

## Testdaten zur Stichprobe Integrierter Arbeitsmarktbiografien (SIAB)

Am Forschungsdatenzentrum der BA im IAB (FDZ) werden den Nutzenden zwei unterschiedliche Wege des Zugangs zu den dort bereitgestellten schwach anonymisierten Daten eröffnet<sup>1</sup>. Die Stichprobe der Integrierten Arbeitsmarktbiografien (SIAB) steht Forschenden im Rahmen eines Gastaufenthalts sowie in der kontrollierten 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 Nutzende 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 2). Die Umsetzung erfolgt durch das zufällige Ziehen eines Wertes aus der entsprechenden

---

<sup>1</sup> Derzeit befinden sich zusätzlich der Remote-Desktop-Zugang, der einen direkten Zugriff auf die Daten vom Büro der Nutzenden ermöglicht, im Aufbau. Im Rahmen dieses Zugangs können sich die Nutzenden auf Basis der Originaldaten eigene Substichproben ziehen, um die Lauffähigkeit ihrer Programme testen zu können. Die von FDZ angebotenen Testdaten werden hier somit nicht benötigt.

Gesamtverteilung der Stichprobe und durch anschließende Zuordnung des gezogenen Tauscherts anstatt des Originalwertes. So werden für Merkmale, die quellspezifisch oder für bestimmte Gültigkeitszeiträume eines Merkmals definiert werden, auch nur Tauscherte 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. Auch ist davon auszugehen, dass zum Teil quellübergreifende Zusammenhänge innerhalb eines Merkmals verloren gehen.

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 2 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 570.012 Sätze zu 25.000 fiktiven, durch Tauschen generierten Personen (vgl. Tabelle 1). Die Testdaten sind als ca. 1,2%-Stichprobe (Personen) aus der SIAB insofern nicht repräsentativ für das Endprodukt, da sie nur Personen beinhalten, deren Erwerbsverlauf mit weniger als 50 Sätzen in den Originaldaten enthalten ist. Außerdem werden Personen, die ausschließlich Beschäftigungssätze in ihrem Konto aufweisen, nicht in die Testdaten übernommen.

**Tabelle 1: Auszählung der Testdaten**

Datenquelle	Anzahl der Spells	Anteil der Spells (%)
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 %

Datenquelle	Anzahl der Spells	Anteil der Spells (%)
XASU	4.990	0,86 %
<b>Gesamt</b>	<b>577.776</b>	<b>100,00 %</b>
<b>Personen</b>	<b>25.000</b>	

**Tabelle 2: Genese der Variablen in den Testdaten**

Bezeichnung	Variable	Genese in den Testdaten
Identifikatoren		
Systemfreie Personennummer	persnr_siab	Zufällige Ersetzung
Systemfreie Betriebsnummer	betnr_siab	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.
Enddatum Originalbeobachtung	endorig	
Beginndatum Episode	begepi	
Enddatum Episode	endeapi	
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
Geburtsmonat (**)	gebmon	Austausch auf Personenebene
Staatsangehörigkeit (**)	nation	Gemeinsamer Austausch auf Personenebene
Staatsangehörigkeit vergrößert	nation_gr	
Familienstand	famst	Austausch auf Personenebene
Kinderzahl	kind	Austausch auf Personenebene
Ausbildung	ausbildung	Gemeinsamer Austausch auf Personenebene
Ausbildung (imputiert)	ausbildung_imp	

Bezeichnung	Variable	Genese in den Testdaten
Schulausbildung	schule	
<b>Information zu Beschäftigung, Leistungsbezug und Arbeitssuche</b>		
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	
Abmeldegrund / Abgabegrund / Beendigungsgrund	grund	Gemeinsamer Austausch auf Satzebene
Tagesentgelt / täglicher Leistungssatz	tentgelt	
Tagesentgelt (inkl. Einmalzahlungen)	tentgelt_bonus	
Tagesentgelt (imputiert)	tentgelt_imp	
Gleitzone	gleitz	
Teilzeit	teilzeit	
Stellung im Beruf	stib	
Erwerbsstatus	erwstat	
Maßnahmeart – Gruppe (**)	mass	
Befristung	befrist	
Leiharbeit	leih	Austausch auf Personenebene
Erwerbsstatus vor Arbeitssuche	estatvor	Austausch auf Personenebene
Status nach Arbeitssuche / SGB-II-Ausschlussgrund / Verfügbarkeit	estatnach	Austausch auf Personenebene
Integrationsprognose	ipo	Austausch auf Personenebene
Art der Kündigung der letzten Tätigkeit	art_kuend	Austausch auf Personenebene
Arbeitszeit des Stellengesuchs	arbzeit	Austausch auf Personenebene
Restanspruch/geplante Dauer	restanspruch	Austausch auf Personenebene
Trägerart	traeger	Austausch auf Personenebene
Beginndatum der Arbeitslosigkeit	alo_beg	Wird für die Testdaten neu berechnet

Bezeichnung	Variable	Genese in den Testdaten
Dauer der Arbeitslosigkeit	alo_dau	Wird für die Testdaten neu berechnet
<b>Betriebsmerkmale</b>		
WZ 73 3-Steller	w73_3	Gemeinsamer Austausch auf Betriebsebene, so dass die zeitliche Struktur und die interne Hierarchie zwischen Wirtschaftsunterklasse und -gruppe erhalten bleibt
w73_3 vervollständigt durch Extrapolation/Imputation	w73_3_gen	
Art der Vervollständigung w73_3	group_w73_3	
WZ 93 5-Steller (**)	w93_5	
WZ 93 3-Steller	w93_3	
W93_3 vervollständigt durch Extrapolation/Imputation	w93_3_gen	
Art der Vervollständigung w93_3	group_w93_3	
WZ 03 5-Steller (**)	w03_5	
WZ 03 3-Steller	w03_3	
WZ 08 5-Steller (**)	w08_5	
WZ 08 3-Steller	w08_3	
W08_3 vervollständigt durch Extrapolation/Imputation	w08_3_gen	
Art der Vervollständigung w08_3	group_w08_3	
Jahr erstes Auftreten Betriebesnummer	grd_jahr	Gemeinsamer Austausch auf Betriebsebene, so dass die zeitliche Struktur erhalten bleibt
Jahr letztes Auftreten Betriebsnummer	lzt_jahr	
Gründungsstatus (*)	eintritt	
Beschäftigte betnr (*)	besch	
Beschäftigte Vorgänger im Vorjahr (*)	besch_vor	
Status Vorgänger (*)	status_vor	
Inflows aus dem Vorgänger zu betnr (*)	Inflow	
Schließungsstatus (*)	Austritt	
Beschäftigte betnr (*)	Besch	
Beschäftigte Nachfolger im Folgejahr (*)	besch_nach	
Status Nachfolger (*)	status_nach	

Bezeichnung	Variable	Genese in den Testdaten
Outflows aus betnr zum Nachfolger (*)	Outflow	
Anzahl Beschäftigte gesamt	az_ges	Gemeinsamer Austausch auf Betriebsebene, so dass die Größenverhältnisse erhalten bleiben
Anzahl Vollzeit (Normalbeschäftigte + sonstige)	az_vz	
Anzahl geringfügig Beschäftigte	az_gf	
Anzahl Frauen (*)	az_f	
Anzahl Normalbeschäftigte (*)	az_reg	
Anzahl Auszubildende Pers.gr. (*)	az_azubi	
Anzahl in Altersteilzeit (*)	az_atz	
Anzahl Teilzeit (Normalbeschäftigte + sonstige) (*)	az_tz	
Anzahl Frauen Vollzeit (*)	az_f_vz	
Anzahl Frauen Teilzeit (*)	az_f_tz	
Anzahl Normalbeschäftigte Vollzeit (*)	az_reg_vz	
Eintritte gesamt (*)	ein_ges	
Eintritte geringfügige Beschäftigte (*)	ein_gf	
Eintritte Vollzeit (Normalbeschäftigte + sonstige) (*)	ein_vz	
Austritte gesamt (*)	aus_ges	
Austritte geringfügige Beschäftigte (*)	aus_gf	
Austritte Vollzeit (Normalbeschäftigte + sonstige) (*)	aus_vz	
Mittelwert imp. Bruttotagesentgelt Vollzeitbeschäftigte	te_imp_mw	Austausch auf Betriebsebene
<b>Ortsangaben</b>		
Wohnort - Kreis (**)	wo_kreis	Gemeinsamer Austausch der Wohnorte, so dass die Hierarchie erhalten bleibt
Wohnort - Bundesland	wo_bula	
Wohnort - Arbeitsagentur (**)	wo_aa	
Wohnort - Regionaldirektion	wo_rd	
Arbeitsort - Kreis (**)	ao_kreis	Gemeinsamer Austausch der Arbeitsorte, so dass die Hierarchie erhalten bleibt
Arbeitsort - Bundesland	ao_bula	



**(\*) Merkmal ist Teil der Erweiterungsdatensätze des BHP und steht nur auf gesonderten Antrag zur Verfügung**

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