Research Institute
Social Cohesion

RDC

SOEP

SCP Documentation

German Social Cohesion Panel

SCP 2021-23 W1-3 Codebook HHRF: Weights for Households (German)



Research Institute Social Cohesion

SOEP The Socio-Economic Panel

German Social Cohesion Panel

Established in 2021, the German Social Cohesion Panel (SCP) is a wide-ranging representative longitudinal study of private households in Germany, carried out in collaboration of the Research Institute Social Cohesion (RISC) and the German Socio-Economic Panel (SOEP).

The aim of the SCP Documentation is to thoroughly document the survey's data collection and data processing.

Recommended Citation

Groh-Samberg, O., Axenfeld, J. B., Gerlitz, J.-Y., Cornesse, C., Kroh, M., Lengfeld, H., Reinecke, J., Teichler, N., Traunmüller, R., & Zinn, S. (2026). SCP 2021-23 W1-3 - Codebook HHRF: Weights for Households (German). *German Social Cohesion Panel 2021-2023 - Wave 1-3*. Bremen and Berlin: RDC-RISC/SOCIUM, SOEP/DIW Berlin. doi:10.60532/scp.2021-23.w1-3.v1

- **Authors:** Olaf Groh-Samberg, Julian B. Axenfeld, Jean-Yves Gerlitz, Carina Cornesse, Martin Kroh, Holger Lengfeld, Jost Reinecke, Nils Teichler, Richard Traunmüller, Sabine Zinn
- **Contributors:** Cosima Adams, Anton Bochert, Martin Gerike, Josefine Kuhrmeier, Anna-Tabea Müller, Eric Nissen, Sebastian Rueda-Uribe, Rainer Siegers, Hans Walter Steinhauer, Knut Wenzig, Julia Witton (Project Members), infas (Data Collector)
- **Publisher:** RDC-RISC
  SOCIUM, University of Bremen
  P.O. Box 330 440
  28334 Bremen
  Germany

  SOEP
  DIW Berlin
  German Socio-Economic Panel (SOEP)
  Anton-Wilhelm-Amo-Str. 58
  10117 Berlin
  Germany

- **DOI:** 10.60532/scp.2021-23.w1-3.v1
- **Website:** www.fgz-risc-data.de

  www.diw.de

The text of this publication is published under the Creative Commons license CC BY-SA 4.0 Attribution-ShareAlike 4.0 International. The exact wording of the license CC BY-SA 4.0 can be found here:

https://creativecommons.org/licenses/by-sa/4.0/

SCP Documentation

German Social Cohesion Panel

SCP 2021-23 W1-3 Codebook HHRF: Weights for Households (German)

Inhaltsverzeichnis

1 Allgemeine Informationen 2

2 Identifikatoren 2
hid – Aktuelle Haushalts-ID . . . 2
cid – Ursprüngliche Haushalts-ID . . . 3

3 Befragungskontext 3
wave – Erhebungswelle . . . 3

4 Statistische Gewichtungsfaktoren 4
design – Inverse Ziehungswahrscheinlichkeit . . . 4
hhrf – Hochrechnungsfaktor . . . 5

5 Inverse Bleibewahrscheinlichkeit 6
hbleib – Inverse Bleibewahrscheinlichkeit . . . 6

1 Allgemeine Informationen

Der HHRF-Datensatz enthält Gewichtungsfaktoren für die Haushalte im SCP. Jeder Haushalt (HID), der in einer bestimmten Erhebungswelle (WAVE) auf den Haushaltsfragebogen geantwortet hat, hat eine Zeile im Datensatz. Haushalte, deren Ankerperson nicht an der Erhebungswelle teilgenommen hat oder die Befragung vor Beginn des Haushaltsfragebogens abgebrochen hat, sind nicht im Datensatz enthalten.
An einigen Stellen in der Dokumentation und in den Daten werden Jahreszahlen z. B. für die Bezeichnung von Variablen und des Fragebogeninstruments verwendet. Diese Jahreszahlen orientieren sich stets am Feldstart der Datenerhebung der entsprechenden Erhebungswelle.

2 Identifikatoren

hid – Aktuelle Haushalts-ID

21100003		3
21100009		4
21100010		1
21100012		1
21100016		4
21100020		4
21100021		4
21100032		1
21100037		1
21100039		4
21100044		1
21100045		1
21100049		4
21100050		1
21100058		1
...	(13071 Zeilen unterdrückt)	31362
22250034		1
22250037		1
22250038		1
22250039		1
22250040		1
22250041		1
22250043		1
22250044		1
22250045		1
22250046		1
22250047		1
22250048		1
22250049		1
22250050		1
22250051		1

Diese Kennziffer gruppiert Individuen in ihre zugehörigen Haushalte zum Zeitpunkt der aktuellen Erhebungswelle. Das bedeutet, dass die HID einer Person sich über die Zeit verän-

dern kann, zum Beispiel wenn ein erwachsenes Kind aus dem elterlichen Haushalt auszieht und einen eigenen Haushalt eröffnet. Die HID gibt dabei den Stand vor Feldstart wieder. Änderungen in der Haushaltszugehörigkeit werden erst in der Folgewelle in der HID erfasst.

cid – Ursprüngliche Haushalts-ID

21100003		3
21100009		4
21100010		1
21100012		1
21100016		4
21100020		4
21100021		4
21100032		1
21100037		1
21100039		4
21100044		1
21100045		1
21100049		4
21100050		1
21100058		1
...	(13023 Zeilen unterdrückt)	31347
21137960		1
21137961		1
21137963		4
21137964		3
21137967		3
21137971		1
21137972		1
21137973		1
21137976		1
21137978		2
21137979		4
21137985		2
21137987		1
21137991		4
21138000		1

Diese Kennziffer gruppiert Individuen in ihre Ursprungshaushalte zu Beginn des Panels. Das bedeutet, dass die CID einer Person zeitkonstant gehalten wird und sie immer mit dem Haushalt verbinden wird, zu dem sie initial gehört hat, selbst wenn sie seitdem den Haushalt gewechselt hat.

3 Befragungskontext

wave – Erhebungswelle

1	[1] Welle 1, Teil 1 (2021/22)	13053
2	[2] Welle 1, Teil 2 (2021/22)	6669
3	[3] Welle 2 (2022/23)	6128
4	[4] Welle 3 (2023)	5562

Diese Variable identifiziert die (Teil-)Welle, in der die Datenerhebung stattgefunden hat.

4 Statistische Gewichtungsfaktoren

design – Inverse Ziehungswahrscheinlichkeit

0	5
70.1461868286133	3
113.673835754395	1
129.841339111328	10
151.398025512695	15
177.21418762207	4
181.577438354492	40
202.015487670898	4
209.343887329102	1
226.846572875977	185
230.22819519043	13
230.803298950195	4
255.753479003906	9
287.660064697266	5
302.295227050781	1000
322.924377441406	6
328.682861328125	13
383.379913330078	49
403.530364990234	55
453.192596435547	6204
459.955841064453	187
537.873657226562	110
574.819702148438	999
766.259521484375	2425
806.560424804688	787
905.884948730469	2402
1149.13916015625	11883
1612.62072753906	481
2297.7783203125	4512

Diese Variable enthält die inversen Ziehungswahrscheinlichkeiten (Design-Gewichte) für die SCP-Stichprobe. Das Design-Gewicht berücksichtigt die aus dem Stichprobenziehungsdesign resultierenden ungleichen Ziehungswahrscheinlichkeiten.
Das SCP hat ein zweistufiges Stichprobenziehungsverfahren. Auf der ersten Stufe werden, stratifiziert nach Region und Urbanitätsgrad, Gemeinden gezogen (primary sampling units; PSUs). Auf der zweiten Stufe werden Personen (secondary sampling units; SSUs) aus den

Registern dieser Gemeinden gezogen. Generell erfolgte die Stichprobenziehung proportional zur Gemeindegröße. Eine Ausnahme ist die beabsichtigte überproportionale Ziehung in Ostdeutschland. Alle in die Stichprobe gezogenen Personen, die an der Befragung teilnehmen, wurden gebeten ihre weiteren volljährigen Haushaltsmitglieder anzugeben. Diese weiteren Haushaltsmitglieder werden dann ebenfalls befragt. Das führt zu einer höheren Ziehungswahrscheinlichkeit für größere Haushalte, was ebenso durch die Design-Gewichte berücksichtigt wird.
Aufgrund der Rundung von Nachkommastellen kann es im Codebuch zu einer Zusammenfassung von Werten kommen.

hhrf – Hochrechnungsfaktor

89.9704132080078		1
102.496826171875		1
103.33854675293		1
114.508514404297		1
117.915008544922		1
121.074142456055		1
121.479843139648		1
125.444351196289		1
132.868362426758		1
137.444686889648		1
138.32536315918		1
149.049224853516		1
150.570892333984		1
156.400650024414		1
156.875350952148		1
...	(31373 Zeilen unterdrückt)	31382
45724.4709507794		1
45834.347811575		1
45859.7743569513		1
46253.0062740231		1
46675.6061271224		1
46891.2510077426		1
47071.5742033779		1
47231.6730588036		1
48348.0996327961		1
49844.7931102933		1
49918.8162558279		1
50033.0723977376		1
50162.1334203015		1
50745.6955823896		1
53255.1817071805		1

Diese Variable enthält die Haushalts-Nonresponse-Gewichte für das SCP, die zur Reduzierung von Verzerrungen durch Unit-Nonresponse dienen. Dieser Gewichtungsfaktor ist eine Kombination aus inverser Stichprobenziehungswahrscheinlichkeit, einem Nonresponse-Adjustierungsfaktor und einer Extrapolation zur Zielpopulation der Befragung.

Die inverse Stichprobenziehungswahrscheinlichkeit (siehe DESIGN-Variable) korrigiert für die ungleichen Ziehungswahrscheinlichkeiten in die Panel-Stichprobe (z.B. durch das beabsichtigte Über-Ziehen von Ostdeutschen).
Der Nonresponse-Adjustierungsfaktor korrigiert für Unit-Nonresponse. Dafür wurden Teilnahmewahrscheinlichkeiten basierend auf einem logistischen Regressionsmodell geschätzt. Zur Schätzung der Teilnahmewahrscheinlichkeit wurden Daten aus dem Sampling-Frame (Altersgruppen, Geschlecht, deutsche Staatsbürgerschaft, Bundesländer) und mikrogeographische Daten als Prädiktoren berücksichtigt. Fehlende Werte in den Prädiktoren wurden mittels multipler Imputation vervollständigt. Prädiktoren wurden durch eine Mischung aus Backward- und Forward-Selection unter Verwendung des Kreuzvalidierungs-Mean-Squared-Errors als Selektionskriterium ausgewählt.
Die Extrapolation basiert auf iterativem proportionalem Fitting (auch als "Raking" bezeichnet) mittels Mikrozensus-Daten zur demographischen Zusammensetzung (Bundesland, Gemeindegröße, Haushaltsgröße) der deutschen Bevölkerung.
Die Gewichte für Erhebungswellen ab Welle 1 Teil 2 wurden durch Multiplikation des initialen Haushalts-Nonresponse-Gewichts der Rekrutierung mit der inversen Teilnahmewahrscheinlichkeit der Ankerperson (AP) an der entsprechenden Erhebungswelle generiert. In Welle 1 Teil 2 wurde diese mittels logistischer Regression geschätzt. Die Prädiktorvariablen umfassen hier zusätzlich Umfragedaten der Vorwellen sowie Interaktionsterme für alle Variablen mit dem Befragtentyp (AP vs. Haushaltsmitglieder). (Dieses Modell zur Schätzung der Bleibewahrscheinlichkeit ist dasselbe wie das Modell, das in phrf zur Schätzung der Bleibewahrscheinlichkeit einzelner Befragter verwendet wurde.) In den folgenden Wellen (ab Welle 2) wurde die Teilnahmewahrscheinlichkeit der AP multiplikativ auf der Grundlage von zwei Modellen geschätzt: (1) der Wahrscheinlichkeit, dass ein Haushalt noch zu dieser Umfragewelle eingeladen wird, und (b) der Wahrscheinlichkeit, dass die AP bei Einladung teilnimmt. Wie in Vorwellen wurde Multiple Imputation angewandt, um fehlende Daten zu vervollständigen, und relevante Prädiktorvariablen wurden via Backward- und Forward-Selection ausgewählt. Anschließend wurden die Gewichte unter Verwendung von Mikrozensus-Informationen erneut geraket.
Aufgrund der Rundung von Nachkommastellen kann es im Codebuch zu einer Zusammenfassung von Werten kommen.

5 Inverse Bleibewahrscheinlichkeit

hbleib – Inverse Bleibewahrscheinlichkeit

0	14042
1.03273797035217	1
1.03302705287933	1
1.03336811065674	1
1.03418147563934	1
1.0357654094696	1
1.03591632843018	1
1.03751683235168	1
1.03825700283051	1
1.0383415222168	1
1.03835797309875	1
1.03866636753082	1
1.03944218158722	1

1.03976380825043		1
1.03996884822845		1
...	(14942 Zeilen unterdrückt)	17341
5.29517936706543		1
5.31301116943359		1
5.42971420288086		1
5.43652248382568		1
5.45314073562622		1
5.51391649246216		1
5.60983037948608		1
5.65867233276367		1
5.72668218612671		1
5.77419185638428		1
5.92092227935791		1
6.20641374588013		1
6.32693386077881		1
6.40488052368164		1
7.49871492385864		1

Diese Variable enthält die inverse Bleibewahrscheinlichkeit des Haushalts in den Wellen nach der Rekrutierung entsprechend einer Modellierung der Bleibewahrscheinlichkeit der Ankerperson (AP) mittels logistischer Regression. Die Prädiktorvariablen umfassen Erhebungsdaten aus früheren Wellen, einschließlich Interaktionsterme mit dem Befragungstyp (AP vs. Haushaltsmitglied). Fehlende Werte in den Prädiktoren wurden mittels multipler Imputation vervollständigt. Prädiktoren wurden durch eine Mischung aus Backward- und Forward-Selection unter Verwendung des Kreuzvalidierungs-Mean-Squared-Errors als Selektionskriterium ausgewählt.