Forschungsdatenzentrum
der Bundesagentur für Arbeit
im Institut für Arbeitsmarkt-
und Berufsforschung

Testdaten zur Stichprobe Integrierter Arbeitsmarktbiografien (SIAB)

Am Forschungsdatenzentrum der BA im IAB (FDZ) werden dem Nutzer zwei unterschiedliche Wege des Zugangs zu den dort bereitgestellten schwach anonymisierten Daten eröffnet. Die Stichprobe der Integrierten Arbeitsmarktbiografien (SIAB) steht Wissenschaftlern im Rahmen eines Gastaufenthalts sowie in der Datenfernverarbeitung für Auswertungen zur Verfügung. Diese Arten des Datenzugangs einerseits und die komplexe Datenstruktur der SIAB andererseits machen es für deren effiziente Verarbeitung unerlässlich, dass für die Vorbereitung von Syntax-Programmen auch entsprechende Testdaten bereitgestellt werden. Auf deren Grundlage können Nutzer die Daten bereits vorab kennenlernen, ihre Syntax-Programme eigenständig vorbereiten, testen und dann entweder zum Gastaufenthalt mitbringen oder zur Datenfernverarbeitung an das FDZ senden.

Die öffentliche Bereitstellung dieser, auf den Originaldaten basierenden, Testdaten kann nur unter Beachtung der rechtlichen Vorgabe einer absoluten Anonymität der Daten erfolgen. Entsprechend müssen die Testdaten als Zufallsstichprobe aus der SIAB noch weitere Bearbeitungs- und Anonymisierungsschritte durchlaufen. Am Ende dieser Schritte stehen Testdaten, welche die Struktur der Originaldaten soweit wie möglich replizieren und dabei durch den Einsatz von Anonymisierungsmethoden dennoch soweit verfremdet sind, dass für Untersuchungseinheiten (Personen bzw. Betriebe) eine Deanonymisierung ausgeschlossen werden kann.
Die wichtigste Eigenschaft der SIAB, die exakte zeitliche Abfolge und gegebenenfalls auch Überlappung von Episoden aus den verschiedenen enthaltenen Datenquellen, bleibt in den Testdaten bestehen. Datumsangaben und Erwerbsstatus der entsprechenden Datensätze werden innerhalb der Personenkonten leicht verfremdet. Die Zuordnung von Personen zu einzelnen Betrieben wird zufällig verändert.

Für die absolute Anonymisierung der Originaldaten wurde ein komplexer „Data-Swapping“ Algorithmus programmiert, mit dessen Hilfe individuelle bzw. betriebliche Charakteristika zufällig innerhalb bestimmter Cluster getauscht werden. Diese Cluster umfassen im einfachsten Fall ein einzelnes Merkmal, können aber auch mehrere Merkmale sowie Dimensionen wie eine spezifische Quellenzuordnung oder bestimmte Gültigkeitszeiträume eines Merkmals berücksichtigen (vgl. Tabelle 3). Die Umsetzung erfolgt durch das zufällige Ziehen eines Wertes aus der entsprechenden Gesamtverteilung der Stichprobe und durch anschließende Zuordnung des gezogenen Tauschwerts anstatt des Originalwertes. So werden für Merkmale, die quellenspezifisch oder für bestimmte Gültigkeitszeiträume eines Merkmals definiert werden, auch nur Tauschwerte für diese Quelle und

diesen Zeitraum herangezogen. Gibt es für Merkmale keine Vorgaben, wird ohne Restriktionen quellenübergreifend sowie über den ganzen Gültigkeitszeitraum der SIAB getauscht.
Durch den Tauschalgorithmus bleiben die univariaten Verteilungen aller im Datensatz enthaltenen Merkmale sowie deren Gültigkeitszeiträume so weit wie möglich originalgetreu erhalten. Zusammenhänge zwischen Variablen im Zeitverlauf gehen dann verloren, wenn diese nicht gemeinsam Bestandteil eines Tauschclusters sind.
Die in den Originaldaten enthaltenen technischen Hilfsmerkmale, die ausschließlich auf Informationen und Werten anderer Variablen beruhen, werden im Original gelöscht und nach Abschluss der Anonymisierungsverfahren für die Testdaten angepasst und erneut generiert.
Für die SIAB werden im Original auch zahlreiche datenschutzrechtlich als sensibel eingestufte Merkmale auf begründeten Antrag hin bereitgestellt. Diese Merkmale sind in den Testdaten enthalten und werden in Tabelle 3 gesondert ausgewiesen.
Die strukturelle Aufteilung der Originaldaten in zwei Datenmodule (Personen- und Betriebsdatei) sowie deren Dateinamen werden auch für die SIAB Testdaten übernommen.

Die Testdaten enthalten insgesamt 161.271 Sätze zu 19.543 fiktiven, durch Tauschen generierten Personen (vgl. Tabelle 1). Die Testdaten sind als ca. 1,1%-Stichprobe (Personen) aus der SIAB insofern nicht repräsentativ für das Endprodukt, da sie nur Personen beinhalten, deren Erwerbsverlauf mit weniger als 20 Sätzen in den Originaldaten enthalten ist. Außerdem werden Personen, die ausschließlich Beschäftigungssätze in ihrem Konto aufweisen, nicht in die Testdaten übernommen. Diese Einschränkungen erklären auch die Abweichungen in der Anzahl der Sätze pro Quelle und Jahr im Vergleich zu den Originaldaten (vgl. Tabelle 2).

Tabelle 1: Auszählung der Testdaten

Datenquelle	Anzahl der Spells	Anteil der Spells (%)
BeH	100.533	62,34 %
LeH	22.730	14,09 %
ASU	26.666	16,53 %
LHG	7.401	4,59 %
XASU	678	0,42 %
MTH	3.263	2,02 %
Gesamt	**161.271**	**100,00 %**
Personen	**19.543**	

Tabelle 2: Anteilswerte der Sätze nach Jahr und Quelle (in Zeilenprozenten)

Beginnjahr des Spells	BeH	LeH	ASU	LHG	XASU	MTH	Gesamt
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
Gesamt	**62,34**	**14,09**	**16,53**	**4,59**	**0,42**	**2,02**	**100,00**

Tabelle 3: Genese der Variablen in den Testdaten

Bezeichnung	Variable	Genese in den Testdaten
Identifikatoren		
Systemfreie Personennummer	persnr	Zufällige Ersetzung
Systemfreie Betriebsnummer	betnr	Zufällige Ersetzung
Gültigkeitszeitraum		
Beginndatum Originalbeobachtung	begorig	Jede Datumsangabe außer dem 1.1. und dem 31.12. werden innerhalb der tatsächlichen Beginn- und Endejahre mit einem fiktiven zufällig generierten Datum ersetzt. Die Reihenfolge der Sätze bleibt erhalten.
Endedatum Originalbeobachtung	endorig	
Beginndatum der Episode	begepi	
Endedatum der Episode	endepi	
Generierte technische Merkmale		
Quelle des Satzes	quelle	Keine Veränderung
Satzzähler pro Konto	spell	Wird für die Testdaten neu berechnet
Jahr	jahr	Keine Veränderung, da Beginnjahr nicht getauscht wird
Informationen zur Person		
Geschlecht	frau	Austausch auf Personenebene
Geburtsjahr	gebjahr	Austausch auf Personenebene
Staatsangehörigkeit (*)	nation	Gemeinsamer Austausch auf Personenebene
Staatsangehörigkeit vergröbert	nation_gr	
Familienstand	famst	Austausch auf Personenebene
Kinderzahl	kind	Austausch auf Personenebene
Ausbildung	ausbildung	Austausch auf Personenebene
Schulausbildung	schule	Austausch auf Personenebene
Information zu Beschäftigung, Leistungsbezug und Arbeitssuche		
Abgabegrund/Beendigungsgrund/SGB-II Einstellungsgrund/Abmeldegrund	grund	Austausch auf Satzebene innerhalb einer Person
Tagesentgelt / täglicher Leistungssatz	tentgelt	Gemeinsamer Austausch auf Satzebene
Gleitzone	gleitz	
Beruf - ausgeübte/letzte Tätigkeit (KldB 1988)	beruf	Gemeinsamer Austausch auf Personenebene

Berufsgruppe – ausgeübte/letzte Tätigkeit (KldB 2010), 3-Steller	beruf2010_3	
Berufsuntergruppe – ausgeübte/letzte Tätigkeit (KldB 2010), 4-Steller	beruf2010_4	
Anforderungsniveau – ausgeübte/letzte Tätigkeit (KldB 2010)	niveau	
Teilzeit	teilzeit	Austausch auf Personenebene
Erwerbsstatus	erwstat	Austausch auf Satzebene innerhalb einer Person
Leiharbeit	leih	Austausch auf Personenebene
Befristung	befrist	Austausch auf Personenebene
Erwerbsstatus vor Arbeitssuche	estatvor	Austausch auf Satzebene innerhalb einer Person
Status nach Arbeitssuche	estatnach	Austausch auf Satzebene innerhalb einer Person
Profillage	profil	Austausch auf Personenebene
Art der Kündigung der letzten Tätigkeit	art_kuend	Austausch auf Personenebene
Arbeitszeit des Stellengesuchs	arbzeit	Austausch auf Personenebene
Restanspruch Arbeitslosengeld	restanspruch	Austausch auf Personenebene
Trägerart	traeger	Austausch auf Personenebene
Beginn der Arbeitslosigkeit	alo_beg	Wird für die Testdaten neu berechnet
Dauer der Arbeitslosigkeit	alo_dau	Wird für die Testdaten neu berechnet
Betriebsmerkmale		
Wirtschaftszweig 73	w73_3	Gemeinsamer Austausch auf Betriebsebene
Wirtschaftszweig 73 generiert – vervollständigt durch Extrapolation	w73_3_gen	
Wirtschaftszweig 73 generiert – Art der Vervollständigung	group_w73_3	
Wirtschaftszweig 93, 5-Steller (*)	w93_5	Gemeinsamer Austausch auf Betriebsebene innerhalb der jeweiligen Klassifikationen, so dass die interne Hierarchie zwischen Wirtschaftsunterklasse und -gruppe erhalten bleibt. Beim Wechsel der Klassifikation können Sprünge auftreten.
Wirtschaftszweig 93, 3-Steller	w93_3	
Wirtschaftszweig 93 generiert – vervollständigt durch Extrapolation	w93_3_gen	
Wirtschaftszweig 93 generiert – Art der Vervollständigung	group_w93_3	
Wirtschaftszweig 03, 5-Steller (*)	w03_5	

Wirtschaftszweig 03, 3-Steller	w03_3	
Wirtschaftszweig 08, 5-Steller(*)	w08_5	
Wirtschaftszweig 08, 3-Steller	w08_3	
Jahr des ersten Auftretens des Betriebes	grd_jahr	Gemeinsamer Austausch auf Betriebsebene
Erstes Auftreten der Betriebsnummer (*)	grd_dat	
Jahr des letzten Auftretens des Betriebs	lzt_jahr	Gemeinsamer Austausch auf Betriebsebene
Letztes Auftreten der Betriebsnummer (*)	lzt_dat	
Anzahl der Beschäftigten gesamt	az_ges	Gemeinsamer Austausch auf Betriebsebene, so dass die Größenverhältnisse erhalten bleiben
Anzahl Vollzeit (Normalbeschäftigte + sonstige)	az_vz	
Anzahl geringfügig Beschäftigte	az_gf	
Mittelwert imputiertes Bruttotagesentgelt Vollzeitbeschäftigte	te_imp_mw	Austausch auf Betriebsebene
Ortsangaben		
Wohnort Kreis (*)	wo_kreis	Gemeinsamer Austausch der Wohnorte, so dass die Hierarchie erhalten bleibt
Wohnort Bundesland	wo_bula	
Wohnort Arbeitsagentur (*)	wo_aa	Gemeinsamer Austausch der Wohnorte, so dass die Hierarchie erhalten bleibt
Wohnort Regionaldirektion	wo_rd	
Arbeitsort Kreis (*)	ao_kreis	Gemeinsamer Austausch der Arbeitsorte, so dass die Hierarchie erhalten bleibt
Arbeitsort Bundesland	ao_bula	

(*) Merkmal steht in den SIAB Originaldaten nur auf gesonderten Antrag zur Verfügung