

1190²⁰²²

SOEP Survey Papers
Series C - Data Documentations (Datendokumentationen)

Gewichtung der CoV-Studie der IAB-BAMF-SOEP-Befragung von Geflüchteten

Ehsan Vallizadeh und Stefan Zins

Running since 1984, the German Socio-Economic Panel (SOEP) is a wide-ranging representative longitudinal study of private households, located at the German Institute for Economic Research, DIW Berlin.

The aim of the SOEP Survey Papers Series is to thoroughly document the survey's data collection and data processing.

The SOEP Survey Papers is comprised of the following series:

- Series A** – Survey Instruments (Erhebungsinstrumente)
- Series B** – Survey Reports (Methodenberichte)
- Series C** – Data Documentation (Datendokumentationen)
- Series D** – Variable Descriptions and Coding
- Series E** – SOEPmonitors
- Series F** – SOEP Newsletters
- Series G** – General Issues and Teaching Materials

The SOEP Survey Papers are available at <http://www.diw.de/soepsurveypapers>

Editors:

Dr. Jan Goebel, DIW Berlin
Prof. Dr. Stefan Liebig, DIW Berlin and Freie Universität Berlin
Prof. Dr. David Richter, DIW Berlin and Freie Universität Berlin
Prof. Dr. Carsten Schröder, DIW Berlin and Freie Universität Berlin
Prof. Dr. Jürgen Schupp, DIW Berlin and Freie Universität Berlin
Prof. Dr. Sabine Zinn, DIW Berlin and Humboldt Universität zu Berlin

Please cite this paper as follows:

Ehsan Vallizadeh und Stefan Zins. 2022. Gewichtung der CoV-Studie der IAB-BAMF-SOEP-Befragung von Geflüchteten. SOEP Survey Papers 1190: Series C. Berlin: DIW/SOEP



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
© 2022 by SOEP

ISSN: 2193-5580 (online)

DIW Berlin
German Socio-Economic Panel (SOEP)
Mohrenstr. 58
10117 Berlin
Germany

soeppapers@diw.de

Gewichtung der CoV-Studie der IAB-BAMF-SOEP-Befragung von Geflüchteten

No. 1190: Series C – Data Documentation

Ehsan Vallizadeh und Stefan Zins

Übersicht

Die IAB-BAMF-SOEP Befragung von Geflüchteten ist eine Längsschnittstudie, die seit 2016 jährlich durchgeführt wird. Befragt werden Geflüchtete (Erwachsene und Minderjährige) unabhängig von ihrem Aufenthaltsstatus, die vom 1. Januar 2013 bis 31. Januar 2016 nach Deutschland eingereist und einen formellen Asylantrag beim BAMF gestellt oder im Rahmen von speziellen Programmen des Bundes oder der Länder als Schutzsuchende aufgenommen wurden. Die IAB-BAMF-SOEP-Befragung von Geflüchteten bietet somit eine umfassende und repräsentative Datengrundlage über die in Deutschland lebenden Geflüchteten (Brücker, Rother, and Schupp 2017). Im Frühjahr 2020 wurden die Haushalte neben der regulären Panelbefragung zusätzlich telefonisch (d.h. in einem CATI-Modus) zu ihren Erfahrungen und zu den Veränderungen ihres Alltags durch die Corona-Krise befragt. Der Fragebogen der CoV-Studie von Geflüchteten (im Folgenden REF-CoV-Studie) wurde so angelegt, dass ein Vergleich mit der ebenfalls im Jahr 2020 gesondert durchgeführten SOEP-CoV-Studie unter der Verantwortung des SOEP am DIW Berlin auch möglich ist (Kühne et al. 2020). Die Befragung umfasst gut 50 Fragen und deckt eine Reihe von Themen ab, wie etwa Änderungen der Arbeitssituation, der persönlichen Gesundheit, der Kinderbetreuung, des sozialen Lebens sowie Lebenszufriedenheit, Einstellung, Sorgen u.v.m. Um den Auswirkungen der Corona-Pandemie auf die besondere Lebenssituation der Geflüchteten Rechnung zu tragen, wurde der Fragebogen um andere wichtige Themen, wie etwa Teilnahme an Integrationskursen und anderen Sprachkursen sowie an Bildungsmaßnahmen, erweitert. Die Feldorganisation und die Durchführung der REF-CoV-Studie wurde vom Erhebungsinstitut Kantar Public verantwortet.

Stichprobe

Die Gewichtung der Studie verlief grundsätzlich anders als die Gewichtung der SOEP-CoV-Studie (Siegers, Steinhauer, and Zinn 2020). Es wurden keine Unterscheidung zwischen Haushalts- und Personengewichte vorgenommen (Es wurde eine Person pro Haushalt befragt.). Zudem wurde die Gewichtung nicht mit der von Vorgängerbefragungen kombiniert, wie dies sonst bei der Gewichtung von Daten des SOEP üblich ist. Die Gewichtung beinhaltet zwei Schritte:

1. Drei aufeinanderfolgende Ausfallschritte, um die Teilnahmewahrscheinlichkeit an der Befragung zu modellieren.
2. Die im ersten Schritt berechneten Propensity Score-Gewichte wurden an verschiedenen Populationsverteilungen der Geflüchteten (Jahr) randangepasst.

Die drei modellierten Ausfallschritte von der Bruttostichprobe zur Nettostichprobe lassen sich wie folgt beschreiben:

1. Vorliegen einer gültigen Telefonnummer.
2. Durchführung der letzten persönlichen Interviews entweder auf Englisch, Arabisch, Farsi oder Deutsch.

Eine systematische Auswahl der Befragungsperson im jeweiligen Haushalt erfolgte nicht. Auch wenn eine Präferenz für den Haushaltsvorstand bestand, konnte jede verfügbare erwachsene Person (ab 18 Jahren) des Haushalts teilnehmen. Unter den Haushalten, welche die Zugangskriterien erfüllt haben, gab es 324 Haushalte, die dennoch nicht Teil der Bruttostichprobe wurden, da sie entweder

- an der Befragungswelle 2019 teilgenommen haben, aber eine weitere Befragung abgelehnt wurde
- oder generell eine weitere Teilnahme verweigert haben
- oder aufgrund von anderen Ausfallgründen, wie z.B. Auslandsverzug, für eine weitere Befragung im Jahr 2020 nicht in Frage kamen.

Feldarbeit

Zur Verbesserung der Teilnahmequote und zur Vorbereitung der Haushalte auf den Anruf wurde ein zweisprachiges, postalisches Anschreiben an die Zielhaushalte verschickt. Die Feldzeit war vom 7. Juli 2020 bis zum 24. August 2020.

Ablauf der Gewichtung

Die Gewichtung der Studie verlief grundsätzlich anders als die Gewichtung der SOEP-CoV-Studie (Siegers, Steinhauer, and Zinn 2020). Es wurden keine Unterscheidung zwischen Haushalts- und Personengewichte vorgenommen (Es wurde eine Person pro Haushalt befragt.). Zudem wurde die Gewichtung nicht mit der von Vorgängerbefragungen kombiniert, wie dies sonst bei der Gewichtung von Daten des SOEP üblich ist. Die Gewichtung beinhaltet zwei Schritte: 1. Drei aufeinanderfolgende Ausfallschritte, um die Teilnahmewahrscheinlichkeit an der Befragung zu modellieren. 2. Die im ersten Schritt berechneten Propensity Score Gewichte wurden an verschiedene Populationsverteilungen der Geflüchteten (Jahr) randangepasst. Die drei modelliert Ausfallschritte von der Bruttostichprobe zur Nettostichprobe lassen sich wie folgt beschreiben

- Schritt 1: Ausgehend von den Haushalten der IAB-BAMF-SOEP-Befragung von Geflüchteten im Jahr 2019, insgesamt 4.133 Haushalte, wurde die Wahrscheinlichkeit dafür modelliert Teil der Auswahlgrundlage für die IAB-BAMF-SOEP-Befragung Geflüchteter im Jahr 2020 zu sein, modelliert.
- Schritt 2: Ausgehend von der Auswahlgrundlage wurde die Wahrscheinlichkeit Teil der Bruttostichprobe, von 2.563 Fällen zu sein, modelliert.
- Schritt 3: Ausgehend von der Bruttostichprobe wurde die Wahrscheinlichkeit Teil der Nettostichprobe, von 1.439 Fällen zu sein, modelliert.

Die folgende Tabelle stellt die Fallzahlen der drei Ausfallschritte dar:

Tabelle 1: Ausfallschritte

Welle 2019	Auswahlgrundlage	Brutto	Netto	Personen
1	0	0	0	4.133
1	1	0	0	3.060
1	1	1	0	2.563
1	1	1	1	1.439

Die Inverse des Produkts der geschätzten Wahrscheinlichkeiten der drei Ausfallschritte, für alle Fälle der Nettostichprobe, bildet die sog. Propensity Score-Gewichte. Diese Propensity Score-Gewichte wurden danach mittels eines linearen Kalibrierungsverfahrens (Deville and Särndal 1992) an die folgenden Randverteilungen angepasst.

Tabelle 2: Alter-Geschlechts Verteilung

Alter	Männer	Frauen
18-24	137.483	44.864
25-34	221.169	84.994
35-44	109.246	57.057
45-54	43.776	24.783
55+	20.914	17.389

Tabelle 3: Bundesländer

Bundesland	Personen
BB_BE	56.120
HH_SH	55.435
HB_NI	92.714
NW	201.479
HE	65.983
SL_RP	47.585
BW	86.934
ST	16.719
BY	90.478

Bundesland	Personen
MV	9.263
TH	14.386
SN	24.579

Tabelle 4: Herkunftsländer

Herkunftsland	Personen
SYR	330.295
AFG	113.730
IRK	76.587
ALB_SRB_KOS	17.409
ERI_SOM	53.934
IRN_PAK_RUS	50.607
OTHER	119.113

Tabelle 5: Ankunftsjahre

Ankunftsjahr	Personen
2013	48.005
2014	104.335
2015	479.125
2016	130.210

Schätzung der Propensity Scores

Zur Schätzung der Wahrscheinlichkeit p der oben beschriebenen Ausfallschritte wird eine generalisierte, additive Regression mit einer logit-Linkfunktion verwendet. Das verwendete Modell hat die folgende Form:

$$g(p) = \beta_0 + X\beta + \sum_j f_j(Z_j)$$

dabei ist $g = \log\left(\frac{p}{1-p}\right)$ (die logit-Funktion), β_0 ein Skalar, X eine Matrix mit fast ausschließlich dichotomen Variablen, β ein Vektor mit Regressionskoeffizienten, Z_j eine

stetige Variable und f_j eine glatte Funktion, deren Parametrisierung spezifiziert werden kann. Die Variablen, die zur Modellierung der Wahrscheinlichkeiten in Betracht gezogen wurden, sind die folgenden:

Tabelle 6: Variablen zur Modellierung der Ausfallschritte

Variable	Datensatz	Variablenlabel
bula	M345_CoV_Brutto_Gewichtung	Bundesland lt. Statistik
bik	M345_CoV_Brutto_Gewichtung	Gemeindetyp nach BIK-Stadtregionen Stand 31.12.2017
ggk	M345_CoV_Brutto_Gewichtung	Gemeindegrößenklasse pol. Stand 31.12.2017
gtyp	M345_CoV_Brutto_Gewichtung	Siedlungsstrukt. Gemeindetyp BFLR Stand 31.12.2008
hhgr	M345_CoV_Brutto_Gewichtung	Aktuelle Haushaltsgröße
hspv	M345_CoV_Brutto_Gewichtung	Sprache
bjwein	bjhbrutto_2020-06-04	Jahr seitdem HH an dieser Adresse geführt wird
bjhtyp	bjhbrutto_2020-06-04	Haushaltstyp
bjintk	bjhbrutto_2020-06-04	Anzahl der Interviewerkontakte
bjsplit	bjhbrutto_2020-06-04	Startbearbeitungsform
bjhfors	bjhbrutto_2020-06-04	Bearbeitungsform Schluß
bjhergs	bjhbrutto_2020-06-04	Bearbeitungsergebnis Schluß
bjudi	bjhbrutto_2020-06-04	Unsachgemäß durchgeführte Interviews
bjhhgr	bjhbrutto_2020-06-04	Aktuelle Haushaltsgröße
bjbefhpmax	bjhbrutto_2020-06-04	Befragungspersonen
bjpteil	bjhbrutto_2020-06-04	Teilnehmer im Haushalt
bjincentive1	bjhbrutto_2020-06-04	Incentivierungsart
bjincentive2	bjhbrutto_2020-06-04	Incentivierungsmodell
bjdat	M345_CoV_Contact_Log/bjhbrutto_2020-06-04	Tage von letztem Kontaktzeit bis Feldzeitbegin
fest_mobil	M345_CoV_Contact_Log	Festnetznummer oder Mobilfunknummer

Variable	Datensatz	Variablenlabel
NumbersOf Contacts	M345_CoV_CATIErg_Brutto	Anzahl der Kontaktversuche

Hinweis: Die IAB-BAMF-SOEP-Befragung von Geflüchteten teilt sich in verschiedene Unterstichproben. Für die REF-CoV-Studie wurden die Stichproben M3, M4 und M5 berücksichtigt. Für eine detaillierte Beschreibung der Stichproben siehe Brücker, Rother, and Schupp (2017). Die Variablen *fest_mobil* und *NumbersOfContact* wurden nur zur Modellierung des dritten Ausfallschritts verwendet.

Mit Hilfe eines LASSO-Verfahrens wurden für die Modelle der jeweiligen Ausfallschritte, eine optimale Teilmenge der oben gelisteten Variablen als Prädiktoren ausgewählt (Chouldechova and Hastie 2015).

Verteilung der kalibrierten Gewichte

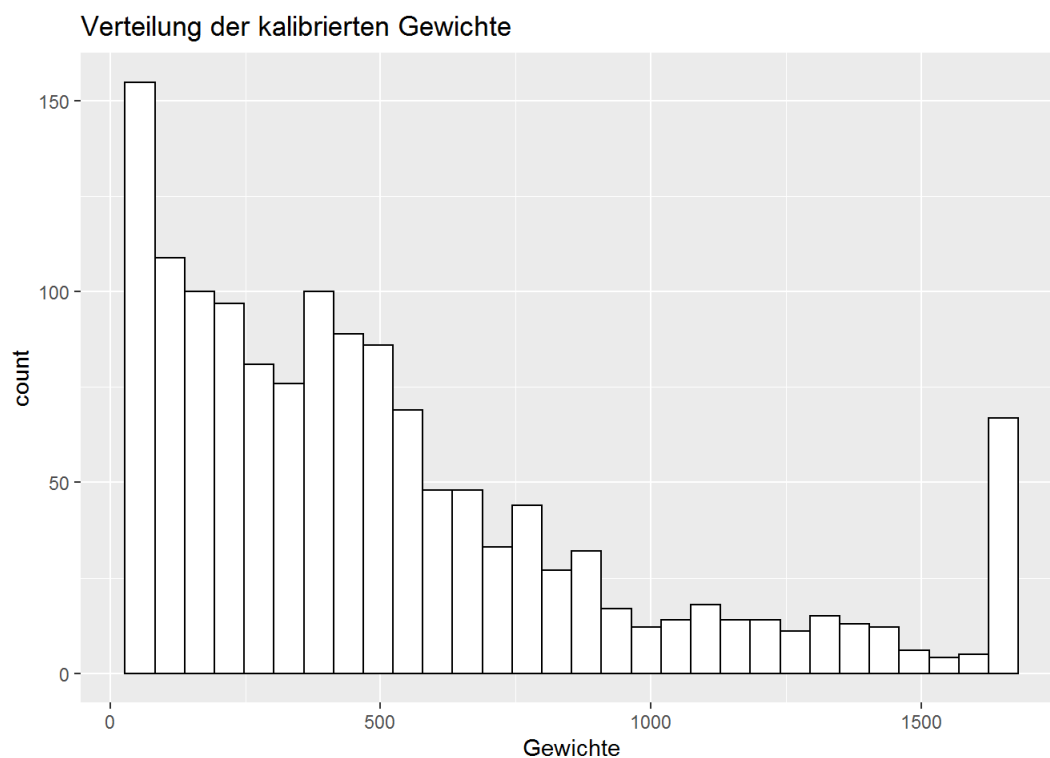


Tabelle 7: Lagemaße der kalibrierten Gewichte

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
40,94606	187,0883	409,6989	514,7362	691,5839	1.637,842

Literatur

Brücker, Herbert, Nina Rother, and Jürgen Schupp. 2017. "IAB-BAMF-SOEP-Befragung von Geflüchteten 2016 * Studiendesign, Feldergebnisse Sowie Analysen Zu Schulischer Wie Beruflicher Qualifikation, Sprachkenntnissen Sowie Kognitiven Potenzialen." IAB.

Chouldechova, Alexandra, and Trevor Hastie. 2015. "Generalized Additive Model Selection." arXiv. <https://doi.org/10.48550/ARXIV.1506.03850>.

Deville, Jean-Claude, and Carl-Erik Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87 (418): 376–82.

Kühne, Simon, Martin Kroh, Stefan Liebig, and Sabine Zinn. 2020. "The Need for Household Panel Surveys in Times of Crisis: The Case of SOEP-CoV" 14 (2): 195–203.

Siegers, Rainer, Hans Walter Steinhauer, and Sabine Zinn. 2020. "Gewichtung Der SOEP-CoV-Studie, SOEP Survey Papers No. 888, Series c – Data Documentation." DIW.