FDZ

Forschungsdatenzentrum

der Bundesagentur für Arbeit
im Institut für Arbeitsmarkt-
und Berufsforschung

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 there. 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 data access (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 anonymisation 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 anonymisation 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 3). This procedure is carried out by drawing a 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. 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 anonymisation 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 3).

The test data contain a total of 161,271 observations concerning 19,543 fictitious individuals generated by means of data swapping (see Table 1). As a 1.1 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 20 observations. Furthermore, individuals whose accounts show only employment observations are not displayed in the test data. These restrictions also explain the differences in the number of observations per source and year compared with the original data (see Table 2).

Tab. 1 Frequencies in the test data

Data source	Number of observations	Shares (%)
BeH	100,533	62.34 %
LeH	22,730	14.09 %
ASU	26,666	16.53 %
LHG	7,401	4.59 %
XASU	678	0.42 %
XLHG	3,263	2.02 %
Total number of obs.	**161,271**	**100.00 %**
Number of individuals	**19,543**	

Tab. 2 Shares of spells per data source and year (row percentages)

Start year of spells	BeH	LeH	ASU	LHG	XASU	XLHG	Total
1975	98.52	1.43	0.05	0.00	0.00	0.00	100.00
1976	83.28	16.72	0.00	0.00	0.00	0.00	100.00
1977	73.64	26.36	0.00	0.00	0.00	0.00	100.00
1978	85.07	14.93	0.00	0.00	0.00	0.00	100.00
1979	84.13	15.87	0.00	0.00	0.00	0.00	100.00
1980	80.91	19.09	0.00	0.00	0.00	0.00	100.00
1981	77.89	22.11	0.00	0.00	0.00	0.00	100.00
1982	77.48	22.49	0.03	0.00	0.00	0.00	100.00
1983	74.82	25.18	0.00	0.00	0.00	0.00	100.00
1984	75.24	24.76	0.00	0.00	0.00	0.00	100.00
1985	75.68	24.32	0.00	0.00	0.00	0.00	100.00
1986	77.37	22.59	0.04	0.00	0.00	0.00	100.00
1987	76.62	23.34	0.04	0.00	0.00	0.00	100.00
1988	73.96	25.95	0.09	0.00	0.00	0.00	100.00
1989	72.15	27.85	0.00	0.00	0.00	0.00	100.00
1990	72.76	26.77	0.38	0.00	0.00	0.10	100.00
1991	65.66	33.63	0.59	0.00	0.00	0.11	100.00
1992	63.29	35.46	1.20	0.00	0.00	0.05	100.00
1993	61.72	35.63	2.55	0.00	0.00	0.11	100.00
1994	59.80	35.98	4.12	0.00	0,00	0.10	100.00
1995	55.24	33.71	10.72	0.00	0.00	0.34	100.00
1996	38.68	21.38	39.88	0.00	0.00	0.06	100.00
1997	23.91	13.93	61.82	0.00	0.00	0.34	100.00
1998	29.81	12.51	56.66	0.00	0.00	1.02	100.00
1999	43.94	10.46	42.10	0.00	0.00	3.50	100.00
2000	48.42	7.99	38.63	0.00	0.00	4.96	100.00
2001	51.64	6.72	37.50	0.00	0.00	4.15	100.00
2002	50.67	8.14	36.57	0.00	0.00	4.63	100.00
2003	55.79	7.73	32.74	0.00	0.00	3.74	100.00
2004	57.30	6.15	31.79	0.00	0.00	4.77	100.00
2005	43.15	2.67	30.00	20.77	1.10	2.31	100.00
2006	51.30	2.11	25.98	16.00	2.02	2.59	100.00
2007	53.67	1.26	23.51	16.20	1.35	4.01	100.00
2008	59.86	2.07	20.35	9.72	1.20	6.80	100.00
2009	57.20	2.53	22.45	9.25	0.67	7.89	100.00
2010	59.85	2.58	21.37	8.88	0.73	6.60	100.00
2011	61.14	2.22	20.16	10.66	0.64	5.18	100.00
2012	60.64	2.42	19.86	12.34	1.27	3.47	100.00
2013	59.27	2.73	21.65	11.74	1.34	3.26	100.00
2014	51.03	3.47	25.34	15.79	1.48	2.89	100.00
Total	**62.34**	**14.09**	**16.53**	**4.59**	**0.42**	**2.02**	**100.00**

Tab. 3 Description of variables

Label	Variable	Data handling
Identifiers		
Artificial individual ID	persnr	Random replacement
Artificial establishment number	betnr	Random replacement
Period of validity		
Original start date of observation	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 of observation	endorig	
Start date of split episode	begepi	
End date of split episode	endepi	
Generated technical variables		
Source of observation	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 Birth	gebjahr	Random replacement on personal level
Nationality (*)	nation	Joint random replacement on personal level
Nationality, aggregated	nation_gr	
Marital status	famst	Random replacement on personal level
Number of children	kind	Random replacement on personal level
Vocational training	ausbildung	Random replacement on personal level
School leaving qualification	schule	Random replacement on personal level
Information on employment. benefit receipt and job search		
Reason for notification/ reason for end of beneift receipt/ reason for discontinuation of unemployment benefit II/ reasonfor deregistration	grund	Random replacement within original data record within person
Daily wage / daily benefit rate	tentgelt	Joint random replacement within original data record
Transition zone	gleitz	
Occupation - current/most recent (KldB 1988)	beruf	Joint random replacement on personal

Occupational group - current/most recent (KldB 2010), 3-digit	beruf2010_3	level
Occupational sub-group - current/most recent (KldB 2010), 4-digit	beruf2010_4	
Level of requirement - current/most recent job (KldB 2010)	niveau	
Part-time	teilzeit	Random replacement on personal level
Employment status	erwstat	Random replacement within original data record within person
Temporary agency work	leih	Random replacement on personal level
Fixed-term job	befrist	Random replacement on personal level
Employment status prior to job-search	estatvor	Random replacement within original data record within person
Employment status after job-search	estatnach	Random replacement within original data record within person
Client profile	profil	Random replacement on personal level
Type of termination to last job	art_kuend	Random replacement on personal level
Desired working hours of the job sought	arbzeit	Random replacement on personal level
Duration of remaining entitlement to unemployment benefit	restanspruch	Random replacement on personal level
Type of institution	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
Establishment variables		
Economic activity 73	w73_3	Joint random replacement on establishment level
Economic activity 73 generated – completed by extrapolation/imputation	w73_3_gen	
Economic activity 73 generated – type of completion	group_w73_3	
Economic activity 93, 5-digit code (*)	w93_5	Joint random replacement on the establishment level within the industry classification hierarchy Please note that the industry codes can
Economic activity 93, 3-digit code	w93_3	
Economic activity 93 generated – completed by extrapolation/imputation	w93_3_gen	

Economic activity 93 generated – type of completion	group_w93_3	change artificially when a new classification becomes valid.
Economic activity 03, 5-digit code	w03_5	
Economic activity 03, 3-digit code	w03_3	
Economic activity 08, 5-digit code (*)	w08_5	
Economic activity 08, 3-digit code	w08_3	
Year of first appearance of establishment	grd_jahr	Joint random replacement on the establishment level
First appearance of establishment (*)	grd_dat	
Year of last appearance of establishment	lzt_jahr	Joint random replacement on the establishment level
Last appearance of establishment (*)	lzt_dat	
Total number of employees	az_ges	Joint replacement on the establishment level so that the proportions are retained
Number of employees with full-time job	az_vz	
Number of marginally employed	az_gf	
Mean imputed daily gross income of employees with full-time job	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	Joint replacement so that the original hierarchy is retained
Place of residence: regional directorate	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 only available upon justified request