and restricting the universe to plants in the IAB-Establishment-Panel has at least two advantages. First, the number of observations is reduced. Second, there are annually approximately 300 additional variables available on plant level. The IAB-Establishment-Panel started in 1993 with approximately 4000 Western Germany plants, since 1996 Eastern Germany States (the "Bundesländer") finance 4000 additional observations. In present, nearly 16.000 plants are asked yearly to employment related matters (2000 about 14.000 plants). Since 2000 the IAB-Establishment-Panel is also representative on the level of (16) States.

With so many plants in the sample, i.e. nearly every larger one, reducing the number of observations for the link to the individual data becomes relative: in 2001 for instance the 16.000 plants cover about 20 percent of the whole workforce in Germany. A data library for

[2] Weighting is done simultaneously for plants and employees in each cell. Employees are considered by the total employment in a specific cell of the size and sector matrix, not (in addition) differentiated by sex, age and education.

the connected individual data would be (still) very large. In the IAB exists an Linked-Employee-Dataset, which contains biographical individual data from 1990 to 1997 of any employee, who works at least one day in a plant, which is part of the Panel in at least one year between 1993 and 1997. In this version, plants are restricted to Western Germany. The resulting database can be prepared only by software tools who compute commands column by column (casewise) like SAS. The running time even from simple programs is long and use a lot of applicable processor resources. The IAB thinks about different and more easy to handle LEE-Data, so called versions for the now available individual data until 2001.

At the moment and in this time window the following version is applicable and used for the analysis. The "yearly version" extracts the specific spell of the individual record, if covering the 30th of June (reference point for all questions in the IAB-Establishment-Panel). The additional condition is being employed in a plant of the IAB-Establishment-Panel in the specific year. Variables concerning unemployment become obsolete for such data. Available individual variables were listed in table 1, a full description can be given on request.

The most important implication of this yearly version concerns panel attrition. Plants and individuals can become completely unobservable in the yearly version. Individual attrition (a missing individual identifier in t+1) are true exits, additional individual identifiers in year t+1 are true entries (in continuing plants). Hence, following the working history of an individual is in this yearly version only possible if the plant has observations over years and if employees do not move between plants. If a plant is panel attrition all their employees become unobservable, too. The change in wages is and can be computed only for continuing workers in continuing plants. In the yearly version it is not possible to compute the change in wages for persons who change plants (with the exception that they change in another plant which is also part of the IAB-Establishment-Panel in year t+1). Please note that this restriction is not a necessary consequence of German LEE-Data, but of the at the moment applicable versions (an exception is the version from 1993 to 1997). In short time there will be yearly versions where individual information are completed by the following event (individual exits) or with information about what was happening before the person is entering a plant which is also part of the IAB-Establishment-Panel (individual entries). Important for the individual level is for instance the wage for workers who change their employer. A quite similar adding of information will be done for panel attrition on plant level by adding forward and backward annual individual observations for dropped out plants including some PDS information.

In order to loose not too many observations in our present applicable yearly version for reasons of panel attrition, all statistics in the main text are calculated on the base of plants in the presented years, which have also an observation in year t-1. Deriving from this, if job tenures were used (change in wages, percentage of workers with job tenure of at least three years) the statistics are based on the maximum of available observations, meaning all employees in plants with at least three years (t, t-1, t-3). The number of observations (not weighted) show, how many individuals get lost for reasons of panel attrition. Five years would produce more panel attrition.

However, there is one consequence for the weighting procedure. If such variables are used, the number of observations is weighted not with cross-sectional, but longitudinal weightings. Such weighting factors are available for all years from 1993 to 2001. The PDS gives the reliable base for both, cross-sectional and longitudinal weighting of the IAB-Establishment-Panel. In 1993 the cross-sectional factor is equal to the longitudinal and we have no observations before this starting year. We negotiated with the data holders to get for the plants in 1993 also the individual records from 1992 and 1990[3]. So Panel attrition has only in 1993 no meaning. This is especially not true for the year 2000 for an additional reason. The number of observations in West Germany increases in 2000 (approximately 5000 plants). Computing the change in wages for 1999/ 2000 is equal with loosing 5000 observations on plant level and all the connected individuals. In 2000 (and in 1995) we use the longitudinal, in 1993 the cross sectional weights for N in tables with weighted values. For reasons of consistency, East German plants are generally excluded, because they would have only one observation in 2000. East Germany joined the survey the first time in 1996.

Especially the wage variable is crucial for the analysis. The original daily wage sum is multiplied times 30 in order to get the total gross wage sum in June of the specific year including all kind of monetary benefits which are related to social security contributions. All wage statistics including those where only one year is needed (f.e. average wage on plant and individual level) are based on continuing full time workers in continuing plants. Apprenticeships and switchers from full to part time and opposite are excluded generally when calculating mean wages on individual and related statistics on plant level.

There is a threshold in the wage variable. It is in each year about times 1.8 of the average earnings of employees. Each year the threshold is a little bit higher. The threshold and the wages in the data are until 1998 in German Marks, later in Euros. All wages are transformed in Euros by dividing the values in German Marks to 1.95583. This is done in the article for

[3] This is a mini version from one of the before mentioned and forthcoming additional datasets.

1993 and 1995. Varying from year to year the threshold is between 3680,- Euros (1993) and about 4000,- Euro (2000).

Between ten and 15 percent of the employees at the upper bound have censored wages[4]. A lower limit exists but is not important for the analysis, because it concerns only part time workers with just a few working hours. Wages at the upper bound are imputed by using a Mincerian earning function augmented by ten occupational and sector dummies. All wages of full-time employees (excluding apprenticeships) were used for a cross-sectional estimate of the predicted values for the censored wages. The correlation of the error-terms is checked on plant level. We follow Gartner (2004) by adding an error term, use the inverse density and add a random error term to impute the wages at the upper bound. A short paper with the mathematics and a STATA ado file is available on request as well as the estimation results for the imputation regression. Please contact for details. The treshhold is a fixed characteristic of the collection process of administrative employee data in Germany. All censored wages are after the estimation and imputation procedure above the upper bound of wages. For the specific analysis, wages are standardized on the year 1993 (nominal wages). The deflators are for the year 1995 about 0.93 and for the year 2000 about 0.92.

The individual data have no information about levels. The switch-rate is the percentage of employees who change their 3-digit occupational code from one year to the other[5]. All rates are now calculated as $2*Event_{it}/(E_t+E_{t-1})$, where E means Employment and events are exits, entries, switches, growth rates etc. . We formed three positions for employees. The lowest are unskilled blue and white collar workers. Their task and responsibility spectrum in the plant is relatively low. Grouping skilled blue and white collar workers in position two takes care of the strong German apprenticeship system. Position three contains several university degrees. The positions can be interpreted as levels, but three are not enough to fill in the statistics for the number of levels. The private sector is created not only via the sector, but also the legal form of a plant. If there is a public ownership, the plant is excluded from the analysis.

There are just a small number of papers existing with which the results can be compared. The mobility rates are comparable f.e. to Bauer/Bender (2002), who use the "old" LEE-Version for turnover-, churning,- job-creation and destruction rates. The results for wages, especially for the standard deviation of log wages, is nearly the same other researchers found with comparable, but nevertheless different (regional) LEE-Data (f.e. Stephan, 2001).

[4] The number of censored and therefore imputed wages decreases over time. It is partly interpreted as a result from panel attrition of the largest plants and the higher wages they pay.

[5] Maybe this column can be scabbed.

Data Access for external researchers

Researchers will have access to several versions presumably at the end of 2004 at the Data Research Center (DRC) of the IAB. The DRC is in the process of implementing the requested infrastructure and clears question concerning the German data protection law in order to give researchers the legal base for using the data. Anyway, in regular it will be not possible to work with LEE-Data outside the IAB. In the IAB, there will be approximately at the end of 2004 four or five work stations with a suitable performance and statistical software (like STATA, SAS[6]) implemented. In the same time it is planned to create detailed descriptions of the data, the several applicable versions[7] and test data in order to limit the needed time in the IAB for working with the original data. The DRC started in 2004, so documentations have to be made first in German, later in English and can in this language not expected earlier than summer 2005. Foreign researchers are recommended to contact the DRC soon, if they like to analyse German LEE-Data via the DRC in order to answer questions about the data protection law[8]. There is still a process of finding standardized access rules for several groups of researchers and datasets, which are conform with the law.

The DRC provides access not only to LEE-Data, but also others, if they are produced by and in the responsibility of the Federal Agency of Employment in Germany, the Bundesagentur für Arbeit and the IAB as a part of it. Concerning LEE-Data, in the year 2005 the following versions will be applicable for external researchers. The link is always done for individual data in connection with the IAB-Establishment-Panel.

(1) Aggregated individual data on plant level (additional variables for the IAB-establishment-Panel)

(2) Yearly versions

(a) like in row 3 of table 1

(b) information before/after an individual entry or exit is added

(c) panel attrition on plant level is compensated is continued on the individual level added by PDS information

[6] Please contact the DRC if you need other software.

[7] For example versions introduced here.

[8] There is for instance a difference between researchers in and outside the EC.

(3) longitudinal version

The individual data library contains daily precise worker information in a period from 1990 to 2001 including individual information about times of subsidized employment and unemployment. There are all persons included, who work at least one day in a plant of a sample drawn from the IAB-Establishment-Panel. It contains each about 2000 plants in East and West Germany. The sample is stratified to four size classes and eight sectors over the whole German economy. The 2000 plants will have at least four observation years in the IAB-Establishment-Panel. There are PDS information added if times of employment are notobservable via the Panel.

Literature:

for an overview over the IAB-Establishment-Panel:

Kölling, A. (2000): European Datawatch: The IAB-Establishment-Panel, Schmollers Jahrbuch, 120 Jg., 291 - 300

for imputing wages at the upper bound:

Gartner, H. (2004): The imputation of wages above the contribution limit in the German IAB employment sample, IAB Working Paper

for comparing wage statistics with other studies (there is no literature existing for the here presented new available data)

Stephan, G. (2001): Firmenlohndifferentiale, eine empirische Untersuchung für die Bundesrepublik Deutschland, Campus Verlag, Frankfurt a. M.

for comparing the mobility of workers with LEE-Data from the IAB in a period of 1993 to 1997:

Bauer, Th./Bender, St. (2002): Technological Change, Organizational Change, and Job Turnover, IZA Discussion Paper No. 570 , Bonn

Annex

In the main text a propensity score matching is used in order to investigate the effects of collective agreement and works councils on a descriptive base. The following annex table shows the results for the underlying probit estimation for the determinants of these two labor market institutions. Plants with more than 1000 employees are excluded. There is just a small number of plants without collective agreement and/ or a works council, too few for a comparison. The plants compared in the text are equal to size and sector. Conditional likelihoods were used for other characteristics shown in the table.