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

Genese der IEBS-Testdaten

Das Ziel der Erstellung der IEBS-Testdaten liegt in der absoluten Anonymisierung der Daten bei gleichzeitigem Erhalt wichtiger Datenstrukturen, damit anhand der Testdaten Auswertungssyntaxen für die Originaldaten vorbereitet werden können.

Als wichtige Eigenschaft der IEBS werden die zeitliche Abfolge und die Überlappung von Spells aus den verschiedenen Datenquellen angesehen. Datumsangaben und Erwerbsstatus werden daher möglichst wenig verfremdet.

Ein Großteil der Variablen wird zufällig ersetzt, indem und ohne Zurücklegen aus der Verteilung des Merkmals in der Stichprobe gezogen wird. Praktisch umgesetzt wird dies, indem die Variablen separat sortiert und anschließend wieder zusammengespielt werden. Zwei Varianten sind dabei zu unterschieden:

Tauschvariante 1: Merkmale, die je nach Quelle unterschiedlich definiert sind, werden getrennt nach Quellen neu zugeordnet.

Tauschvariante 2: Merkmale, die über mehrere Quellen gleich definiert sind, werden quellenübergreifend neu zugeordnet.

Für beide Varianten gilt: Ist eine Merkmalsausprägung in den Originaldaten über mehrere Spells konstant, dann ist sie das auch in den Testdaten.

Durch diese Art der Ersetzung bleiben die univariaten Verteilungen weitgehend erhalten. Zusammenhänge zwischen Variablen oder einer Variablen im Zeitablauf (steigendes Ausbildungsniveau) gehen jedoch verloren.

Eine dritte Gruppe von Variablen sind technische Hilfsmerkmale, die ausschließlich auf den Informationen der anderen Variablen beruhen. Diese Merkmale werden im Anschluss an die Ersetzung der anderen Merkmale neu berechnet.

Die Testdaten enthalten 49.985 Spells zu 5.468 fiktiven Personen. Die Testdaten sind insofern nicht repräsentativ für die IEBS, da Personen mit mehr als 20 Spells und Personen mit ausschließlich BeH-Spells nicht in den Testdaten enthalten sind.

Anzahl der Spells nach Geschlecht und Datenquelle

Datenquelle	-9	-7	Frauen	Männer	Gesamt
BeH			16.701	15.453	32.154
LeH			5.349	4.506	9.855
MTG		1	337	336	674
BewA	2		3.669	3.631	7.302
Gesamt	2	1	26.056	23.926	49.985

Genese der einzelnen Variablen in den Testdaten

Variable	Bezeichnung	Ersetzungsmaßnahme
	Identifikationsnummern	
persnr	Personennummer	zufällig ersetzt
satznr	Satznummer	zufällig ersetzt
masnr	Maßnahmenummer	zufällig ersetzt, Maßnahmewechsel werden richtig angezeigt, aber nicht, welche Personen in der selben Maßnahme ist
betnr	Betriebsnummer	zufällig ersetzt, Betriebswechsel werden richtig angezeigt, aber nicht, welche Personen im selben Betrieb arbeiten.
	Spellanfang und -ende	
begorig	Beginndatum Originalspell	Datumsangaben werden zum Teil zufällig ersetzt. Erhalten bleiben die Reihenfolge der Spells und das Beginn- und Endjahr. Vollständig erhalten bleiben Daten am 1.1. und 31.12. Die Struktur der Spelldauern entspricht dadurch nicht deren Struktur in den Originaldaten.
endorig	Endedatum Originalspell	
begepi	Beginndatum der gesplitteten Episode	
endepi	Endedatum der gesplitteten Episode	
	Generierte technische Merkmale	
quelle	Quelle	nicht verändert
kom_quel	Kombination der Quellen	wird neu berechnet
spell	Spellzähler pro Konto	wird neu berechnet
nspell	Anzahl der Spells pro Konto	wird neu berechnet
level2	Spellzähler pro Episode	wird neu berechnet
nlevel2	Anzahl der Spells pro Episode	wird neu berechnet
level1	Spellzähler pro Episode und Quelle	wird neu berechnet
nlevel1	Anzahl der Spells pro Episode und Quelle	wird neu berechnet
berknz	Bereinigungskennzeichen	Tauschvariante 1
stendat	Status des Ende-Datums	Tauschvariante 1
	Personenstatus vor, während und nach dem aktuellen Spell	
estatvor	Erwerbsstatus vor Arbeitsuche	Tauschvariante 1
krankvor	Fortsetzung der Arbeitslosigkeit nach Arbeitsunfähigkeit	Tauschvariante 1
erwstat	Erwerbsstatus: Personengruppe, Leistungsart, Maßnahmeart, Arbeitsuche-Status	nicht verändert
grund	Grund des Spellendes	nicht verändert
sna	Status nach Abgang	Tauschvariante 1
	Persönliche Merkmale	
gebjahr	Geburtsjahr	Zufälliger Tausch der Differenz zum Jahr des ersten Spells. Alle Personen sind zu Beginn des ersten Spells mind. 13 Jahre alt. Mehr Personen als in den Originaldaten haben ein unplausibel hohes Alter.
sex	Geschlecht	Tauschvariante 2

Variable	Bezeichnung	Ersetzungsmaßnahme
nation *nation_org*	Staatsangehörigkeit	Tauschvariante 2
schweb	Schwerbehindertenstatus	Tauschvariante 2
schbild	Schulabschluss	Tauschvariante 2
bild	Ausbildung	Tauschvariante 1
	Angaben zu Beschäftigungsverhältnis und Arbeitsuche	
stib	Stellung im Beruf und Arbeitszeit	Tauschvariante 1
beruf	Beruf	Tauschvariante 2
vstyp	Rentenversicherungsträger	Tauschvariante 1
beitgr	Beitragsgruppe	Tauschvariante 1
tentgelt	Tagesentgelt / täglicher Leistungssatz	Tauschvariante 1
gleitz	Gleitzone	Tauschvariante 1
w93	Wirtschaftszweig 93	Tauschvariante 1
begalo	Beginndatum der Arbeitslosigkeit	wird neu berechnet, für MTG-Spells unvollständig.
daualo	Dauer der Arbeitslosigkeit	
endplan	Geplantes Ende der Maßnahme	zufälliger Tausch der Differenz bis zum Ende des Originalspells. Im Gegensatz zu den Originaldaten liegt das geplante Ende in den Testdaten z.T. vor dem Beginn der Maßnahme.
	Ortsangaben	
ao_bula *ao_kreis* *ao_gemei*	Arbeitsort Bundesland (Kreis, Gemeinde)	Tauschvariante 2 Die Ziehung erfolgt jeweils unabhängig voneinander, so das die Ortsangaben nicht konsistent sind.
wo_bula *wo_kreis* *wo_gemei*	Wohnort Bundesland (Kreis, Gemeinde)	
ao_rd *ao_aa* *ao_gest*	Arbeitsort Regionaldirektion (Arbeitsagentur, Geschäftsstelle)	
wo_rd *wo_aa* *wo_gest*	Wohnort Regionaldirektion (Arbeitsagentur, Geschäftsstelle)	
wo_aatyp	Wohnort Regionaltyp der Arbeitsagentur	wird neu berechnet auf Basis von wo_aa

Sensible Merkmale, deren Nutzung im Antrag gesondert zu begründen ist, sind in den Testdaten enthalten (in der Tabelle *kursiv* gekennzeichnet).