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

FDZ-DATENREPORT

Dokumentation zu Arbeitsmarktdaten

04|2021 DE

Dokumentation und Codebuch für das Hochfrequente Online Personen Panel "Leben und Erwerbstätigkeit in Zeiten von Corona" (IAB-HOPP, Welle 1–7)

Marieke Volkert, Georg-Christoph Haas, Stefan Zins, Lisa Bellmann, Sandra Dummert, Sophie Hensgen, Johannes Ludsteck, Bettina Müller, Dana Müller, Christopher Osiander, Julia Schmidtke, Gesine Stephan, Annette Trahms

Dokumentation und Codebuch für das Hochfrequente Online Personen Panel ”Leben und Erwerbstätigkeit in Zeiten von Corona” (IAB-HOPP, Welle 1–7)

Marieke Volkert (Institut für Arbeitsmarkt- und Berufsforschung)
Georg-Christoph Haas (Institut für Arbeitsmarkt- und Berufsforschung; Universität Mannheim)
Stefan Zins (Institut für Arbeitsmarkt- und Berufsforschung)
Lisa Bellmann (Institut für Arbeitsmarkt- und Berufsforschung)
Sandra Dummert (Institut für Arbeitsmarkt- und Berufsforschung)
Sophie Hensgen (Institut für Arbeitsmarkt- und Berufsforschung)
Bettina Müller (Institut für Arbeitsmarkt- und Berufsforschung)
Dana Müller (Institut für Arbeitsmarkt- und Berufsforschung)
Johannes Ludsteck (Institut für Arbeitsmarkt- und Berufsforschung)
Christopher Osiander (Institut für Arbeitsmarkt- und Berufsforschung)
Julia Schmidtke (Institut für Arbeitsmarkt- und Berufsforschung)
Gesine Stephan (Institut für Arbeitsmarkt- und Berufsforschung; Universität Erlangen-Nürnberg)
Annette Trahms (Institut für Arbeitsmarkt- und Berufsforschung)

Die FDZ-Datenreporte beschreiben die Daten des FDZ im Detail. Diese Reihe hat somit eine doppelte Funktion: zum einen stellen Nutzerinnen und Nutzer fest, ob die angebotenen Daten für das Forschungsvorhaben geeignet sind, zum anderen dienen sie zur Vorbereitung der Auswertungen.

FDZ-Datenreporte (FDZ data reports) describe FDZ data in detail. As a result, this series of reports has a dual function: on the one hand, those using the reports can ascertain whether the data offered is suitable for their research task; on the other, the data can be used to prepare evaluations.

Inhaltsverzeichnis

1 Überblick 6

2 Stichprobenkonzept 9

3 Datenerhebung 10

3.1 Survey Operations 10

3.1.1 Welle 1 10

3.1.2 Welle 2 bis 4 11

3.1.3 Überführung auf die Plattform Panelingress 11

3.1.4 Welle 5 bis 7 12

3.2 Ergebnisse 12

3.2.1 Response Raten 12

3.2.2 Linkage Consent 13

4 Datenbeschreibung 13

4.1 Fehlende Werte (Missings) 14

4.2 Offene Angaben 15

4.3 Variablenbenennung und Änderungsverlauf 15

4.4 Variablen mit geändertem Wertebereich 16

4.5 Generierte Variablen 17

4.5.1 Metadaten 17

4.5.2 Inhaltlich komplexe generierte Variablen 19

4.5.3 Randomisierte Variablen 25

4.6 Übertragung von Werten auf Folgewellen (Fortschreibung) 26

4.7 Einschränkung ausgelieferter Fälle im SUF 28

4.7.1 Definition vollständiger Fälle 28

4.7.2 Plausibilisierung von Fällen 29

4.8 Fragenprogramm 30

5 Gewichtung 30

5.1 Beschreibung Gewichtungskonzept 31

5.1.1 Wahrscheinlichkeitsgewichtung 31

5.1.2 Kalibrierung 33

5.1.3 Integration der *zweiten* Rekrutierungsstichprobe (ab Welle 5) 34

5.1.4 Variablen zur Schätzung von Response-, Record-Linkage-, und Panel-Consent-Wahrscheinlichkeiten 35

5.2 Auslieferung der Gewichte 37

Literatur 39

Tabellenverzeichnis

Tabelle 1: Kurzbeschreibung 8
Tabelle 2: Response Raten (RR) für neu Rekrutierte, Welle-1-Panelisten und Auffrisch-Panelisten 13
Tabelle 3: Generierte Variablen und Meta-Daten 18
Tabelle 4: Fortgeschriebene Variablen: Originalzeitpunkt Erfassung 27
Tabelle 5: Fragen aus anderen Studien 30

Abstract

Seit sich das Corona-Virus SARS-CoV-2 ausbreitet, hat sich das Leben in Deutschland stark verändert. Viele Menschen müssen neue Anforderungen wie Homeoffice oder Kinderbetreuung zu Hause bewältigen, sind von Kurzarbeit betroffen oder haben ihre Stelle verloren. Die Regierungen auf Bundes- und Länderebene versuchen dieser Lage mit verschiedensten Maßnahmen – wie Ausgangsbeschränkungen oder Maskenpflicht – Herr zu werden. Aber wie wirken sich das Virus und die Maßnahmen auf das Sozial- und Arbeitsleben der Menschen aus? Das IAB hat mit dem Hochfrequenten Online-Personen-Panel „Leben und Erwerbstätigkeit in Zeiten von Corona“ (IAB-HOPP) eine neue Befragung aufgesetzt, die im Zeitraum von Mai 2020 bis März 2021 monatlich Personen zu ihrer derzeitigen Lebens- und Arbeitssituation befragte. Die Daten sollen Wissensschaftlerinnen und Wissenschaftlern die Möglichkeit geben, die Auswirkungen der Covid-19-Pandemie zu erforschen. Dieser Datenreport gibt einen Überblick über den Datensatz und die dazu notwendigen Aufbereitungsschritte.

Since the Corona virus SARS-CoV-2 started spreading, life in Germany has changed. Many people have to cope with new situations such as home office or child care at home, may be affected by short-time work or have lost their jobs. Governments at the federal and state level are trying to cope with this situation with a wide variety of measures – such as exit restrictions or mandatory masks. But how do the virus and the measures affect people's social and working lives? The IAB has set up a new survey, the High-frequency Online Personal Panel "Life and Work Situations in Times of Corona" (IAB-HOPP), which asked people about their current living and working situation every month from May 2020 to March 2021. The data will allow researchers to explore the impact of the Covid-19 pandemic. This data report provides an overview of the prepared data.

Keywords

Covid-19, Hochfrequente Datenerhebung, Panelerhebung, Erwerbstätigkeit

1 Überblick

Die Ausbreitung des Corona-Virus SARS-CoV-2 hat das Leben in Deutschland stark beeinflusst. Viele Menschen mussten neue Anforderungen wie Homeoffice oder Kinderbetreuung zu Hause bewältigen, waren von Kurzarbeit betroffen oder haben ihre Stelle verloren. Die Regierungen auf Bundes- und Länderebene haben versucht, dieser Lage mit verschiedensten Maßnahmen – wie Ausgangsbeschränkungen oder Maskenpflicht – Herr zu werden. Aber wie wirken sich das Virus und die Maßnahmen auf das Sozial- und Arbeitsleben der Menschen aus? Konventionelle Panelbefragungen haben lange Erhebungsdauern, welche die hoch dynamische Situation der sich ständig ändernden Lebensumstände kaum erfassen können. Das IAB hat mit dem Hochfrequenten Online Personen Panel „Leben und Erwerbstätigkeit in Zeiten von Corona“ (IAB-HOPP) daher eine neue Befragung aufgesetzt, die dazu beitragen soll, einen besseren Blick auf die Lebens- und Arbeitswelt der Menschen während der Corona-Krise zu bekommen und längerfristig die Forschungsmöglichkeiten zu den Auswirkungen der Covid-19-Pandemie zu verbessern.

Anfang Mai 2020 hat das IAB 200.000 Personen per Post zur Teilnahme an einer standardisierten Online-Befragung mit dem oben genannten Titel eingeladen. Mit der Zustimmung der Befragten wurden diese in einem monatlichen (Welle 2 bis 4) und später zweimonatlichen Rhythmus zu einer Wiederbefragung eingeladen (Welle 5 bis 7). Um Veränderungen auch innerhalb von kurzen Zeiträumen feststellen zu können, wurden die Befragten ab Welle 2 in vier Gruppen eingeteilt. In jeder Welle erhalten die vier Befragtengruppen das gleiche Fragenprogramm, werden aber wöchentlich (ab Welle 5: zweiwöchentlich) versetzt zur Befragung eingeladen. Dieses Design ermöglicht es, flexibel auf plötzlich eintretende Änderungen zu reagieren.

Für die Befragung wurde eine proportional geschichtete Zufallsstichprobe aus den Integrierten Erwerbsbiografien (IEB) mit dem Stand vom 31.12.2018 gezogen. In den IEB sind detaillierte tagesgenaue Informationen in Kontenform zu Personen enthalten, die einer sozialversicherungspflichtigen Beschäftigung nachgehen, Arbeitslosengeld aus der Arbeitslosenversicherung oder steuerfinanzierte Leistungen der Grundsicherung beziehen, an arbeitsmarktpolitischen Maßnahmen teilnehmen und/oder sich bei der Agentur für Arbeit arbeitsuchend gemeldet haben. Damit können Personen mit unterschiedlichem Erwerbsstatus befragt werden, wie sie die Herausforderungen der Corona-Krise bewältigen. Darüber hinaus können repräsentative Aussagen für Erwerbspersonen (mit Ausnahme Selbstständiger und Beamter) getroffen werden, für die bei der Bundesagentur für Arbeit Informationen vorliegen.

Ziel der Befragung ist es, die Entwicklung der Lebens- und Erwerbssituation während und nach der Corona-Krise zu erfassen. Die gewonnenen Daten sollen es der Forschung ermögli-

chen, die Auswirkungen der Covid-19-Pandemie besser zu verstehen und politische Akteure angemessen zu möglichen Maßnahmen zu beraten.

Dieser Datenreport enthält die Beschreibung der Erhebungswellen 1 bis 7, die im Zeitraum von Mai 2020 bis März 2021 erhoben wurden. Der bereitgestellte Datensatz beinhaltet dementsprechend die Fälle von Welle 1 bis 7. Tabelle 1 enthält eine Kurzbeschreibung des Datensatzes.

Tabelle 1: Kurzbeschreibung

Kategorien	Erläuterungen
Kurzbeschreibung des Inhalts	Folgende Themen werden abgedeckt: Sorgen und Zufriedenheit, Erwerbsstatus, Arbeitszeit, Homeoffice, Kurzarbeit, Haushalt und Kinderbetreuung; für mehr Details siehe die Variablenübersicht.
Untersuchungseinheit(en)	Sozialversicherungspflichtig Beschäftigte, Arbeitslose, Arbeitsuchende, Leistungsempfängerinnen und -empfänger, Teilnehmende an Maßnahmen
Fallzahlen	Welle 1 = 11.311 Welle 2 = 4.746 Welle 3 = 4.071 Welle 4 = 3.682 Welle 5 = 11.072 Welle 6 = 6.659 Welle 7 = 6.334
Zeitraum	08.05.2020 - 15.03.2021
Erhebungsdesign	online
Dateityp	Stata-File
Dateigröße	• HOPP_W1-W7_v1: 25,3 MB • HOPP_Weights4Waves_W1-W7_v1: 2,75 MB • HOPP_Weights4Months_W1-W7_v1: 2,75 MB
Dateiorganisation	Die Daten werden in drei separaten Datensätzen ausgeliefert (Befragungsdaten, Wellengewichte, Monatsgewichte)
Datenzugang	Scientific Use File (SUF)
Anonymisierungsgrad	Faktisch anonymisierte Form
Sensible Merkmale	Keine
Zitierung der Daten und Datendokumentation	• Haas, Georg-Christoph; Müller, Bettina; Osiander, Christopher; Schmidtke, Julia; Trahms, Annette; Volkert, Marieke; Zins, Stephan (2021): Development of a New COVID-19 Panel Survey: The IAB High-frequency Online Personal Panel (HOPP): Journal for Labour Market Research. DOI: 10.1186/s12651-021-00295-z • **Daten**: Die Datengrundlage dieses Beitrags bilden die faktisch anonymisierten Daten des Hochfrequenten Online Personen Panel „Leben und Erwerbstätigkeit in Zeiten von Corona“ (IAB-HOPP), Wellen 1 bis 7. Der Datenzugang erfolgte über ein Scientific Use File, das über das Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung zur Verfügung gestellt wurde. DOI: 10.5164/IAB.HOPP_W01-W07.de.en.v1 • **Datendokumentation:** Volkert, Marieke; Haas, Georg-Christoph; Zins, Stefan; Bellmann, Lisa; Dummert, Sandra; Hensgen, Sophie; Ludsteck, Johannes; Müller, Bettina; Müller, Dana; Osiander, Christopher; Schmidtke, Julia; Stephan, Gesine; Trahms, Annette (2021): Dokumentation und Codebuch für das Hochfrequente Online Personen Panel “Leben und Erwerbstätigkeit in Zeiten von Corona” (IAB-HOPP), wellen 1 bis 7. FDZ-Datenreport, 04/2021 (de), Nürnberg. DOI: 10.5164/IAB.FDZD.2104.de.v1
Datensatzversion	Hochfrequente Online Personen Panel „Leben und Erwerbstätigkeit in Zeiten von Corona“ (IAB-HOPP), Wellen 1 bis 7; DOI: 10.5164/IAB.HOPP_W01-W07.de.en.v1

Anmerkung:
Details zu verschiedenen Zugangswegen, zu den Voraussetzungen der Nutzung sowie zur Antragstellung des Datensatzes können der FDZ-Internetseite unter https://fdz.iab.de/ entnommen werden.

2 Stichprobenkonzept

Die Bruttostichprobe für die HOPP-Studie wurde aus den Integrierten Erwerbsbiografien (IEB) gezogen (siehe z.B. (Dorner u. a., 2010)). Die IEB enthalten detaillierte tagesgenaue administrative Informationen, die Arbeitgeber, Jobcenter und Arbeitsagenturen an die Bundesagentur für Arbeit melden. Diese Datensätze enthalten alle Personen, die mindestens eine der folgenden Meldungen aufweisen: sozialversicherungspflichtige Beschäftigung, geringfügige Beschäftigung, Leistungsbezug, Teilnahme an einer Maßnahme der aktiven Arbeitsmarktpolitik (z.B. Beschäftigung schaffende Maßnahmen oder Förderungen beruflicher Weiterbildung) oder eine Arbeitsuchendmeldung bei der Bundesagentur für Arbeit (Frodermann u. a., 2021). Das schließt Personen aus, die Beamte oder selbstständig sind.

Als Grundlage für die Stichprobe wurden die IEB mit dem Meldedatum bis zum 31.12.2018 berücksichtigt. Es wurden nur Personen, die am 01.05.2020 oder früher ihr 18. Lebensjahr vollendet haben und im Jahr 2018 mindestens eine Meldung (Spell) in den oben genannten Kategorien hatten, bei der Ziehung berücksichtigt. Nicht berücksichtigt wurden Personen mit unvollständiger oder ausländischer Adresse.

Beim Stichprobendesign handelt es sich um eine geschichtete Stichprobe mit einfacher Zufallsstichprobe innerhalb der Schichten, wobei die Schichten nach Region, Alter, Geschlecht und Erwerbsstatus im Jahr 2018 definiert wurden. Für die Schichtung nach Region wurden die Regionaldirektionen der Bundesagentur für Arbeit herangezogen (Wohnort-Prinzip). Die Personen wurden anhand ihres Alters (Stand: 1. Mai 2020) in fünf Kategorien eingeteilt: 18 bis 29 Jahre, 30 bis 39 Jahre, 40 bis 49 Jahre, 50 bis 59 Jahre und 60 bis 99 Jahre. Die Schichten zum Erwerbsstatus im Jahr 2018 umfassen vier Kategorien: a) Personen, die im Jahr 2018 ausschließlich Beschäftigungsmeldungen hatten, wobei sich mindestens eine Meldung auf eine geringfügige Beschäftigung bezog, b) Personen, die im Jahr 2018 ausschließlich Beschäftigungsmeldungen hatten, wobei sich keine Meldung auf eine geringfügige Beschäftigung bezog, c) Personen, die im Jahr 2018 mindestens einmal Arbeitslosengeld II (Grundsicherung) bezogen, und d) Personen, die im Jahr 2018 kein Arbeitslosengeld II bezogen, aber aus anderen Gründen bei der Bundesagentur für Arbeit registriert waren (Bezug von Arbeitslosengeld aus der Arbeitslosenversicherung, Teilnahme an einer arbeitsmarktpolitischen Maßnahme, Arbeitsuchendmeldung).

Der Bruttostichprobenumfang wurde proportional zur Gesamtzahl der Personen in den jeweiligen Schichten des Stichprobenrahmens vergeben, d.h. die Einschlusswahrscheinlichkeiten waren für alle Personen im Stichprobenrahmen gleich ($p = 0{,}0043$). Eine Ausnahme von dieser Regel bildeten ältere Arbeitnehmerinnen und Arbeitnehmer in der Altersgruppe 60 bis 99 Jahre, die im Jahr 2018 beschäftigt waren, sowie geringfügig Beschäftigte, die einen höheren

Stichprobenanteil (p = 0,0063) aufwiesen als Personen in den anderen Schichten.

Da die Verteilung der Schichtungsvariablen größtenteils proportional zu ihrer Verteilung innerhalb des IEB-Stichprobenrahmens ist, ist der Anteil der Personen, die im Jahr 2018 nicht ausschließlich Beschäftigungsmeldungen hatten und damit zu Schicht c) oder d) gehören, in der Stichprobe relativ gering. Dies vermindert die statistische Aussagekraft jeder Analyse, die sich auf eine dieser Untergruppen bezieht, im Vergleich zu den Beschäftigten. Im Hinblick auf inferenzstatistische Aussagen ist es daher empfehlenswert, nur Analysen für den gesamten deutschen Arbeitsmarkt oder nur für Beschäftigte durchzuführen.

Die Panelteilnehmenden wurden zu zwei Zeitpunkten rekrutiert: Für Welle 1 im Mai 2020 und für Welle 5 im September/Oktober 2020. Ziel war es, für Welle 1 eine Nettostichprobe von etwa 10.000 vollständigen Interviews zu erreichen. Basierend auf Erfahrungen aus anderen Studien des IAB mit ähnlicher Zielpopulation, ähnlichem Modus, Stichproben- und Kontaktstrategie wurde eine Rücklaufquote von etwa fünf Prozent erwartet. Daher wurde für Welle 1 eine Stichprobe von 200.000 Personen gezogen. Für Welle 5 wurde eine Auffrischungsstichprobe von 99.188 Fällen mit dem gleichen Design gezogen, um netto wiederum etwa 5.000 neue Fälle zu generieren.

3 Datenerhebung

3.1 Survey Operations

3.1.1 Welle 1

Am 7. Mai 2020 wurden insgesamt 200.000 Personen postalisch eingeladen, an der Befragung teilzunehmen, d.h. die Einladungsschreiben wurden an diesem Tag versandt. Die ersten Rückläufe geschahen am 8. Mai. Das Feld für die erste Welle wurde am 25.05.2020 um 09:30 Uhr formal geschlossen. Über diesen Zeitpunkt hinaus war es eingeladenen Personen zwar immer noch möglich, an der Befragung teilzunehmen. Personen, die später an der Befragung teilnahmen, bekamen jedoch keine Einladung mehr für die zweite Welle. Ausgelieferte Daten, Gewichte und Beschreibungen in diesem Datenreport beziehen sich auf Befragte, die bis einschließlich zum 25.05.2020 um 09:30 Uhr teilgenommen haben.

3.1.2 Welle 2 bis 4

Um Aussagen über die wöchentliche Entwicklung der Auswirkungen der Corona-Krise treffen zu können, wurden alle Teilnehmenden, die bis zum 25.05.2020 an der Befragung teilgenommen haben, in vier Teilstichproben eingeteilt, die als T1, T2, T3 und T4 bezeichnet werden. Die Einteilung in die Teilstichproben erfolgte randomisiert. Bei den Wellen 2 bis 4 wurden Personen in diesen Teilstichproben wöchentlich versetzt eingeladen. Jede Woche startete eine neue Teilstichprobe ins Feld, während die Befragung der vorherigen Teilstichprobe weiterlief. Teilnehmende aus Welle 1 hatten die Wahl, ob sie postalisch, per E-Mail oder gar nicht zu einer Folgebefragung eingeladen werden wollten. Für Personen, die einer postalischen Kontaktierung zugestimmt hatten, wurden am Donnerstag der Einladungswoche Einladungsschreiben verschickt, so dass die potenziellen Befragten die Einladung im Regelfall am Freitag oder Samstag erhielten. Personen, die zuvor einer Kontaktierung per E-Mail zugestimmt hatten, wurden am Freitag der Befragungswoche mit einer E-Mail eingeladen.

Um Überschneidungen zwischen den Wellen zu vermeiden, hatte jede Teilstichprobe eine Feldzeit von 23 Tagen. Die Befragungen konnte aus technischen Gründen allerdings nicht für einzelne Teilstichproben geschlossen werden, so dass es Personen nach Ablauf von 23 Tagen immer noch möglich war, an der Befragung teilzunehmen. Für den Scientific Use File (SUF) werden allerdings nur Personen berücksichtigt, die innerhalb von 23 Tagen nach dem Einladungszeitpunkt antworteten.

3.1.3 Überführung auf die Plattform Panelingress

Aufgrund des engen Zeitplanes wurde die Befragung mit der Software „KeyIngress“ gestartet, die über die Bundesagentur für Arbeit zur Verfügung gestellt wird. Über KeyIngress besteht allerdings keine Möglichkeit, softwarebasiert Incentives an die Teilnehmenden auszuzahlen. Um die Teilnahmebereitschaft zu sichern, wurde deswegen während der laufenden Feldzeit entschieden, die Panelisten in die Software „PanelIngress“ zu überführen und mittels dieser Software weiter zu befragen[1]. Diese Plattform ermöglicht die softwarebasierte Ausgabe von Gutscheinen als Incentives für die Teilnahme an der Befragung und weitere Funktionen zur Paneladministration.

Panelisten, die eine E-Mail-Adresse angegeben hatten, wurden wie nachfolgend beschrieben eingeladen, sich auf der Panelwebsite zu registrieren: Die Mitglieder der Teilstichproben T3 und T4 bekamen am Ende ihrer Befragung im Rahmen der dritten Welle eine Einladung mit der Bitte, sich in PanelIngress zu registrieren. Als Dankeschön für ihre bisherige Teilnahme

[1] Die Befragungsseite ist unter http://www.iab-online-befragung.de abrufbar.

und die Registrierung wurden diesen Personen 500 Punkte mittels des Gutscheinsystems zugesagt. 100 Punkte entsprechen einem Wert von einem Euro, 500 Punkte demzufolge einem Wert von fünf Euro. Die Gutscheine können ab einem Gutscheinwert von fünf Euro bei verschiedenen kommerziellen Anbietern für Produkte eigener Wahl eingelöst werden. Panelisten in den Teilstichproben T1 und T2 wurden zu Beginn der vierten Welle eingeladen, sich auf der Befragungsseite zu registrieren. Dies geschah, bevor sie zur Befragung in Welle 4 gelangten. Auch diese Teilnehmenden erhielten bei Registrierung 500 Punkte.

Allen Teilnehmenden wurden zudem für die Teilnahme an jeder weiteren Befragungswelle 200 Punkte angeboten. Panelisten aus den Teilstichproben T3 und T4 der vierten Welle, die sich nicht in Welle 3 registriert hatten, erhielten eine E-Mail-Einladung von PanelIngress, sich auf der Panelwebsite zu registrieren. Panelisten, die keine E-Mail-Adresse angegeben hatten, wurden am Ende der Befragung von Welle 4 eingeladen, sich in PanelIngress zu registrieren.

3.1.4 Welle 5 bis 7

Zusätzlich zu den Panelisten aus Welle 1 bis 4 wurde in Welle 5 eine Auffrischungsstichprobe von 99.188 Personen eingeladen. Die Auffrischungsstichprobe wurde an das Design der Wellen 2-4 angelehnt, d.h. alle Auffrischungsfälle wurden ebenfalls in eine der bereits existierenden vier Teilstichproben T1-T4 eingeteilt. Der zeitliche Abstand, in dem die Teilstichproben befragt wurden, wurde ab Welle 5 von einer Woche auf zwei Wochen erhöht und bis zum Ende der Welle 7 beibehalten.

Auffrischungsfälle hatten die Möglichkeit, sich am Ende ihrer ersten Befragung – also nach Welle 5 – in PanelIngress zu registrieren. Als Dankeschön für ihre Registrierung erhielten diese Personen bei erfolgter Registrierung ebenfalls 500 Punkte. Die Möglichkeit, für Folgewellen postalisch eingeladen zu werden gab es, anders als für Befragte der Welle 1, nicht mehr. Ab Welle 6 wurden alle registrierten Panelisten über die Panelwebsite eingeladen, an weiteren Befragungswellen teilzunehmen.

3.2 Ergebnisse

3.2.1 Response Raten

Response-Raten werden auf Grundlage der AAPOR-Standarddefinitionen für Rücklaufquoten 1 (RR1) berechnet: die Anzahl der vollständigen Interviews wird durch die Anzahl der eingeladenen Fälle geteilt (AAPOR, 2016). Als vollständig werden Interviews definiert, in denen

die Befragten eine Antwort auf die letzte inhaltliche Frage gegeben haben. Falls die Befragten ihr Einverständnis für die Verknüpfung mit den administrativen Daten gegeben haben, wurden bei der Datenaufbereitung Alter und Geschlecht zwischen den Befragungsdaten und den administrativen Daten abgeglichen. Alle Fälle, bei denen Alter und Geschlecht in den Befragungs- und den administrativen Daten nicht übereinstimmten, wurden als unplausible Fälle gekennzeichnet und nicht in den SUF mitaufgenommen.

Response-Raten für alle Wellen können Tabelle 2 entnommen werden.

Tabelle 2: Response Raten (RR) für neu Rekrutierte, Welle-1-Panelisten und Auffrisch-Panelisten

Welle	**Neu Rekrutiert**		Welle-1-Panelisten		Auffrisch-Panelisten		Anzahl realisierter Interviews	
	$N_{eingeladen}$	RR^a	$N_{eingeladen}$	RR^a	$N_{eingeladen}$	RR^a	*Alle*	*mit Zuspielerlaubnis*
1	200.000	5,7	-	-	-	-	11.311	9.548
2	-	-	9.751	48,7	-	-	4.746	4.258
3	-	-	9.751	41,7	-	-	4.071	3.673
4	-	-	9.751	37,8	-	-	3.682	3.339
5	99.188	8,2	3.739	79,5	-	-	11.072	9.595
6	-	-	3.744	82,4	4.939	72,3	6.659	6.141
7	-	-	3.737	80,6	4.941	67,3	6.334	5.836

[a] (AAPOR, 2016)
Tabelle übernommen aus (Haas u. a., 2021)

3.2.2 Linkage Consent

Für eine Verknüpfung von Befragungs- und administrativen Daten ist aus datenschutzrechtlichen Gründen die explizite Zustimmung der Befragungsteilnehmenden notwendig. Erstbefragte wurden deshalb in Welle 1 und 5 gebeten, ihre Zustimmung zur Zuspielung der administrativen Daten zu den Befragungsdaten zu geben. Die letzte Spalte von Tabelle 2 gibt Auskunft über die Zahl der Personen, die ihre Zustimmung gegeben haben.

4 Datenbeschreibung

Der Datensatz wird als Stata-Datei im Long-Format bereitgestellt. Fälle der gleichen Befragungsperson können über die ID-Variable „iab_id“ identifiziert und für längsschnittliche Analysen in ein Wide-Format transformiert werden. Im Folgenden werden wichtige Eigenschaften des Datensatzes wie der Umgang mit fehlenden Werten, generierten Variablen, Fortschreibungen u.Ä. detailliert erläutert.

4.1 Fehlende Werte (Missings)

Fehlende Werte im Datensatz sind mit negativen Vorzeichen gekennzeichnet und schlüsseln sich inhaltlich wie folgt auf:

-1 „Weiß nicht“
Dieser fehlende Wert wird gesetzt, wenn die entsprechende Antwortmöglichkeit im Fragebogen explizit ausgewählt werden konnte.

-2 „Keine Angabe“
Dieser fehlende Wert wird gesetzt, wenn eine Frage von der befragten Person unbeantwortet übersprungen wurde.

-3 „Trifft nicht zu (Filter)“
Hierunter fallen alle durch die Frageprogrammierung gesetzten Filter.

-5 „Inkonsistenter Wert“
Der Wert wurde im Rahmen der Datenbereinigung gesetzt.

-6 „Vom SUF ausgeschlossene offene Angabe“
Offene Angaben werden generell nicht im SUF ausgeliefert. Durch diesen fehlenden Wert ist allerdings feststellbar, ob eine offene Angabe gemacht wurde.

-7 „Unvollständige Angabe“
Dieser Wert wird ausschließlich bei generierten Variablen verwendet, die aus verschiedenen Gründen nicht durch valide Angaben befüllt werden können.

-8 „Vorzeitiges Beenden des FraBo“
Dieser fehlende Wert trifft auf wenige freiwillige Angaben am Ende der Befragung zu, die für ein vollständiges Interview nicht notwendig waren.

-11 „Frage wurde nicht erhoben“
Dieser fehlende Wert gibt an, wenn eine Frage nicht in jeder Welle erhoben wurde.

Der Regelfall in der Befragung waren nicht-verpflichtende Fragen. Das bedeutet, dass Personen Fragen überspringen konnten, wenn sie auf den Button „Weiter“ klickten. War dies der Fall, erhalten die entsprechenden Variablen den Wert „-2“. Bei wenigen Fragen wurde eine aktive Verneinung als Antwortmöglichkeit eingeräumt, die in diesem Fall nicht als fehlender Wert kodiert wird, sondern als eigener Antwort-Code im Datensatz auftaucht:

Beispiel 1
Teilnehmende Personen konnten auf die Frage, wie viele Überstunden sie in der letzten Woche geleistet haben, mit „gar nicht“ antworten. Die Antwort ist in Variable AZ3000 entsprechend mit dem Antwort-Code „991“ versehen.

Beispiel 2
Auf die Frage, ob die im Homeoffice absolvierte Arbeit innerhalb normaler Arbeitszeiten oder

auch in der Freizeit stattfindet, konnten Personen auch ein „Trifft nicht zu“ angeben. Diese Antwort wird als „7“ kodiert, da sie aktiv von der befragten Person ausgewählt wurde.

4.2 Offene Angaben

Offene Angaben können personenbezogene Daten enthalten. Daher sind diese *nicht* im SUF enthalten. Offene Angaben, die nicht zur Verfügung stehen, betreffen z.B. Angaben zu Schulbildung und zur beruflichen Ausbildung (z.B. SD4000_o), aber auch freiwillige Anmerkungen am Ende des Fragebogens (AN1000). Solche Antworten wurden bislang nicht nachträglich kodiert und kategorisiert. Bei der Datennutzung kann jedoch durch den fehlenden Wert „-6“ identifiziert werden, ob eine offene Angabe gemacht wurde.

4.3 Variablenbenennung und Änderungsverlauf

Bei der Benennung der Variablen orientierte sich das Projektteam an dem Variablenkonzept des „Panel Arbeitsmarkt und soziale Sicherung (PASS)“ (Berg u. a., 2020). Demnach werden Veränderungen an Variablen von einer Welle zur nächsten direkt aus dem Variablennamen ersichtlich, etwa anhand der Formulierungen von Frage- oder Antworttexten. Die erste Version einer Variable endet zunächst immer mit drei Nullen. Wurde die Formulierung einer Frage geändert oder die Variable im Rahmen der Datenaufbereitung angepasst, wird dies durch „Hochzählen“ der Nullstellen deutlich gemacht. Im folgenden Text werden diese Nullstellen als „Veränderungs-Kenner“ bezeichnet. Zur Veranschaulichung werden die unterschiedlichen Typen einer Veränderung am fiktiven Beispiel der Variable EW2000 erläutert:

EW2000
: Drei laufende Nullen beschreiben die erste Version der erhobenen Variable EW2000.

EW2100
: Verändert sich die Null an der ersten Stelle, wurde die der Variable zugrundeliegende Frage verändert, z.B. durch eine angepasste Frageformulierung oder geänderte Wertebereiche. EW2100 entspricht also der zweiten, EW2200 der dritten Version der gleichen inhaltlichen Frage usw..

EW2010
: Verändert sich die Null an der zweiten Stelle, beschreibt dies Variablen, die im Zuge der Datenaufbereitung verändert wurden. Dies betrifft z.B. die Inklusion offener Angaben oder das nachträgliche Zusammenfassen bestimmter Kategorien.

EW2001
: Verändert sich die dritte Nullstelle, beschreibt dies fortgeschriebene Variablen, in denen

Werte aus Vor- oder Folgewellen im Zuge der Datenaufbereitung übertragen wurden. Dies betrifft z.B. die Abfrage der Arbeitszeiten vor der Covid-19-Pandemie und damit Variablen, die für Analysen in allen Wellen gleichermaßen relevant sein dürften, aber nur einmal abgefragt wurden.

Alle Veränderungs-Kenner einer Variable können beliebig miteinander kombiniert werden. Beispielsweise entspräche die Variable mit der Bezeichnung EW2110 einer Variable, die im Laufe der Befragung einmal angepasst wurde und bei der im Rahmen der Datenaufbereitung zusätzlich offene Nennungen integriert wurden.

4.4 Variablen mit geändertem Wertebereich

Während der Fragebogenentwicklung wurden teils kleinere Veränderungen an wiederholt gestellten Fragen vorgenommen, so dass sich Eingabemöglichkeiten für befragte Personen veränderten. Hierzu zählten insbesondere veränderte Wertebereiche bei offenen numerischen Angaben. Solche Veränderungen können zu Änderungen im Antwortverhalten führen, beispielsweise wenn eine Person eher gar keine Angabe macht, weil der Wert „0“ als mögliche Antwort ausgeschlossen ist. Um Änderungen an Variablen im Datensatz direkt kenntlich zu machen, wurden Variablennamen entsprechend der Beschreibung in Absatz 4.3 angepasst. Mit Ausnahme von zeitkonstanten Variablen, die fortgeschrieben wurden[2], wurde darauf verzichtet, generierte Variablen zu erstellen. Falls von Veränderung betroffene Variablen in Zeitreihenanalysen verwendet werden sollen, müssen folgende Variablen wieder in eine Variable zusammengespielt werden.

- **AZ2***: Wenn Sie an Ihre letzte Arbeitswoche denken: Wie viele Stunden haben Sie tatsächlich gearbeitet, einschließlich regelmäßig geleisteter Überstunden, Mehrarbeit usw.?
  - AZ2000 Wertebereich: 0-120
  - AZ2100 Wertebereich: 1-120 ab Welle 5
- **AZP2***: Und wie viele Stunden hat Ihr Partner/Ihre Partnerin in der letzten Arbeitswoche gearbeitet, einschließlich regelmäßig geleisteter Überstunden, Mehrarbeit usw.?
  - AZP2000 Wertebereich: 0-120
  - AZP2100 Wertebereich: 1-120 ab Welle 5
- **HO3***: Wenn Sie an Ihre letzte Arbeitswoche denken: Wie viele Stunden haben Sie zu Hause gearbeitet?
  - HO3000 Wertebereich: 0-120
  - HO3100 Wertebereich: 1-120 ab Welle 5

[2] Siehe dazu Absatz 4.6

- **HOP3***: Und wie viele Stunden hat Ihr Partner/Ihre Partnerin in der letzten Arbeitswoche zu Hause gearbeitet?
  - HOP3000 Wertebereich: 0-120
  - HOP3100 Wertebereich: 1-120 ab Welle 5
- **KA3***: Zahlt Ihr Betrieb Ihnen eine Zulage zum Kurzarbeitergeld?
  - KA3010 Zusammenfassung der Antwortkategorien 2 und 3
  - KA3100 Vereinfachte Ja/Nein Abfrage
- **Z5***: […] Zukünftig möchten wir Sie zu Befragungen über eine neue Befragungsseite einladen. Damit Sie uns weiter unterstützen können, bitten wir Sie, sich auf der Seite zu registrieren. […]
  - Z5000
  - Z5100 Veränderter Fragetext

4.5 Generierte Variablen

Im SUF sind Variablen enthalten, die im Zuge der Datenaufbereitung generiert und in der vorliegenden Form nicht in der Befragung selbst erhoben wurden. Die Variablennamen so erstellter Variablen beginnen entweder mit „GEN“ oder „PA“ und lassen sich dementsprechend in zwei Gruppen aufteilen. Zum einen zählen dazu Meta-Daten des Datenexports (siehe hierzu Abschnitt 4.5.1). Zum anderen fallen darunter inhaltliche Variablen, bei denen sich entweder die Filterführung oder die Formulierung der Frage verändert hat. Diese werden anschließend in Abschnitt 4.5.2 näher beschrieben. Die entsprechenden Variablen sind in Tabelle 3 aufgelistet.

4.5.1 Metadaten

Im Datensatz werden Metainformationen pro Interview bereitgestellt, die zu den generierten Variablen gezählt werden. Die folgende Liste enthält ausschließlich Variablen, die nähere Erläuterungen erforderlich machen. Darüber hinaus wurden weitere Metainformationen zum Interview bereitgestellt, die der Variablenübersicht[3] entnommen werden können.

GEN1000

Die Variable GEN1000 enthält Informationen zur aktuellen Befragungswelle. Sie hat aktuell die Ausprägungen 1 bis 7, die für die jeweilige Welle stehen. Eine eingehende Beschreibung des Wellenplans kann Abschnitt 3.1 entnommen werden.

[3] Die Variablenübersicht ist in einer gesonderten Excel-Datei über die Seite des FDZ verfügbar.

Tabelle 3: Generierte Variablen und Meta-Daten

Name	Bedeutung
GEN1000	Erhebungswelle
GEN4000	Kalenderwoche Interview beendet
GEN5000	Monatliche Stichprobe
GEN7000a	Einladungsmonat (numerisch), Wertebereich: 1-12
GEN7000b	Responsemonat entspricht dem Einladungsmonat Indikator für Monatsgewichte
GEN8000a	Einladungswoche (numerisch), Wertebereich: 1-51
GEN8000b	Responsewoche entspricht der Einladungswoche Indikator für Wochengewichte
GEN_allein	Alleine wohnend
GEN_Partner	Zusammenleben mit festem Partner/fester Partnerin
GEN_Kind_u18	Kinder unter 18 Jahren im Haushalt
GEN_andPers	Andere verwandte/nicht-verwandte Personen im Haushalt
GEN_Haushalt	Haushaltsgröße
GEN_GebKinda	Geburtsjahr Kind 1 (unter 18 Jahren im Haushalt)
GEN_GebKindb	Geburtsjahr Kind 2 (unter 18 Jahren im Haushalt)
GEN_GebKindc	Geburtsjahr Kind 3 (unter 18 Jahren im Haushalt)
GEN_GebKindd	Geburtsjahr Kind 4 (unter 18 Jahren im Haushalt)
IMP_GebKind	Korriegierte Angaben GEN_GebKind
GEN_selbst	Erwerbsstatus aktuell: Selbstständigkeit
GEN_konsum	Konsum Einschätzung
PA1000	Interviewdauer
PA8000	Datum Start Interview
PA9000	Datum Ende Interview

GEN4000

Die Variable GEN4000 beschreibt, in welcher Kalenderwoche ein Interview beendet wurde. Sie kann entsprechend die Ausprägungen 1-52 annehmen.

GEN5000

Ab Welle 2 wurden die Befragten in vier Teilstichproben untergliedert, die bis einschließlich Welle 4 im Abstand von einer Woche und ab Welle 5 im Abstand von zwei Wochen zu der Befragung eingeladen wurden (Vergleiche auch Abschnitt 3.1.2). Die Zuweisung zu einer Teilstichprobe ist der Variable GEN5000 zu entnehmen. Sie kann die Ausprägungen 1-4 annehmen, die für die Teilstichproben T1 bis T4 stehen. Personen, die ausschließlich in Welle 1 teilgenommen haben, werden keiner Teilstichprobe zugeordnet und weisen in Welle 1 entsprechend fehlende Werte auf. Die Zuweisung zu einer Teilstichprobe bleibt über die Wellen hinweg konstant, d.h. Personen wechseln nicht von einer Teilstichprobe in eine andere.

4.5.2 Inhaltlich komplexe generierte Variablen

Inhaltlich generierte Variablen werden dann bereit gestellt, wenn sich die Abfrage aufgrund veränderter Filterführung oder einer Anpassung der Frageformulierung zwischen den Wellen deutlich geändert hat. Die ursprünglichen Originalvariablen werden im SUF nicht mehr zur Verfügung gestellt. Für Analysen ist es hilfreich, über die Variablenübersicht[4] in der Spalte „Anmerkungen“ zu überprüfen, in welche generierte Variable eine Variable überführt wurde. Bei generierten Variablen werden fehlende Werte nicht mehr gesondert aufgeschlüsselt, sondern mit dem fehlenden Wert „-7“ ausgewiesen, falls ein Wert aufgrund fehlender oder implausibler Angaben nicht ermittelt werden kann.

Im Folgenden wird die Erstellung einzelner generierter Variable beschrieben und dokumentiert, aus welchen Originalvariablen sich die jeweilige generierte Variable zusammensetzt.

GEN_allein

Kurzbeschreibung:

Die Variable GEN_allein gibt an, ob eine befragte Person alleine lebt. Mit der Variable lassen sich Single-Haushalte identifizieren.

Originalvariablen:

- HA1000, HA1100, HA1200: Wie viele Personen leben ständig in Ihrem Haushalt, Sie selbst eingeschlossen?
- HA7000a: Aktuell leben Sie in Ihrem Haushalt: allein

Generierung:

Die Variable GEN_allein speist sich zum einen aus Informationen der Variable HA1000 (Welle 1)[5], zum anderen aus Angaben aus Variable HA7000a. Diese wurde regulär in Welle 2-3, in Welle 5 der Auffrischer-Stichprobe und allen Befragten in Welle 6 gestellt. In den Wellen 4-5 und 7 wurde die HA7000a nur jenen Personen gestellt, die zuvor keine Angaben gemacht hatten.
Die generierte Variable GEN_allein wird in allen Wellen bereitgestellt, fehlende Werte werden erst aus Angaben angrenzender vorausgegangener und danach aus nachgelagerten Wellen imputiert. Falls es zu Widersprüchen[6] zwischen HA7000a (wohne alleine) und Angaben in der

[4] Die Variablenübersicht ist in einer gesonderten Excel-Datei über die Seite des FDZ verfügbar.
[5] In Welle 5 wird HA1000 aufgrund des geänderten Eingabe-Wertebereichs zu HA1100 und in Welle 6 zu HA1200
[6] Die HA1*-Abfragen sind zwar durch die Filterführung mit der HA7*-Abfrage plausibilisiert, vereinzelt kann es aber zu widersprüchlichen Angaben kommen, wenn Befragte im Fragebogen zurückgehen und Angaben nachträglich korrigieren.

Anzahl der Haushaltsmitglieder (HA1*) kommt, wird die Angabe in der Variable HA7000a als wahrer Wert angenommen.

GEN_Partner

Kurzbeschreibung:

Die Variable GEN_Partner gibt an, ob die befragte Person mit einem Partner oder einer Partnerin in einem Haushalt zusammen lebt.

Originalvariablen:

- HA2000: Leben Sie in Ihrem Haushalt mit einem festen Partner/einer festen Partnerin zusammen?
- HA2100: Bitte geben Sie alle zutreffenden Kategorien an. Aktuell leben Sie in Ihrem Haushalt...: mit einem festen Partner/einer festen Partnerin

Generierung:

Angaben zu einem im Haushalt lebenden Partner oder einer Partnerin wurden in Welle 1 in Variable HA2000, ab Welle 2 im Rahmen einer Mehrfach-Abfrage in Variable HA2100 gemacht, die regulär in Welle 2, in Welle 3, in Welle 5 der Auffrischer-Stichprobe und in Welle 6 allen Befragten gestellt wurde. In den Wellen 4, 5 und 7 wurde die Frage nur jenen Personen gestellt, die in Welle 2, 3 oder 6 keine regulären Angaben gemacht hatten. Plausible Angaben aus zeitlich angrenzenden Vorwellen werden im Abgleich mit Angaben aus der Variable GEN_allein in Folgewellen imputiert.

GEN_Kind_u18

Kurzbeschreibung:

Die Variable GEN_Kind_u18 gibt an, ob ein oder mehrere minderjährige leibliche Kinder, Stief-, Adoptiv-, oder Pflegekinder im Haushalt leben.

Originalvariablen:

- HA3000: Wie viele Kinder unter 18 Jahren leben in Ihrem Haushalt?
- HA7000c: Bitte geben Sie alle zutreffenden Kategorien an. Aktuell leben Sie in Ihrem Haushalt…: mit leiblichen Kindern, Stief-, Adoptiv-, oder Pflegekindern (unter 18 Jahren)

Generierung:

Die Variable GEN_Kind_u18 wurde über Informationen zur Anzahl der im Haushalt lebenden leiblichen, Stief-, Adoptiv-, oder Pflegekindern aus der Variable HA3000 in Welle 1, 5 und 6 sowie ab Welle 2 aus der Bejahung der Variable HA7000c gebildet. Variable HA7000c wurde regulär in Welle 2, Welle 3, in Welle 5 der Auffrischer-Stichprobe und in Welle 6 allen Befragten gestellt. In den Wellen 4, 5 und 7 wurde die Frage nur jenen Personen gestellt, die in Welle 2, 3 oder 6 keine regulären Angaben gemacht hatten. Plausible Angaben aus Vorwellen werden im Abgleich mit Angaben aus der Variable GEN_allein in Folgewellen imputiert.

GEN_andPers

Kurzbeschreibung:

Die Variable GEN_andPers gibt an, ob weitere verwandte oder nicht-verwandte Personen im Haushalt leben.

Originalvariablen:

- HA7000d: Bitte geben Sie alle zutreffenden Kategorien an. Aktuell leben Sie in Ihrem Haushalt...: mit anderen verwandten Personen
- HA7000e: Bitte geben Sie alle zutreffenden Kategorien an. Aktuell leben Sie in Ihrem Haushalt...: mit anderen nicht verwandten Personen

Generierung:

Die Variable GEN_andPers kann erst ab Welle 2 bereitgestellt werden und setzt sich aus Angaben aus der Variable HA7000d und HA7000e zusammen. In diesen Variablen ist die Information vorhanden, ob „weitere“ verwandte oder nicht-verwandte Personen (neben Partnerin bzw. Partner oder Kindern unter 18 Jahren) im Haushalt leben. Auf eine Rekonstruktion aus Welle 1 durch andere Variablen wurde verzichtet. Die Fragen zu den Variablen HA7000d und HA7000e wurden in Welle 2, Welle 3, in Welle 5 der Auffrischer-Stichprobe und in Welle 6 allen Befragten gestellt. In den Wellen 4, 5 und 7 wurde die Frage nur jenen Personen gestellt, die in Welle 2, 3 und 6 keine Angaben gemacht hatten. Plausible Angaben aus Vorwellen werden im Abgleich mit Angaben aus der Variable GEN_allein in Folgewellen imputiert. Im Gegensatz zu anderen Haushaltsvariablen werden die Variablen HA7000d und HA7000e weiterhin als Originalvariablen im SUF mitgeliefert.

GEN_Haushalt

Kurzbeschreibung:

Aktuelle oder plausible Angaben vorausgegangener Wellen zu der Anzahl der im Haushalt lebenden Personen einschließlich der befragten Person selbst.

Originalvariablen:

- HA1000 - HA1200: Wie viele Personen leben ständig in Ihrem Haushalt, Sie selbst eingeschlossen?
- HA7000a-HA7000e, HA2100: Bitte geben Sie alle zutreffenden Kategorien an. Aktuell leben Sie in Ihrem Haushalt...

Generierung:

Die Variable GEN_Haushalt basiert auf der Variable HA1000/ HA1100/ HA1200 (Haushaltsgröße), die in Welle 1 allen teilnehmenden Personen, in Welle 5 der Auffrischstichprobe und in Welle 6 erneut allen Personen gestellt wurde. Da die Haushaltsgröße HA1100 ab Welle 5 gefiltert nur noch Personen gestellt wird, die neben einem Partner weitere im Haushalt lebende Personen im Haushalt angeben, werden fehlende Werte in Welle 5 und 6 auf Basis der Filterfragen aufgefüllt: Allein-lebende Personen, identifiziert durch die Variable HA7000a, erhalten den Wert „1“, Personen in Partnerschaften, die durch die Variable HA2100 identifiziert werden, erhalten den Wert „2“.

Übertragungen von Werten der GEN_Haushalt auf Folgewellen, in denen die Frage *nicht* gestellt oder beantwortet wurde, erfolgt in Abstimmung mit Angaben zu im Haushalt lebenden Personen, erfasst mit den Variablen HA7000a bis HA7000e sowie HA2100. Diese Angaben wurden in Welle 2, Welle 3 und Welle 6 allen Personen gestellt und in direkt anschließenden Wellen nur jenen Befragten, die die Antwort versäumt hatten oder zur Auffrischstichprobe in Welle 5 zählten. Die Werte der Variable GEN_Haushalt werden nur dann in Folgewellen übertragen, wenn sich die Angaben der kategorialen Variablen (HA7000a bis HA7000e und HA2100) im Vergleich zu vorherigen Angaben nicht geändert haben. Sobald eine Veränderung in der Haushaltszusammensetzung durch veränderte Angaben in den Variablen HA7000a bis HA7000e sowie HA2100 angegeben wurde, wird die GEN_Haushalt auf „-7“ gesetzt.

Beispiel:

Eine befragte Person gibt an, mit einem Partner oder einer Partnerin sowie zwei Kindern zusammen zu leben, die beide noch nicht volljährig sind. Entsprechend gibt die Person in Welle 1 vier im Haushalt lebende Personen (HA1000) an. Diese Information wird in die Variable GEN_Haushalt geschrieben. In Welle 2 bestätigt die befragte Person ihre Angaben, indem sie angibt, mit einem Partner oder einer Partnerin sowie Kindern unter 18 Jahren zusammen zu leben (HA2100 und HA7000c). Somit erhält die Person auch in Welle 2 in der Variable GEN_-Haushalt den Wert „4“. In Welle 3 ändern sich jedoch die Angaben. Ein Partner wird nicht mehr

angegeben. Daher wird die Fortschreibung entsprechend beendet und mit „-7“ kenntlich gemacht. Erst in Welle 6 wird die Person erneut regulär zur Haushaltsgröße befragt (Variable HA1200). Durch diese neue Angabe kann dann auch die Variable GEN_Haushalt wieder aktualisiert und in Welle 7 fortgeschrieben werden.

GEN_GebKinda-GEN_GebKindd und IMP_GebKind

Kurzbeschreibung:

Geburtsjahr der vier jüngsten im Haushalt lebenden Kinder, beginnend mit dem jüngsten Kind.

Originalvariablen:

- HA8000a-d: In welchem Jahr wurde das Kind/wurden die Kinder geboren? Bitte geben Sie für jedes Kind unter 18 Jahren in Ihrem Haushalt das Geburtsjahr an (vier Ziffern). Wenn Sie mehr als vier Kinder haben, geben Sie bitte die Geburtsjahre für die jüngsten vier Kinder an.

Generierung:

Die Fragen HA8000a-d wurden allen teilnehmenden Personen regulär in Welle 2, Welle 3 und Welle 6 gestellt. Ab Welle 4 wurden nur von jenen Befragten Angaben eingeholt, die in den Wellen zuvor keine Angaben gemacht hatten, was auch alle neu-befragten Personen in Welle 5 betraf (Auffrischer). Bei der wiederholten Abfrage wurden Kinder teilweise in einer veränderten Reihenfolge oder mit anderen Geburtsjahren eingetragen, z.B. wurde diesmal das älteste Kind als erstes genannt oder ein Kind mehr oder weniger eingetragen. Entsprechend wurden die Geburtsjahre der Kinder pro Nennung sortiert, so dass in Variable GEN_GebKinda das jüngste und letztgeborene Kind, in Variable GEN_GebKindb das zweitjüngste Kind usw. erfasst ist und keine Lücken in der Variable entstehen. In einem weiteren Schritt wurden Kinder, die während der Erhebung geboren wurden, auf alle vorherigen Wellen rückwirkend übertragen, um den Datenschutz der befragten Person zu wahren.[7]

Soweit möglich wurden zudem inkonsistente Nennungen korrigiert. Da Personen maximal in drei Wellen Angaben zu den im Haushalt lebenden Kindern machen konnten, wurde eine abweichende Nennung korrigiert, wenn zwei weitere Nennungen identisch und plausibel waren. Dabei wurden jeweils mit einer Ausnahme[8] nur Abweichungen von einem Jahr früher

[7] Ansonsten könnten Haushalte leicht identifiziert werden, deren Kinder in einem bestimmten Monat geboren wurden.

[8] In diesem Fall wurden einmal vier Kinder und zwei mal zwei Kinder eingegeben, wobei deutlich war, dass sich die Person bei den Angabe der vier Kinder bei einer Ziffer vertippt hatte und statt den Fehler zu korrigieren, die Geburtsdaten erneut eingegeben hatte

oder später korrigiert.

Beispiel:

Eine Person gibt in Welle 2 und 6 ein Kind in mit dem Geburtsjahr „2019“, ein Kind mit dem Geburtsjahr „2017“ und ein Kind mit dem Geburtsjahr „2020“ an – in Welle 3 verändert sich jedoch die Zahlenreihe in „2018“, „2017“ und „2020“. Der Wert „2018“ wird in Welle 3 durch den Wert „2019“ ersetzt. Beibehalten werden jedoch Veränderungen, die größere Abweichungen beschreiben oder neue Nennungen bei Geburtsjahren beinhalten.

Wurden Werte manuell im Rahmen der Datenaufbereitung angepasst oder verändert, ist dies in der Variable IMP_GebKind gekennzeichnet.

Die Übertragung der Werte in Vor- und Folgewellen, in denen die Frage nicht gestellt wurde, erfolgte immer für die komplette Nennung der Variablengruppe GEN_GebKinda-GEN_GebKindd. Dabei gilt, dass komplett neue Nennungen nie durch vorherige Angaben aufgefüllt werden, wenn etwa ein Kind weniger genannt wird.

GEN_selbst

Kurzbeschreibung:

Die Variable GEN_selbst enthält die Information, inwiefern eine Person zum Befragungszeitpunkt selbstständig erwerbstätig ist. Dies umfasst sowohl soloselbstständige Personen als auch Personen mit abhängig Beschäftigten im eigenen Unternehmen.

Originalvariablen:

- EW1000c: Tragen Sie eine Beschäftigung auch dann ein, wenn Sie aufgrund der Corona-Krise weniger arbeiten, ihre Beschäftigung aber weiterbesteht: Selbstständig
- EW1100c: Tragen Sie eine Beschäftigung auch dann ein, wenn Sie aufgrund der Corona-Krise weniger arbeiten, Ihre Beschäftigung aber weiterbesteht: Selbstständig, mit Beschäftigten
- EW1000n: Tragen Sie eine Beschäftigung auch dann ein, wenn Sie aufgrund der Corona-Krise weniger arbeiten, Ihre Beschäftigung aber weiterbesteht: Selbstständig, ohne Beschäftigte (Solo-Selbstständig)

Generierung:

In Welle 1-5 wird diese Angabe direkt aus der Variable EW1000c generiert. Ab Welle 6 wurde

die Abfrage differenziert in die Kategorien Selbstständigkeit mit Beschäftigten (EW1100c) und ohne Beschäftigte (EW1000n). Entsprechend trifft die GEN_selbst zu, wenn entweder die Variable EW1100c oder EW1000n positiv beantwortet wurde. Die Variable EW1000c wird im SUF nicht mehr ausgeliefert.

GEN_konsum

Kurzbeschreibung: Die Variable GEN_konsum enthält harmonisierte Angaben auf die Frage, wie viel eine Person anteilig sofort ausgeben oder sparen würde, falls sie unerwartet einen Geldbetrag in Höhe ihres derzeitigen Nettoeinkommens erhalten würde.

Originalvariablen:

- KN1000: Stellen Sie sich vor, Sie erhalten unerwartet einen Geldbetrag in Höhe Ihres monatlichen Netto-Haushaltseinkommen. Wie viel davon würden Sie im nächsten Monat ausgeben? Antworten Sie bitte anhand der folgenden Skala, bei der 0 „gar nichts ausgeben“ und 10 „alles ausgeben“ bedeutet.
- KN1100: Stellen Sie sich vor, Sie erhalten unerwartet einen Geldbetrag in Höhe Ihres monatlichen Netto-Haushaltseinkommen. Wie viel davon würden Sie im nächsten Monat ausgeben? Antworten Sie bitte anhand der folgenden Skala, bei der 0 „gar nichts ausgeben“ und 10 „alles ausgeben“ bedeutet.

Generierung:

Die Variable GEN_konsum fasst die Variablen KN1000 (Welle 3) und KN1100 (ab Welle 4) zusammen. In der KN1000 wurde randomisiert entweder eine Skala von 0 bis 10 präsentiert, in der ausschließlich die Enden beschriftet waren, oder eine Skala, in der zusätzlich die Mitte beschriftet war. Ab Welle 4 wurde die in der Mitte beschriftete Skala für alle Befragten beibehalten.

4.5.3 Randomisierte Variablen

Während der Befragung wurden wiederholt einzelne Fragen oder Fragegruppen randomisiert, z.B. hinsichtlich der Position im Fragebogen oder der Frageformulierung. Für alle randomisierten Fragen wurden Indikatorvariablen bereitgestellt, die der Variablenübersicht entnommen werden können. Hierzu zählen:

IND_AS2000
: Kennzeichen, das angibt, ob die Fragegruppe AS2* zu Beginn oder am Ende des Fragebogens gestellt wurde.

IND_AS3000

Kennzeichen, das angibt, ob die Frage AS3000 zu Beginn oder am Ende des Fragebogens gestellt wurde.

IND_MK1000

Kennzeichen, das angibt, welche Fragegruppen einer Person gezeigt wurde.

IND_KN1000

Kennzeichen, das angibt, ob die Variable KN1000 mit oder ohne Hinweistext bei Antwort 5 gezeigt wurde. Später wurde eine einheitliche Kennzeichnung gewählt (KN1100). Die Variablen KN1* gehen daher in die GEN_konsum ein.

4.6 Übertragung von Werten auf Folgewellen (Fortschreibung)

Bestimmte Variablen wurden im Befragungszeitraum nur zu einem Zeitpunkt erfasst. Es handelt sich dabei in der Regel entweder um

- zeitlich konstante Merkmale,
- Angaben, die sich auf den Zeitraum vor Beginn der Pandemie im März 2020 beziehen, oder
- Merkmale, die sich nur sehr selten verändern.

Dazu zählen etwa Variablen wie das Geschlecht und das Geburtsjahr, aber auch die aktuelle Wohnsituation einer Person. Da derartige Informationen für Analysen in allen Wellen relevant sind, werden diese Merkmale sowohl in vorangegangene Wellen als auch Folgewellen übertragen. Die so generierten Variablen werden im Variablennamen durch das Hochzählen der letzten Nullstelle gekennzeichnet (vergleiche hierzu auch Abschnitt 4.3). Entsprechend wird etwa die Frage nach dem Geschlecht einer Person SD1000 als SD1001 ausgewiesen, um hervorzuheben, dass es sich um eine fortgeschriebene Variable handelt und das Merkmal ggf. zu einem anderen Zeitpunkt erhoben wurde. Die Originalvariablen werden entsprechend nicht mehr zusätzlich ausgeliefert. Die tatsächlichen Messzeitpunkte und ursprünglichen Bezeichnungen können der Tabelle 4 auf Seite 27 entnommen werden. Darüber hinaus werden auch die meisten der generierten Variablen fortgeschrieben.[9] Bei der Fortschreibung von Merkmalen, die in mehreren Wellen abgefragt wurden, wird die Variable aktualisiert, sobald eine Personen in einer der Folgewellen eine von der bisherigen Antwort abweichende Antwort angibt. Fehlende Werte werden hingegen durch Angaben aus vorherigen Wellen ersetzt. Aus der Variablenübersicht beziehungsweise aus Tabelle 4 kann der Zeitpunkt entnommen werden, zu dem ein Merkmal regulär abgefragt wurde. Durch die Fortschreibung der Variablen ist es bei

[9] In welchen Wellen generierte Variablen ausgewiesen werden, kann der Variablenübersicht entnommen werden, die in einer gesonderten Excel-Datei auf der Seite des FDZs verfügbar ist.

Personen, die die Antwort auf mehrfach erhobene Variablen verweigert haben, nicht mehr möglich zu identifizieren, aus welcher Welle die ursprüngliche Angabe stammt. Beispielsweise wurde die Frage HA5000 für Panelisten aus Welle 1 zu zwei und für Panelisten aus Welle 5 nur zu einem Zeitpunkt erhoben (vergleiche Tabelle 4 auf Seite 27). In Welle 6 basiert der fortgeschriebene Wert in der Variable HA5001 für Panelisten aus Welle 5 auf Angaben aus Welle 5. Für Personen, die bereits ab Welle 1 teilgenommen haben, wurde die Angabe mit hoher Wahrscheinlichkeit aus Welle 4 übernommen. Falls die Person hingegen in Welle 4 die Frage nicht beantwortet hat, wurde der Wert aus Welle 1 fortgeschrieben.

Tabelle 4: Fortgeschriebene Variablen: Originalzeitpunkt Erfassung

			Welle						
Neu	**Variablenlabel**	**Variablenname**	1	2	3	4	5	6	7
AT1001a	Arbeitsteilung vor Corona-Krise: Hausarbeit	*AT1000a*		x					
AT1001b	Arbeitsteilung vor Corona-Krise: Einkaufen	*AT1000b*		x					
AT1001c	Arbeitsteilung vor Corona-Krise: Reparaturen	*AT1000c*		x					
AT1001d	Arbeitsteilung vor Corona-Krise: Finanzen und Behörden	*AT1000d*		x					
AT1001e	Arbeitsteilung vor Corona-Krise: Kinderbetreuung	*AT1000e*		x					
AZ1001	Arbeitszeit vor Corona, pro Woche	*AZ1000, AZ1100*	x				x		
HA3001	Anzahl Kinder unter 18 Jahren im Haushalt	*HA3000, HA3100, HA3200*	x				x	x	
HA5001	Monatliches Netto-Haushaltseinkommen (Kategorien)	*HA5000*	x			x	x		
HA7001d	Haushaltszusammensetzung: Mit anderen verwandten Personen	*HA7000d*		x	x			x	
HA7001e	Haushaltszusammensetzung: Mit anderen nicht verwandten Personen	*HA7000e*		x	x			x	
HO2001	Arbeitszeit Home-Office vor Corona, pro Woche	*HO2000, HO2100*	x				x		
KB1001	Kinderbetreuung vor Corona-Krise: Ganztags	*KB1000*		x					
KB5001_A	Orga Kinderbetreuung vor Corona-Krise (Version mit Partner)	*KB5000_A*		x					
KB5001_B	Orga Kinderbetreuung vor Corona-Krise (Version ohne Partner)	*KB5000_B*		x					
SD1001	Geschlecht	*SD1000*	x				x		
SD2001	Geburtsjahr	*SD2000*	x				x		
SD3001	Höchster allgemeinbildender Schulabschluss	*SD3000*	x				x		
SD4001	Höchster beruflicher Ausbildungsabschluss	*SD4000*	x				x		

Fortführung Tabelle 4

Neu	Variablenlabel	Variablenname	Welle						
			1	2	3	4	5	6	7
SD5001	Art des Hochschulabschlusses	*SD5000*	x				x		
SD6001	In Deutschland geboren	*SD6000*	x				x		
SD7001	Zuzugsjahr	*SD7000*	x				x		
WS1001	Wohnfläche	*WS1000*		x					
WS2001a	Nutzung von zusätzlichen Flächen: Balkon/Terrasse	*WS2000a*		x					
WS2001b	Nutzung von zusätzlichen Flächen: Hof	*WS2000b*		x					
WS2001c	Nutzung von zusätzlichen Flächen: Garten	*WS2000c*		x					
WS2001d	Nutzung von zusätzlichen Flächen: Nichts von alledem	*WS2000d*		x					
WS3001	Wohnort Bundesland	*WS3000*						x	
Z1001	Zuspielbereitschaft	*Z1000*	x				x		

4.7 Einschränkung ausgelieferter Fälle im SUF

4.7.1 Definition vollständiger Fälle

Im Scientific Use File werden nur Interviews ausgeliefert, die als inhaltlich vollständig markiert sind. Ein Interview gilt als vollständig, wenn die befragte Person alle als inhaltlich relevant festgelegten Fragen gesehen hat, auch wenn diese ggf. nicht beantwortet wurden. Zu den inhaltlich nicht relevanten Fragen zählen weitere ergänzende Angaben oder Angaben, welche für die Feldsteuerung benötigt werden. Solche Angaben wurden immer am Ende der Befragung platziert. Die folgende Übersicht listet die Fragen auf, bei denen keine Antwort mehr notwendig ist und Interviews ggf. bei einem Abbruch an dieser Stelle dennoch als vollständig definiert werden:

- **Welle 1-2, 5-7**
  AN1000 (Anmerkungen der Befragungsperson): „Wir wollten die Befragung für Sie so kurz wie möglich halten und konnten wahrscheinlich nicht alle Themen aufgreifen, die Sie betreffen bzw. zu denen Sie uns gerne Auskunft gegeben hätten. Wenn Sie uns noch etwas mitteilen möchten, schreiben Sie es bitte in das untenstehende Textfeld."
- **Welle 3:**
  Z5000 (500 Punkte Annahme): „[…] Zukünftig möchten wir Sie zu Befragungen über eine neue Befragungsseite einladen. Damit Sie uns weiter unterstützen können, bitten wir Sie, sich auf der Seite zu registrieren. Mit jeder abgeschlossenen Befragung bekommen Sie 200 Punkte gutgeschrieben, dabei entsprechen 100 Punkte einem Euro. Sobald Sie

500 Punkte gesammelt haben, können Sie diese in Gutscheine von 19 verschiedenen Anbietern umwandeln (z.B. Karstadt, Thalia, Media Markt …). […]“

- **Welle 4:**
  Z5100 (500 Punkte Annahme): „[…]Zukünftig möchten wir Sie zu Befragungen über eine neue Befragungsseite einladen. Damit Sie uns weiter unterstützen können, bitten wir Sie, sich auf der Seite zu registrieren. Mit jeder abgeschlossenen Befragung bekommen Sie 200 Punkte gutgeschrieben, dabei entsprechen 100 Punkte einem Euro. Sobald Sie 500 Punkte gesammelt haben, können Sie diese in Gutscheine von 19 verschiedenen Anbietern umwandeln (z.B. Karstadt, Thalia, Media Markt …). […]“

4.7.2 Plausibilisierung von Fällen

Die im SUF enthaltenen Fälle wurden durch die in den Integrierten Erwerbsbiografien (IEB) enthaltenen Angaben zu Alter und Geschlecht plausibilisiert. Personen, die ein anderes Geschlecht als das in den administrativen Daten gelistete angegeben haben oder bei denen das angegebene Alter um mehr als drei Jahre von dem Alter in den administrativen Daten abweicht, sind im SUF nicht mehr enthalten.

Darüber hinaus wurde geprüft, ob Befragte erst mit deutlichem Zeitabstand zur Einladung an der Befragung teilgenommen haben oder ggf. schon vor dem Versand einer Einladung an der Befragung teilnehmen konnten. In wenigen Fällen kam es z.B. vor, dass eingeladene Personen der Welle 1 erst ein halbes Jahr später teilnahmen. Während der Wellen 1-4 konnten Personen bis zu 23 Tage nach Einladung an der Befragung teilnehmen. Mit der Erhöhung des Abstandes zwischen zwei Wellen auf zwei Monate ab Welle 5 können Personen bis zu 52 Tage nach Einladung teilnehmen. Dadurch wurde sichergestellt, dass sich die Beantwortung der Wellen für Befragungspersonen nicht überschneiden und z.B. Welle 4 nicht vor Welle 3 beantwortet wird. Außerdem ist damit sichergestellt, dass Angaben innerhalb einer Welle sich auf ähnliche Zeitperioden beziehen.

Bei postalisch eingeladenen Personen konnte es beim Abtippen des individualisierten Einwahlcodes in seltenen Fällen zu Tippfehlern kommen, wodurch fälschlicherweise der Einwahlcode einer anderen Person verwendet wurde, beispielsweise einer Person aus einer anderen Teilstichprobe. Dadurch kam es zu Fällen, die vor der geplanten Einladung zur Befragung schon an der Befragung teilgenommen haben. Diese Fälle wurden aus dem SUF entfernt.

4.8 Fragenprogramm

Das Fragenprogramm wurde überwiegend durch das IAB, aber auch in Zusammenarbeit mit externen Forscherinnen und Forschern entwickelt. Alle Beteiligten werden im Kopf der Fragebögen der einzelnen Wellen namentlich genannt. Einzelne Fragen wurden aus anderen Studien übernommen (Tabelle 5). Die Fragebögen können unter https://fdz.iab.de/de/FDZ_Individual_Data/HOPP.aspx heruntergeladen werden.

Nicht alle Fragen aus dem Fragebogen wurden unverändert in den SUF übernommen, sondern bei Bedarf für die Nutzung aufbereitet. Alle Änderungen zwischen Fragebögen und SUF können der Variablenliste entnommen werden (siehe https://fdz.iab.\\de/de/FDZ_Individual_Data/HOPP.aspx).

Tabelle 5: Fragen aus anderen Studien

Variable	Welle	Quellen
AS2000, AS3000	1-5, 7	Sozio-ökonomisches Panel (GSOEP)[a]
GS2000	2-5	German Internet Panel (GIP)[b]
GS3000	2-6	German Internet Panel (GIP)[b]
AT2000	2-5	Beziehungs- und Familienpanel (pairfam)[c]
WL1000a-c	5	Panel Arbeitsmarkt und Soziale Sicherung (PASS)[d]
GS6000a-1	6-7	Sozio-ökonomisches Panel (SOEP)[e]

[a] https://www.diw.de/en/soep
[b] https://www.uni-mannheim.de/gip/das-gip/
[c] https://fdz.iab.de/de/FDZ_Individual_Data/PASS.aspx
[d] https://www.pairfam.de/
[e] Nübling, M., Andersen, H. H., and Mühlbacher, A. (2006). Entwicklung eines Verfahrens zur Berechnung der körperlichen und psychischen Summenskalen auf Basis der SOEP-Version des SF 12 (Algorithmus). Retrieved from https://www.diw.de/documents/publikationen/73/diw_01.c.44987.de/diw_datadoc_2006-016.pdf

Im SUF von HOPP fehlen zum Zeitpunkt der Veröffentlichung einige Variablen aus Welle 6 und 7 der Befragung. Es handelt sich hierbei um ein Modul zum unintendierten Gebrauch von Kurzarbeit, das von einer Gruppe von Forschenden innerhalb des IAB erstellt wurde, und umfasst die Variablen MK1000-MK7000. Wegen der besonderen inhaltlichen Sensibilität der Fragestellung werden die entsprechenden Variablen erst nach interner Überprüfung und Analyse am 01.10.2021 veröffentlicht.

5 Gewichtung

5.1 Beschreibung Gewichtungskonzept

Die Gewichtung der HOPP-Daten setzt sich aus drei Faktoren zusammen:

1. **Der Inklusionswahrscheinlichkeit** π, d.h. der Wahrscheinlichkeit, in die Bruttostichprobe S_g der Größe n_g zu gelangen,
2. **Der Teilnahmewahrscheinlichkeit** p_t, d.h. der Wahrscheinlichkeit, Teil von S_{nt} , der Nettostichprobe von Welle t mit Umfang n_{nt} zu sein, gegeben der Inklusion in die Bruttostichprobe,
3. **Der Kalibrierung**; Dabei wird die Wahrscheinlichkeitsgewichtung, also die Inverse des Produkts aus Inklusions- und Teilnahmewahrscheinlichkeit, auf die Verteilung der Schichtvariablen und Angaben zur letzten Beschäftigung, im Ziehungsrahmen, also der IEB, angepasst.

Die Gewichtungsschritte 2 und 3 werden für jede Welle jeweils für alle Personen in der Nettostichprobe und noch einmal nur für die Personen mit Record Linkage Consent (RLC) durchgeführt. Das bedeutet, es werden Gewichte getrennt für zwei Gruppen berechnet, einmal für alle Personen des SUF und einmal für alle Personen des SUF mit RLC.

5.1.1 Wahrscheinlichkeitsgewichtung

Für Welle 1 lässt sich die Wahrscheinlichkeitsgewichtung p_{1k}, für alle Personen des SUF, sowie p_{RLC1k} für die Personen mit RLC des SUF, wie folgt beschreiben:

$$p_{1k} = (\pi_k \hat{\theta}_{1k})^{-1}$$

$$p_{RLC1k} = (\pi_k \hat{\theta}_{RLC\,k})^{-1}$$

wobei $\pi_k = Pr(k \in S_g)$, d.h. π_k ist die Inklusionswahrscheinlichkeit des k-ten Elements, θ_{1k} ist eine Schätzung der Teilnahmewahrscheinlichkeit $\theta_{1k} = Pr(k \in S_{n1})$, mit S_{n1} als der Welle 1 Nettostichprobe, d.h. $S_{n1} \subset S_g$. Des Weiteren ist $\hat{\theta}_{RLC\,1k}$ ein Schätzer für $\theta_{RLC\,k} =$

$Pr(k \in S_{nRLC\,1})$), $S_{n1} \subset S_{n1RLC}$, die Teilnahmewahrscheinlichkeit des k-ten Elements mit RLC und $S_{n1RLC\,1}$ als der Nettostichprobe der ersten Welle für Personen mit RLC.

Um θ_{1k} zu schätzen wird eine generalisierte additive Regression mit einer logit-Linkfunktion verwendet. Das verwendete Modell hat die folgende Form:

$$g(\theta_{1k}) = \beta_0 + X_{IEB}\beta + \sum_{j=1}^{m_{lIBE}} f_i(Z_{IEB\,j})$$

dabei ist $g = \log(\frac{p}{1-p})$ (die logit-Funktion), β_0 ist ein Skalar, X_{IEB} ist eine $n_g \times m_{dIEB}$ Matrix mit der linearen Prädiktoren (kategoriale Variablen) aus der IEB, β ein $m_{dIEB} \times 1$ Vekor mit Regressionkoeffizienten, $Z_{IEBj},\ j = 1,\ ...,\ m_{lIBE}$ sind glatte Prädiktoren (stetige Variablen) und $f_j,\ j = 1,\ ...,\ m_{lIBE}$ sind glatte Funktionen, deren Parametrisierung spezifiziert werden kann. Das Modell zur Schätzung von $\theta_{RLC\,k}$ ist analog zu dem Modell für θ_{1k}.

Für alle weiteren Wellen , d.h. $t > 1$, stellt sich die Wahrscheinlichkeitsgewichtung für alle Personen des SUF, p_{tk}, wie folgt dar:

$$p_{tk} = (\pi_k \hat{\theta}_{1k} \hat{\theta}_{PC\,k} \hat{\theta}_{tk})^{-1}$$

Dabei ist $\hat{\theta}_{PCk}$ ein Schätzer für $\theta_{PCk} = Pr(k \in S_{nPC1})$, die Wahrscheinlichkeit des $k-ten$ Elements in der Nettostichprobe von Welle 1 zu sein und $Panel\ Consent\ (PC)$ zu geben. Das Modell zur Schätzung von $\theta_{PC\,k}$ ist analog zu dem Modell für θ_{1k} bzw. $\theta_{RLC\,k}$.θ_{tk} ist ein Schätzer für $\theta_{tk} = Pr(k \in S_{nt} | k \in S_{npc1})$, die Wahrscheinlichkeit Teil der Nettostichprobe in Welle t zu sein, unter der Bedingung auch Teil der Nettostichprobe von Welle 1 mit PC zu sein. Für Schätzer θ_{tk} gibt es eine Fallunterscheidung, ob Element k RLC gegeben hat oder nicht, d.h.

$$\hat{\theta}_{tk} = \begin{cases} f_{RLC}(D_{IEB}, D_{svy\,1}, D_{msvy\,t-1}) \text{ für } k \in S_{nRLC1}, \\ f_{all}(D_{svy\,1}, D_{msvy\,t-1}) \text{ für } k \notin S_{nRLC1} \end{cases}$$

dabei ist D_{IEB} eine $n_{npc1} \times m_{IEB}$ Matrix mit m_{IEB} Prädiktoren aus der IEB. Die Matrizen $D^{'}_{svy\,1}$ und $D_{svy\,1}$ haben die Dimensionen $n_{npc1} \times m_{svy1}$ bzw. $n_{npc1} \times m^{'}_{svy\,1}$ und enthalten $m_{svy\,1}$ bzw. $m^{'}_{svy\,1}$ Prädiktoren aus der Welle 1 Befragung. Für t > 2 hat Matrix $D_{msvy\,t-1}$ die Dimension $n_{npc1} \times m^{'}_{msvy\,t-1}$ und enthält Metadaten aus der Kontakt und Responsehistorie der Elemente $k \in S_{npc1}$. Ist t=2 enthält $D_{msvy\,t-1}$ keine Variablen. Die Funktionen f_{RLC} und f_{all} sind Modelle der Art wie sie für die Schätzung von θ_{1k} , $\theta_{RLC\,k}$ und $\theta_{PC\,k}$ verwendet werden.

Sie unterscheiden sich untereinander dadurch, dass f_{RLC} mit IEB- und Befragungsdaten aus der ersten Welle und f_{all} nur mit und Befragungsdaten aus der ersten Welle geschätzt wird.

Das Wahrscheinlichkeitsgewicht $p_{RLC\,tk}$ nur für Personen mit RLC in Wellen t>1, stellt sich wie folgt dar:

$$p_{RLC\,tk} = (\pi_k \hat{\theta}_{RLC\,k} \hat{\theta}_{PC\,RLC\,k} \hat{\theta}_{RLC\,tk})^{-1}$$

Dabei ist $\hat{\theta}_{PC\,RLC\,k}$ ein Schätzer für $\hat{\theta}_{RLC\,PC\,k} = Pr(k \in S_{nPC\,1} | k \in S_{nRLC\,1})$, die Wahrscheinlichkeit des $k-ten$ Elements in der Nettostichprobe von Welle 1 zu sein und Panel Consent (PC) zu geben, unter der Bedingung auch RLC gegeben zu haben. Das Modell zur Schätzung von $\theta_{PC\,RLC\,k}$ ist analog zu dem Modell für θ_{1k} bzw. $\theta_{RLC\,k}$. Des Weiteren ist $\hat{\theta}_{RLC\,tk}$ ein Schätzer für $\theta_{RLC\,tk} = Pr(k \in S_{nt} | \{S_{nRLC\,1} \cap S_{nPC\,1}\})$, die Wahrscheinlichkeit Teil der Nettostichprobe in Welle t zu sein, unter der Bedingung auch Teil der Nettostichprobe von Welle 1 mit PC und RLC zu sein. Für Schätzer $\hat{\theta}_{RLC\,tk}$ wird, in Analogie zu Schätzer $\hat{\theta}_{tk}$, folgendes Modell verwendet:

$$\hat{\theta}_{RLC\,tk} = f_{RLC}(D_{RLC\,IEB}, D'_{RLC\,svy\,1}, D_{RLC\,msvy\,t-1})$$

Matrizen $D_{RLC\,IEB}$, $D'_{RLC\,svy\,1}$ und $D_{RLC\,msvy\,t-1}$ enthalten dabei die gleichen Variablen wie D_{IEB}, $D'_{svy\,1}$ und $D_{msvy\,t-1}$ haben jedoch nur Einträge für Elemente $k \in \{S_{nRLC\,1} \cap S_{nPC\,1}\}$.

5.1.2 Kalibrierung

Der dritte Faktor der Gewichtung ist die Kalibrierung oder Anpassung der Wahrscheinlichkeitsgewichte auf Randverteilungen von Variablen im Ziehungsrahmen. Das kalibrierte Wahrscheinlichkeitsgewichte in Welle t für allem Elemente im SUF stellt sich wie folgt dar:

$$w_{tk} = p_{tk} c_{tk}$$

wobei c_{tk} der Kalibrierungsfaktor des Wahrscheinlichkeitsgewichts p_{tk} ist. Dabei wird c_{tk} für alle $k \in S_{nt}$ so gewählt, dass gilt:

$$\sum_{k \in S_{nt}} w_k h_k \approx N$$

mit N als $H \times 1$ Vektor mit den Totalwerten im Ziehungsrahmen folgender Variablen aus der IEB:

Name	Bedeutung
erw_alt_sex	Kreuzkombination aus Erwerbsstatus, Alterskategorie, und Geschlecht wie es auch bei der Schichtung der Welle 1 Bruttostichprobe gesetzt wurde
bula	Bundesland
beruf_last	Berufsklassifikation der letzten SV-Beschäftigung
pos_last	Position in der letzten SV-Beschäftigung
sektor_last	Sektor des Betriebs mit der letzten SV-Beschäftigung

Zur Berechnung von c_{tk} wird eine lineares Kalibrierungsverfahren verwendet (Deville/Särndal, 1992). Ziel ist es dabei c_{tk} für alle $k \in S_{nt}$ so nahe bei 1 zu halten wie möglich, unter der Bedingung, dass die oben genannte Kalibrierungsbedingung erfüllt wird. Für die Gewichtung nach Erhebungsmonaten werden die Wahrscheinlichkeitsgewichte der Personen in den entsprechen Teilstichproben kalibriert.

5.1.3 Integration der *zweiten* Rekrutierungsstichprobe (ab Welle 5)

Mit Welle 5 wurde die Stichprobe mit einer neuen Bruttostichprobe von 99.188 Personen aufgefrischt. Diese 99.188 Personen wurden zusammen mit den ersten Bruttostrichpobe von 200.000 Personen gezogen (siehe Kapitel 2). Die Wahrscheinlichkeit für ein Element k in die Welle 5 Nettostichprobe zu gelangen, lässt sich wie folgt beschrieben:

$$\theta_{5k} = \pi_{\text{rec1}\,k}\theta_{\text{rec1}\,1k}\theta_{\text{rec1}\,PC\,k}\theta_{\text{rec1}\,PIC\,k}\theta_{\text{rec1}\,5k} + \pi_{\text{rec2}\,k}\theta_{\text{rec2}\,5k} \cdot$$

Dabei bezeichnet *rec1* die Konditionierung auf $k \in S_{g\ \text{rec1}}$, mit $S_{g\ \text{rec1}}$ als Bruttostichprobe der ersten Rekrutierung. Entsprechend bezeichnet *rec2* die Konditionierung der auf $k \in S_{g\ \text{rec2}}$, mit $S_{g\ \text{rec2}}$ als Bruttostichprobe der zweiten Rekrutierung. Des Weiteren bezeichnet $\theta_{\text{rec1}\,PIC\,k}$ die Wahrscheinlichkeit sich für Panelingress zu registrieren (siehe Kapitel 3.1.3. Die Registration in Panelingress war eine Voraussetzung für Personen aus der ersten Rekrutierungsstichprobe für Welle 5 und alle Folgewellen eingeladen zu werden. Für Personen aus der zweiten Rekrutierungsstichprobe war die Registration notwendig, um zu Welle 6 und 7 eingeladen zu werden.)

Die Wahrscheinlichkeit für ein Element k in die Welle t > 5 Nettostrichprobe zu gelangen, lässt sich dann wie folgt beschreiben:

$$\theta_{tk} = \pi_{\text{rec1}\,k}\theta_{\text{rec1}\,1k}\theta_{\text{rec1}\,PC\,k}\theta_{\text{rec1}\,PIC\,k}\theta_{\text{rec1}\,tk} + \pi_{\text{rec2}\,k}\theta_{\text{rec2}\,5k}\theta_{\text{rec2}\,PIC\,k}\theta_{\text{rec2}\,tk} \quad \text{für } t > 5\ .$$

Die Wahrscheinlichkeiten bezüglich der *ersten* Rekrutierungsstichprobe, $\theta_{\text{rec1}\,.}$, für alle $k \in S_{g\,\text{rec2}}$, werden geschätzt durch eine Prädiktion basierend auf Modellen, die für die Personen aus der *ersten* Rekrutierungsstichprobe geschätzt wurden. Umgekehrtes gilt für die Wahrscheinlichkeiten bezüglich der *zweiten* Rekrutierungsstichprobe.

5.1.4 Variablen zur Schätzung von Response-, Record-Linkage-, und Panel-Consent-Wahrscheinlichkeiten

Matrix X_{IEB} enthält die folgenden lineare Prädiktoren der IEB:

Name	Bedeutung
sex	Geschlecht
ausbildung	Ausbildung
bula	Bundesland
deutsch	Deutsche Staatsangehörigkeit
algj	Jemals_ALG/ALHI/UHG_(SGB_III)_erhalten
alg2j	Jemals_ALG2_(SGB_II)_erhalten
mini_5d	Kein_Minijob_in_5-Jahres-Historie
alg2_5d	Kein_ALG2_in_5-Jahres-Historie
asu_5d	Keine_Arbeitsuchend_Meldung_in_5-Jahres-Historie
mass_5d	Keine_Arbeitsmarktpolitische_Massnahme_in_5-Jahres-Historie
emp_1d	Keine_Sozialversicherungspflichtige_Beschäftigung_in_1-Jahres-Historie
mini_1d	Kein_Minijob_in_1-Jahres-Historie
alg_1d	Kein_ALG_(SGB_III)_in_1-Jahres-Historie
alg2_1d	Kein_ALG2_(SGB_II)_in_1-Jahres-Historie
asu_1d	Keine_Arbeitsmarktpolitische_Massnahme_in_1-Jahres-Historie
post_last	Position in der letzten SV-Beschäftigung
sector_last	Sektor des Betriebs mit der letzten SV-Beschäftigung

Die nicht linearen Prädiktoren $Z_{IEB\,j}, j = 1, ..., m_{lIBE}$ sind folgende Variablen aus der IEB:

Name	Bedeutung
emp_5 | ausbildung	Sozialversicherungspflichtige_Beschäftigung_in_5-Jahres-Historie_in_Jahren_(anteilig),_gegeben_Ausbildung
mini_5	Mini-Job-Beschäftigung_in_5-Jahres-Historie_in_Jahren_(anteilig)
alg2_5	Dauer_ALG2_(SGB_II)_in_5-Jahres-Historie_in_Jahren_(anteilig)
age20 | sex	Alter_in_Jahren,_gegeben_Geschlecht
age20 | ausbildung	Alter_in_Jahren,_gegeben_Ausbildung
age20 | deutsch	Alter_in_Jahren,_gegeben_Deutsche_Staatsangehörigkeit

Matrix $D_{svy\,1}$ enthält die folgenden Variablen aus der Welle 1 Befragung:
GEN3000, SD1000, SD2000, SD3000, SD6000, SD4000, SD8000a, AS1000a, AS1000b, AS1000c, AS1000d, AS2000a, AS2000b, AS2000c, AS2000d, AS2000e, AS2000f, AS2000g, EW1000a, EW1000b, EW1000c, EW1000d, EW1000e, EW1000f, EW1000g, EW1000h, EW1000i, EW1000j, EW1000k, EW1000l, EW1000m

Matrix $D^{'}_{svy\,1}$ enthält die gleichen Variablen wie Matrix $D_{svy\,1}$, jedoch ohne Variablen SD1000, SD2000, SD3000, SD4000, SD6000, SD8000a.

Matrix D_{IEB} enthält die folgenden Prädiktoren aus der IEB:

Name	Bedeutung
sex	Geschlecht
ausbildung	Ausbildung
laa	Regionaldirektion_der_BA
deutsch	Deutsche_Staatsangehörigkeit
algj	Jemals_ALG/ALHI/UHG_(SGB_III)_erhalten
alg2j	Jemals_ALG2_(SGB_II)_erhalten
mini_5d	Kein_Minijob_in_5-Jahres-Historie
alg2_5d	Kein_ALG2_in_5-Jahres-Historie
asu_5d	Keine_Arbeitsuchend_Meldung_in_5-Jahres-Historie
mass_5d	Keine_Arbeitsmarktpolitische_Massnahme_in_5-Jahres-Historie
emp_1d	Keine_Sozialversicherungspflichtige_Beschäftigung_in_1-Jahres-Historie
mini_1d	Kein_Minijob_in_1-Jahres-Historie
alg_1d	Kein_ALG_(SGB_III)_in_1-Jahres-Historie
alg2_1d	Kein_ALG2_(SGB_II)_in_1-Jahres-Historie
asu_1d	Keine_Arbeitsmarktpolitische_Massnahme_in_1-Jahres-Historie
post_last	Position in der letzten SV-Beschäftigung
sector_last	Sektor des Betriebs mit der letzten SV-Beschäftigung
emp_5	Sozialversicherungspflichtige_Beschäftigung_in_5-Jahres-Historie_in_Jahren_(anteilig)
mini_5	Mini-Job-Beschäftigung_in_5-Jahres-Historie_in_Jahren_(anteilig)
alg2_5	Dauer_ALG2_(SGB_II)_in_5-Jahres-Historie_in_Jahren_(anteilig)
alg20	Alter_in_Jahren

Für $t > 2$ enthält Matrix $D_{msvy\,t-1}$ enthält die folgenden Metadaten:

Name	Bedeutung
w(2)_suf	Teil der Nettostichprobe in Welle 2
⋮	⋮
w(t-1)_suf	Teil der Nettostichprobe in Welle $t-1$
w(2)_pt	Partielles Interview in Welle 2
⋮	⋮
w(t-1)_pt	Partielles Interview in Welle $t-1$

5.2 Auslieferung der Gewichte

Die Gewichte werden in zwei gesonderten Datensätzen ausgeliefert:

- HOPP_Weights4Waves_W1-W7_v1.dta
- HOPP_Weights4Months_W1-W7_v1.dta

Um die Handhabung der Gewichte zu vereinfachen, wird hier kurz der Inhalt der Gewichtungsdatensätze beschrieben.

HOPP_Weights4Waves_W1-W7_v1.dta enthält die Gewichte um jede Welle auf die Population der IEB hochzurechnen. Der Datensatz besteht aus 15 Variablen: zwei Indikatoren, sechs Gewichten und sieben Kalibrierungsvariablen.

Die zwei Indikatoren sind die ID für den Panelisten (iab_id) und der Wellenindikator (welle), welcher angibt, auf welche Welle sich das Gewicht bezieht.

Die sechs Gewichte sind durch die folgenden Abkürzungen gekennzeichnet und enthalten folgende Informationen:

- ***dweight***: Designgewichte (Inverse Inklusionswahrscheinlichkeit der Bruttostichprobe)
- ***pweight***: Wahrscheinlichkeitsgewicht (Designgewicht mal Teilnahmewahrscheinlichkeit)
- ***cweight***: Kalibrierungsgewicht (Wahrscheinlichkeitsgewicht kalibriert auf Erwerbsstatus $\times$ Alter $\times$ Geschlecht, Bula, Beruf_last, Pos_last, Sektor_last)
- ***_all***: Gewichte mit dieser Endung beziehen sich auf alle Teilnehmenden einer Welle
- ***_rlc***: Gewichte mit dieser Endung beziehen sich auf alle Teilnehmenden einer Welle, die record linkage consent gegeben haben

Das Gewicht cweight_all liegt jeweils für alle Teilnehmenden einer Welle vor, d.h. wenn nur Befragungsdaten ausgewertet werden, sollte dieses Gewicht verwendet werden. Werden für die Auswertungen administrativen Daten zugespielt, sollte das Gewichte cweight_rlc verwendet werden, welches für die Gruppen der Personen mit Record Linkage Consent erstellt wurde. Bei der Auswertung nach Gruppen ist zu beachten, dass die Gewichte nur zuverlässige Schätzungen für Gruppenvariablen geben können, auf die die Gewichte kalibriert sind, d.h. Erwerbsstatus, Alter, Geschlecht, Bundesland, Beruf_last, die letzte berufliche Position (Pos_last), Sektor_last.

Durch das Studiendesign überlappen sich die Erhebungswellen teilweise und sind nur bedingt für eine Darstellung von Analysen im Zeitverlauf geeignet.
HOPP_Weights4Months_W1-W7_v1.dta enthält die Gewichte, um jeden Erhebungsmonat auf die Population der IEB hochzurechnen. Diese Gewichte eignen sich, um Aussagen über bestimmte Monate zu machen, z.B. die Darstellung der Homeofficequote vom Mai 2020 bis Februar 2021. Der Datensatz besteht aus 16 Variablen: drei Indikatoren, sechs Gewichten und sieben Kalibrierungsvariablen.

Die drei Indikatoren sind die ID für den Panelisten (iab_id), der Wellenindikator (welle) und der Monat, indem der Befragte geantwortet hat (month). Die letzte Indikatorvariable gibt an, auf welchen Monat sich das Gewicht bezieht. Nur Fälle mit einem gültigen Gewicht haben im selben Monat geantwortet wie sie eingeladen wurden und werden für die Monatsgewichte berücksichtigt.

Die sechs Gewichte und sieben Kalibrierungsvariablen beinhalten die gleichen Informationen wie im Datensatz HOPP_Weights4Waves_W1-W7_v1.dta.

Um die Gewichte and an den Scientific Use File zu spielen, sollten die Variablen iab_id und GEN1000 = welle verwendet werden.

Literatur

AAPOR (2016): Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys.

Berg, Marco; Cramer, Ralph; Dickmann, Christian; Gilberg, Reiner; Jesske, Birgit; Kleudgen, Martin; Beste, Jonas; Dummert, Sandra; Frodermann, Corinna; Schwarz, Stefan; Trappmann, Mark; Bähr, Sebastian; Coban, Mustafa; Friedrich, Martin; Gundert, Stefanie; Müller, Bettina; Teichler, Nils; Unger, Stefanie; Wenzig, Claudia (2020): Codebuch und Dokumentation des Panel 'Arbeitsmarkt und soziale Sicherung' (PASS) – Datenreport Welle 13. FDZ-Datenreport, 12/2020 (de). In: .

Deville, Jean-Claude; Särndal, Carl-Erik (1992): Calibration Estimators in Survey Sampling. In: Journal of the American Statistical Association, Bd. 87, Nr. 418, S. 376–382, URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475217.

Dorner, Matthias; Heining, J örg; Jacobebbinghaus, Peter; Seth, Stefan (2010): The sample of integrated labour market biographies. In: Journal of Contextual Economics: Schmollers Jahrbuch, Bd. 130, Nr. 4, S. 599–608.

Frodermann, Corinna; Schmucker, Alexandra; Seth, Stefan; vom Berge, Philipp (2021): Stichprobe der Integrierten Arbeitsmarktbiografien (SIAB) 1975–2019. FDZ Datenreport, 01/2021 (de). In: .

Haas, Georg-Christoph; Müller, Bettina; Osiander, Christopher; Schmidtke, Julia; Trahms, Annette; Volkert, Marieke; Zins, Stefan (2021): Development of a New COVID-19 Panel Survey: The IAB High-frequency Online Personal Panel (HOPP). In: Journal for Labour Market Research.

Impressum

FDZ-Datenreport Nr 04|2021

Veröffentlichungsdatum
30. Juni 2021

Herausgeber
Forschungsdatenzentrum
der Bundesagentur für Arbeit
am Institut für Arbeitsmarkt- und Berufsforschung
Regensburger Straße 104
90478 Nürnberg

Bezugsmöglichkeit dieses Dokuments
http://doku.iab.de/fdz/reporte/2021/DR_04-21.pdf

Dokumentation Version
HOPP_W01-W07_DE_v1_dok1, DOI: 10.5164/IAB.FDZD.2104.de.v1

Dataset Version
HOPP W01-W07 v1, DOI: 10.5164/IAB.HOPP_W01-W07.de.en.v1

Bezugsmöglichkeit aller Veröffentlichung der Reihe „FDZ-Datenreport“
https://fdz.iab.de/de/FDZ_Publications/FDZ_Publication_Series/FDZ-Datenreporte.aspx

Website
https://fdz.iab.de

Rückfragen zum Inhalt
Marieke Volkert
Telefon: 0911 179 2628
E-Mail: marieke.volkert@iab.de
Georg-Christoph Haas
Telefon: 0911 179 2919
E-Mail: georg-christoph.haas@iab.de