

Test Data of the Sample of Integrated Labour Market Biographies (SIAB)

At the Research Data Centre (FDZ) of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB) users are offered two different modes of access to the weakly anonymous data made available there¹. The Sample of Integrated Labour Market Biographies (SIAB) is available to researchers for analysis during a stay as a guest researcher at the FDZ or via remote execution (see FDZ-Datenreport 1/2013). These types of data access on the one hand and the complex data structure of the SIAB on the other hand make it essential to provide relevant test data for the preparation of programs if the data are to be processed efficiently. On the basis of these test data users can already familiarise themselves with the data in advance, prepare their programs independently, test them and then either bring them along when they visit the FDZ as guest researchers or send them to the FDZ for remote execution.

These test data, which are based on the original data, can only be made available to the public in compliance with the legal requirement that the data be absolutely anonymous. Accordingly, the test data, as a random sample drawn from the SIAB, have to undergo further processing and anonymization steps. At the end of these procedures, test data are available that replicate the structure of the original data as far as possible but have nonetheless been modified using anonymization methods to the extent that any identification of data units (individuals or establishments) can be ruled out.

The most important characteristic of the SIAB, the precise chronological order and, where applicable, the overlapping of episodes from the various data sources included, is retained in the test data. The dates and employment statuses of the corresponding observations are slightly modified within the individual accounts. The allocation of individuals to establishments is randomly modified. The division of the original data into two modules (Individual Data and Establishment Data) is also retained in this form for the test data.

For the absolute anonymisation of the original data a complex “data swapping” algorithm was programmed, with which individual or establishment characteristics can be exchanged randomly within certain clusters. In the simplest case these clusters comprise one single variable, but they may also take into account several variables and dimensions such as a specific source allocation or certain periods of validity of a variable (see Table 2). This procedure is carried out by drawing a

¹ In addition, remote desktop access, which enables direct access to the data from the user's office, is currently being set up. As part of this access, users can draw their own subsamples based on the original data in order to test the operability of their programmes. The test data offered by the FDZ is therefore not required here.

value randomly from the corresponding overall distribution of the sample and then assigning the exchange value instead of the original value. Hence for characteristics that are defined for a specific data source or for certain periods of validity of a variable, only exchange values are used for this data source and this period. If there are no guidelines for variables, data swapping is conducted without restrictions across all data sources and across the entire period of validity of the SIAB.

As a result of the data swapping algorithm the univariate distributions of all of the variables contained as far as possible in the dataset and their periods of validity are retained in virtually the same form as the original data. Relationships between variables over time are lost if the variables do not belong to the same exchange cluster. In some cases, correlations across different data sources within one variable might also be lost.

The technical auxiliary variables that are contained in the original data, which are based solely on information and values concerning other variables, are deleted in the original data and are adapted and generated again after the anonymization procedure for the test data.

For the SIAB numerous variables which are classified as sensitive from the viewpoint of data protection legislation are also provided in their original form following a justified application. These variables are included in the test data and are shown separately in the attached table (see Table 2).

The test data contain a total of 573,319 observations concerning 25,000 fictitious individuals generated by means of data swapping (see Table 1). As a 1.2 percent sample drawn from the SIAB, the test data are not representative of the final product in so far as they only contain individuals whose employment histories are included in the original data with fewer than 50 observations. Furthermore, individuals whose accounts show only employment observations are not displayed in the test data.

Tab. 1 Frequencies in the test data

Data source	Number of observations	Shares (%)
BeH	408,359	70.68 %
LeH	54,238	9.39 %
LHG	23,550	4.08 %
MTH	8,768	1.52 %
XMTH	696	0.12 %
ASU	77,184	13.36 %
XASU	4,990	0.86 %
Total	408,359	70.68 %
Individuals	25.000	

Tab. 2 Description of variables in the test data

Label	Variable	Data handling
Identifiers		
Individual ID	persnr_siab	Random replacement
Establishment ID	betnr_siab	Random replacement
Period of validity		
Original start date	begorig	Dates are randomly modified within the years of start and end dates of each observation. Exceptions are January 1 and December 31. The chronological order remains unchanged.
Original end date	endorig	
Episode start date	begepi	
Episode end date	endepe	
Generated technical variables		
Source of spell	quelle	No modification
Observation counter per person	spell	Generated after data swapping
Year	jahr	No modification
Personal information		
Gender	frau	Random replacement on personal level
Year of irth	gebjahr	Random replacement on personal level
Month of birth	gebmon	Random replacement on personal level
Nationality (**)	nation	Joint random replacement on personal level
Nationality, grouped	nation_gr	
Marital status	famst	Random replacement on personal level
Number of children	kind	Random replacement on personal level
Professional training	ausbildung	Joint random replacement on personal
Professional training (imputed)	ausbildung_imp	
School leaving qualification	schule	
Information on employment. benefit receipt and job search		
Occupation - current/most recent (KIdB 1988)	beruf	Joint random replacement on personal level
Occupational group - current/most recent (KIdB 2010), 3-digit	beruf2010_3	
Occupational sub-group - current/most recent (KIdB 2010), 4-digit (**)	beruf2010_4	

Label	Variable	Data handling
Level of requirement - current/most recent job (KIdB 2010)	niveau	
Reason for cancellation/notification/termination	grund	Joint random replacement on spell level
Daily wage/daily benefit	tentgelt	
Daily wage (incl. one-off special payment)	tentgelt_bonus	
Daily wage (imputed)	tentgelt_imp	
Transition zone	gleitz	
Part-time	teilzeit	
Occupational status and working hours	stib	
Employment status	erwstat	
Measure type - group	mass	
Fixed-term contract	befrist	
Temporary agency work	leih	Random replacement on personal level
Employment status prior to job-search	estatvor	Random replacement on personal level
Employment status after job-search / SGB-II-ground-of-exclusion / availability	estatnach	Random replacement on personal level
Integration forecast	ipo	Random replacement on personal level
Reason for end of previous job	art_kuend	Random replacement on personal level
Working hours of job application	arbeitszeit	Random replacement on personal level
Residual claim/planned duration	restanspruch	Random replacement on personal level
Type of provider	traeger	Random replacement on personal level
Start date of unemployment	alo_beg	Generated after data swapping
Duration of unemployment	alo_dau	Generated after data swapping

Label	Variable	Data handling
Establishment variables		
Classification of economic activity 73	w73_3	Joint random replacement on the establishment level, trying to keep the temporal structure and internal hierarchy within the industry classifications intact
w73_3 completed by extrapolation/imputation	w73_3_gen	
Type of imputation w73_3	group_w73_3	
Classification of economic activity 93, sub-classes (**)	w93_5	
Classification of Economic activity 93, groups	w93_3	
w93_3 completed by extrapolation/imputation	w93_3_gen	
Type of imputation w93_3	group_w93_3	
Classification of economic activity 03, sub-classes (**)	w03_5	
Classification of economic activity 03, groups	w03_3	
Classification of economic activity 08, sub-classes (**)	w08_5	
Classification of economic activity 08, groups	w08_3	
w08_3 completed by extrapolation/imputation	w08_3_gen	
Type of imputation w08_3	group_w08_3	
Year of first appearance	grd_jahr	Joint random replacement on the establishment level, trying to keep the temporal structure intact
Year of last appearance	lzt_jahr	
Entry type (*)	eintritt	
Employment of betnr in year(*)	besch	
Employment predecessor one year before (*)	besch_vor	
Predecessor exits (*)	status_vor	
Inflows from the predecessor to betnr (*)	inflow	
Exit type (*)	austritt	

Label	Variable	Data handling
Employment of betnr in year(*)	besch	
Employment successor one year later (*)	besch_nach	
Successor is entrant (*)	status_nach	
Outflow from betbr to successor (*)	outflow	
No. employees total	az_ges	Joint replacement on the establishment level so that the proportions are retained
No. full-time (regular workers + others)	az_vz	
No. marginal part-time workers	az_gf	
No. female employees (*)	az_f	
No. regular workers (*)	az_reg	
No. trainees/apprentices (*)	az_azubi	
No. employees in partial retirement (*)	az_atz	
No. part-time (regular workers + others) (*)	az_tz	
No. full-time female employees (*)	az_f_vz	
No. part-time female employees (*)	az_f_tz	
No. full-time regular workers (*)	az_reg_vz	
Inflows (*)	ein_ges	
Inflows marginal part-time (*)	ein_gf	
Inflows full-time (regular workers + others) (*)	ein_vz	
Outflows (*)	aus_ges	
Outflows marginal part-time (*)	aus_gf	
Outflows full-time (regular workers + others) (*)	aus_vz	

Label	Variable	Data handling
Mean imp. wage all full-time employees	te_imp_mw	Random replacement on establishment level
Regional Codes		
Place of residence - district (Kreis) (**)	wo_kreis	Joint replacement so that the original hierarchy is retained
Place of residence - federal state (Bundesland)	wo_bula	
Place of residence - employment agency (**)	wo_aa	
Place of residence - regional directorate (Regionaldirektion)	wo_rd	
Place of work - district (Kreis) (*)	ao_kreis	Joint replacement so that the original hierarchy is retained
Place of work - federal state (Bundesland)	ao_bula	

(*) Variable is part of the extension data-set and only available upon justified request

(**) Variable is only available upon justified request