

Anonymisierung der MLD-Daten für den für das GESIS-Datenarchiv bestimmten Datensatz**1. Hintergrund**

Voraussetzung für die Bereitstellung der Daten im GESIS-Datenarchiv ist, dass keine Rückschlüsse auf die Identität der Befragten möglich sind, dass also deren Anonymität gewährleistet ist. Zwar sind Name, Anschrift o.ä. personenbezogene Daten im Originaldatensatz nicht enthalten, allerdings sind seltene Nationalitäten vertreten. In Kombination mit Merkmalen wie Bundesland, Alter, Zusammensetzung des Haushalts sowie Alter der Haushaltsmitglieder kann daher nicht ausgeschlossen werden, dass in Einzelfällen eine Deanonymisierung von Personen möglich ist. Aus diesem Grund wurden in dem zur Datenarchivierung bestimmten Datensatz alle Variablen, die Länderangaben enthalten, umcodiert. Die Nennungen der konkreten Länder wurden durch die auch im Forschungsbericht 6 des BAMF verwendeten acht zusammenfassenden Kategorien ersetzt.

Tabelle 1: Übersicht über die unter den regionalen Herkunftskategorien berücksichtigten Länder

Südost-europa	Türkei	Zentralasi-en/GUS	Iran	Süd-/ Süd-ostasien	Naher Osten	Nord-afrika	Restliches Afrika
- Albanien - Bulgarien - Nachfolge-staaten des ehemaligen Jugoslawien (ohne Kroatien und Slowenien)	- Türkei	- Aserbaidshan - Kasachstan - Kirgisistan - Russische Föderation - Turkmenistan - Usbekistan	- Iran	- Afghanistan - Bangladesch - Indien - Indonesien - Malaysia - Pakistan	- Ägypten - Irak - Israel - Jemen - Jordanien - Libanon - Saudi Arabien - Syrien	- Algerien - Libyen - Marokko - Tunesien	- Äthiopien - Cote d'Ivoire - Eritrea - Gambia - Ghana - Guinea - Kamerun - Liberia - Mosambik - Nigeria - Senegal - Sierra Leone - Somalia - Sudan - Togo

2. Variablentransformation im zur Weitergabe bestimmten Datensatz

In einem ersten Schritt wurde aus den relevanten Angaben Variablen über den Hauptmigrationshintergrund des Haushalts sowie des Befragten gebildet. Hiermit soll vermieden werden, dass durch die in einem zweiten Schritt erfolgte Zusammenfassung von Ländervariablen zu Regionalvariablen und dem damit verbundenen Informationsverlust Unklarheiten entstehen.

Bei der Umsetzung der Ländervariablen in Regionalvariablen (einschließlich der selbst gebildeten Variablen zum Migrationshintergrund) wurden folgende Kategorien berücksichtigt: die acht im Forschungsbericht ausgewiesenen Herkunftsregionen (s. Tabelle 1), ein anderer (nicht relevanter) Migrationshintergrund sowie kein Migrationshintergrund.

Bei Abfragen mit möglichen Mehrfachantworten (z.B. Staatsangehörigkeiten), bei denen die Länder durch Dummy-Variablen erfasst wurden, wurden neue Dummy-Variablen für die Regionen geschaffen und besetzt. Bei Variablen, bei denen die Länder durch Kategorien erfasst wurden, wurden diese in einer neuen Variable umcodiert. Dies war etwa bei den Geburtslandvariablen der Fall, bei denen keine Mehrfachantworten möglich sind. Alle Variablen mit differenzierten Ländernennungen wurden anschließend gelöscht (s. hierzu Tabelle 2).

Einige weitere sehr differenzierte Variablen wurden aus dem zur Weitergabe bestimmten Datensatz ersatzlos gelöscht, so etwa offene Angaben über die Interviewsprache, der Regierungsbezirk o.ä.

Um folgende selbst gebildete Variablen wurde der Datensatz ergänzt:

- Migrationshintergrund des Befragten / des Haushalts, recodiert nach Regionen,
- Generationenzugehörigkeit des Befragten,
- Zahl der genannten (ehemaligen) Staatsangehörigkeiten (als Ersatz für den diesbezüglichen Informationsverlust durch die Löschung ehemals differenzierterer Dummy-Variablen).

Zu den vorgenommen Transformationen siehe auch Tabelle 2.

Tabelle 2: Übersicht über die transformierten und umcodierte Variablen:

alte (gelöschte) Variablen	neue Variablen	Inhalt
transformierte Dummy-Variablen bei Ländernennungen (Länder werden Regionen zugeordnet, offene Angaben gelöscht)		
S1_11 – S1_97o	s1_101 – s1_199	Screening, Geburtsland
S2_11 – S2_97o	s2_101 – s2_199	Screening, Geburtsland Eltern
S3_11 – S3_97o	s3_101 – s3_199	Screening, Staatsangehörigkeit
v7_1 – v7_97o	v7_101 – v9_199	Staatsangehörigkeit(en) Befragter
v9_11 – v9_97o	v9_101 – v9_199	frühere Staatsangehörigkeit(en), Befragter
v77_1 – v77_97o	v77_101 – v77_199	Staatsangehörigkeit(en) Partner
v78_1 – v78_97o	v78_100 – v78_199	frühere Staatsangehörigkeit(en) Partner
v103p1_1 – v103p1_97o.	103p1_101 – v103p1_199	Staatsangehörigkeit(en) Person 1
v103p1_2 – v103p2_97o.	103p2_101 – v103p2_199	Staatsangehörigkeit(en) Person 2
v103p1_3 – v103p3_97o.	103p3_101 – v103p3_199	Staatsangehörigkeit(en) Person 3
v103p1_4 – v103p3_97o.	103p4_101 – v103p4_199	Staatsangehörigkeit(en) Person 4
v103p1_5 – v103p5_97o.	103p5_101 – v103p5_199	Staatsangehörigkeit(en) Person 5
v103p1_6 – v103p6_97o.	103p6_101 – v103p6_199	Staatsangehörigkeit(en) Person 6
v103p1_7 – v103p7_97o.	103p7_101 – v103p7_199	Staatsangehörigkeit(en) Person 7
v103p1_8 – v103p8_97o.	103p8_101 – v103p8_199	Staatsangehörigkeit(en) Person 8
v103p1_9 – v103p9_97o.	103p9_101 – v103p9_199	Staatsangehörigkeit(en) Person 9
v103p1_10 – v103p10_99.	103p10_101 – v103p10_199	Staatsangehörigkeit(en) Person 10
v103p1_11 – v103p11_99.	103p11_101 – v103p11_199	Staatsangehörigkeit(en) Person 11
v103p1_12 – v103p12_99.	103p12_101 – v103p12_199	Staatsangehörigkeit(en) Person 12
v103p1_13 – v103p13_99.	103p13_101 – v103p13_199	Staatsangehörigkeit(en) Person 13
v103p1_14 – v103p14_99.	103p14_101 – v103p14_199	Staatsangehörigkeit(en) Person 14
v103p1_15 – v103p15_99.	103p15_101 – v103p15_199	Staatsangehörigkeit(en) Person 15
v103p1_16 – v103p16_99.	103p16_101 – v103p16_199	Staatsangehörigkeit(en) Person 16
v103p1_17 – v103p17_99.	103p17_101 – v103p17_199	Staatsangehörigkeit(en) Person 17
v103p1_18 – v103p18_99.	103p18_101 – v103p18_199	Staatsangehörigkeit(en) Person 18
v103p1_19 – v103p19_99.	103p19_101 – v103p19_199	Staatsangehörigkeit(en) Person 19
umcodierte Ländervariablen (Länder werden in neuer Variable Regionen zugeordnet, alte Ländervariablen und offene Angaben gelöscht)		
v2, v2o, v5, v5o, v6, v6o	v2reg, v5reg, v6reg	Geburtsland Befragter, Mutter, Vater
v76, v76o	v76reg	Geburtsland Partner
ersatzlos gelöschte Variablen		
migr_11m – migr_62m		von Infratest gebildete Multi-Variablen über MigHg
erwsz_s		von Infratest gebildete Variablen mit Ländernennungen
regbez		Von Infratest zugespielte Variable über Regierungsbezirk
v110_9o		offene Nennungen für Interviewsprache
neu gebildete Variablen		
	s1_sum, s2_sum, s3_sum	Zahl der Ländernennungen in Screeningfragen
	v7_sum, v9_sum	Zahl der Ländernennungen bei (ehemaliger) Staatsangehörigkeit des Befragten
	V77_sum, v78_sum	Zahl der Ländernennungen bei (ehemaliger) Staatsangehörigkeit des Partners
	v103p1_sum – v103p19_sum	Zahl der Staatsangehörigkeiten von Person 1 – 19
	gen1	Generationenzugehörigkeit (4 Ausprägungen)
	region	Hauptherkunftsregion des Haushalts
	region_b, region_b1, region_b2	Hauptherkunftsregion des Befragten