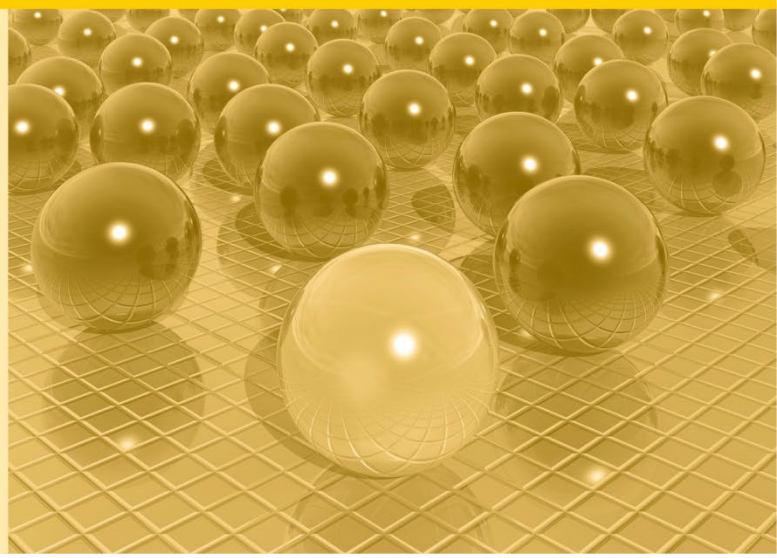
# Metadatenreport



Teil II: Produktspezifische Informationen zur Nutzung des Mikrozensus Scientific Use Files 2020

DOI: 10.21242/12211.2020.00.00.3.1.0

Version 1

gesis
Leibniz-Institut
für Sozialwissenschaften

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

**Impressum**

Herausgeber: Statistische Ämter des Bundes und der Länder
GESIS - Leibniz-Institut für Sozialwissenschaften, German Microdata Lab
Herstellung: Information und Technik Nordrhein-Westfalen
Telefon 0211 9449-01 • Telefax 0211 9449-8000
Internet: www.forschungsdatenzentrum.de
E-Mail: forschungsdatenzentrum@it.nrw.de

**Fachliche Informationen**
zu dieser Veröffentlichung:

Forschungsdatenzentrum der
Statistischen Ämter der Länder
– Düsseldorf –
Tel.: 0211 9449-2871
Fax: 0211 9449-8087
forschungsdatenzentrum@it.nrw.de

**Informationen zum Datenangebot**:

Statistisches Bundesamt
Forschungsdatenzentrum

Tel.: 0611 75-2420
Fax: 0611 75-3915
forschungsdatenzentrum@destatis.de

Forschungsdatenzentrum der
Statistischen Ämter der Länder
– Geschäftsstelle –
Tel.: 0211 9449-2883
Fax: 0211 9449-8087
forschungsdatenzentrum@it.nrw.de

Erscheinungsfolge: unregelmäßig
Erschienen im Mai 2023

Diese Publikation wird kostenlos als PDF-Datei zum Download unter www.forschungsdatenzentrum.de angeboten.



**Empfohlene Zitierung:**

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, Statistisches Bundesamt, GESIS - Leibniz-Institut für Sozialwissenschaften: Metadatenreport. Teil II: Produktspezifische Informationen zur Nutzung des Mikrozensus Scientific Use Files 2020 (EVAS-Nummer: 12211, 12231, 12241, 12251). Version 1. DOI: 10.21242/12211.2020.00.00.3.1.0. Düsseldorf 2023.

# Metadatenreport

Teil II: Produktspezifische Informationen zur Nutzung des Mikrozensus Scientific Use Files 2020

DOI: 10.21242/12211.2020.00.00.3.1.0

Version 1

## Inhalt

# Einleitung

Dieser Metadatenreport soll Forschenden dabei helfen, die Daten des Mikrozensus Scientific Use Files (MZ-SUF) 2020[1] sachgerecht auszuwerten. Er gibt einen Überblick über Datenaufbereitung, Dokumentation, Qualität, bereitgestellte Merkmale, Eckwerte relevanter Merkmale und praktische Hinweise.

Weitere Informationen zum Datenangebot und zum Datenzugang sind zudem auf den Seiten der Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder abrufbar. Hier findet sich u.a. der Metadatenreport Teil I Statistik, der diesen Metadatenreport mit allgemeinen und methodischen Informationen zum Mikrozensus ab dem Berichtsjahr 2020 ergänzt. Ausführliche Informationen und Auswertungshilfen zum MZ-SUF, u. a. Masterfragebogen, Datenhandbuch mit Randauszählungen, Tools zur Umsetzung sozialwissenschaftlicher Konzepte, Variablen-Zeitpunkte-Matrix, Verknüpfung von MZ-Querschnitterhebungen zu Panels, stehen auf dem Mikrodaten-Informationssystem (MISSY) der GESIS zur Verfügung.

Bei weiteren Fragen können sich interessierte Personen und Nutzende des Mikrozensus an das Forschungsdatenzentrum der Statistischen Ämter der Länder – Standort NRW (insbesondere bei Fragen zum Datenzugang und zur Datenaufbereitung), das Forschungsdatenzentrum des Statistischen Bundesamtes und an das German Microdata Lab der GESIS (insbesondere bei inhaltlichen Fragen und Fragen zum Angebot in MISSY) wenden.

Neben der Nutzung der Scientific-Use-Files am eigenen Arbeitsplatz in der wissenschaftlichen Einrichtung gibt es auch die Möglichkeit, weniger stark anonymisierte Datensätze an einem Gastwissenschaftsarbeitsplatz oder über eine kontrollierte Datenfernverarbeitung auszuwerten. Produktspezifische Informationen zur On-Site-Nutzung werden im Metadatenreport Teil II zur On-Site-Nutzung bereitgestellt.

Weitere allgemeine Informationen zum Mikrozensus sind auf den Seiten des Statistischen Bundesamtes (Was ist der Mikrozensus?) und der Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder abrufbar (Häufige Fragen zum Mikrozensus).

[1] Die Aufbereitung und Dokumentation des faktisch anonymisierten Einzelmaterials erfolgt durch das Forschungsdatenzentrum der Statistischen Ämter der Länder – Standort NRW und das Statistische Bundesamt, Referat F 36 – Mikrozensus - Auswertung und Analyse in Kooperation mit dem German Microdata Lab der GESIS.

# 1. Datenaufbereitung in den FDZ

## 1.1 Datenaufbereitung

Das Datenmaterial wird einer Vollzähligkeits- und Vollständigkeitskontrolle unterzogen, bei der geprüft wird, ob alle Erhebungsmerkmale und die dazugehörigen typisierten Merkmale, die für das Mikrozensusjahr vorgesehen sind, belegt sind. Nicht belegte Merkmale werden aus dem Datenmaterial entfernt. Gleichzeitig wird darauf geachtet, dass die einzelnen Variablen die korrekten Formate aufweisen.

Außerdem sind in den gesamten Mikrozensus nur vollständige Haushalte miteingeflossen. Das bedeutet, dass bei einer fehlenden Angabe einer Person der ganze Haushalt aus dem entsprechenden Befragungsteil entfernt wurde. Missingkodierung

### 1.1.1 Missingkodierung

Im MZ-SUF wird eine Ausfallkennung zur Kenntlichmachung der Ausfallursache umgesetzt. Weiterhin wird geprüft, ob sich die Filterführung des Fragebogens im Datenmaterial widerspiegelt. Für Personengruppen, denen gemäß Ausfalltypisierung bestimmte Fragen nicht gestellt wurden, werden für die betroffenen Merkmale die in Tabelle 1 beschriebenen Werte codiert. Liegt ein Fehlwert in den Daten vor, wird der dazu passende Missingcode hinterlegt. Treten mehrere Ausfallgründe beim selben Befragten auf, wird der Missingcode eingetragen, der in Tabelle 1 als erstes referenziert ist. Wird beispielsweise eine Person unter 15 Jahren aus einer Gemeinschaftsunterkunft (GU) befragt, wird die Ausfallkennung -1 zugewiesen. Bei einer unter 15-jährigen Person, die nicht erwerbstätig ist, wird der Missingcode -3 hinterlegt. Die Abfolge der Ausfalltypisierung wird vorgenommen, um die Zuordnung der Ausfalltypisierung zu vereinfachen, da andernfalls verschiedene Kombinationen eines Ausfalls geprüft werden müssten.

| **Missingcode** | **Beschreibung** |
|---|---|
| -1 | Gemeinschaftsunterkunft (GU) |
| -4 | Nicht in der Substichprobe des Ad-hoc Moduls bzw. <Benennung der jeweiligen Unterstichprobe bzw. des Zusatzprogramms> |
| -3 | Person unter 15 (bzw. 16) Jahre[2] |

[2] Der Code -3 wird auch dort verwendet, wo die Altersgrenze nicht 15 Jahre, sondern 16 Jahre ist (z.B. im Erhebungsteil SILC).

| | |
|---|---|
| -2 | Nichterwerbstätige |
| -6 | Nebenwohnsitze für SILC: Personen am Nebenwohnsitz entfallen in der SILC-Unterstichprobe, bei Haushaltsvariablen entfallen im SILC-Teil Haushalte, in denen keine Person ab 16 Jahren den Hauptwohnsitz hat |
| -5 | Spezifischer Filterausfall, Ausfallgrund wird variablenspezifisch benannt |
| -9 | Filterfehler oder technisch zugelassene Fehlwerte bei Fragen mit Auskunftspflicht, Personen ohne substichprobenspezifischen Hochrechnungsfaktor bei SILC- oder LFS-Unterstichprobe |

*Tabelle 1: Ausfallkennungen*

Im MZ-SUF vor 2020 wurden zusätzlich der Missingcode -8 für Leerstand und ausgefallene Privathaushalte sowie der Missingcode -7 für Auswahlbezirke (AWBs) ohne potenziell zu befragende Haushalte vergeben. Im MZ-SUF 2020 werden diese beiden Ausfälle nicht programmiert, da im Datenmaterial 2020 nur freigegebene Haushalte (kein Leerstand oder unbewohnte Wohnungen und nur AWBs mit zu befragenden Haushalten) vorhanden sind.

Item Nonresponse bei freiwilligen Fragen wurden mit 9, 99, 999 oder Ähnlichem codiert.

### 1.1.2 Hochrechnungs-/Gewichtungsvariablen im MZ-SUF

Fälle ohne einen Kernhochrechnungsfaktor werden aus dem Datenmaterial entfernt. Die Hochrechnungsfaktoren werden so angeglichen, dass sie ohne zusätzliche Multiplikation (mit 1000) auf die Gesamtbevölkerung hochgerechnet werden können. Zudem werden die Hochrechnungsfaktoren im MZ-SUF skaliert, indem alle Hochrechnungsfaktoren im gesamten MZ durch 0,7 geteilt werden. Durch diese Skalierung kommt es zu minimalen Abweichungen zwischen Ergebnissen des MZ-SUF und den veröffentlichten Ergebnissen.

## 1.2 Anonymisierungsmaßnahmen

Die angewendeten Anonymisierungsverfahren im MZ-SUF bestehen aus Verfahren zur Informationsreduktion für einzelne Merkmalstragende und für einzelne Merkmale.

### 1.2.1 Substichprobenziehung

Der MZ-SUF ist eine faktisch anonymisierte 70% Substichprobe der AWBs des Mikrozensus. Seit 2012 werden als Auswahleinheiten für die Substichprobe die AWBs innerhalb eines Rotationsviertels herangezogen.

Um zu gewährleisten, dass die Substichprobe in bestimmten Merkmalen nur geringe zufallsbedingte Abweichungen aufweist, wird das gesamte Mikrozensusmaterial zunächst nach folgenden Merkmalen sortiert, um eine geschichtete Substichprobenziehung zu erreichen:

- Kennung Rotationsviertel (AWBRotationsviertel)
- Bundesland (Land)
- Summe der Befragungen im AWB (Dezile)
- Kennung über Grundauswahl und Aktualisierung der AWBs (AWBANSGROESSENKLASSE)
- Kennung über die Unterstichprobe (ABBZUSATZPROGRAMM)
- Regierungsbezirk (GBTRegierungsbezirk)
- Regionale Anpassungsschicht (AWBLFDNRANPASSUNGSSCHICHTSTICH)
- Regionale Schicht (AWBLFDNRREGSCHICHTSTICHPROBE)
- Regionale Untergruppe (AWBLFDNRREGREGUNTERGRUPPESTICH)
- Gemeindegrößenklasse (GBTGEMEINDEGROESSENKLASSEAKTUE)
- Nr. des AWB (AWBNummerFremd)

Zwar wurde bei der Grundauswahl bzw. wird bei der Aktualisierung der AWBs eine einheitliche Größe der jeweiligen Anschriftengrößenklassen angestrebt, sie unterscheiden sich jedoch fluktuationsbedingt erheblich. Um eine daraus resultierende Erhöhung der Fehlervarianz zu begrenzen, erfolgt die zusätzliche Aufnahme einer Sortierung nach der Anzahl der Befragungen in einem AWB. Die Größenklassen der AWB (Dezile) werden berechnet, indem man die Anzahl der Befragungen pro AWB ermittelt und diese dann auf AWB-Ebene aggregiert.

Jeweils zehn in der Reihenfolge der Sortierung aufeinanderfolgende AWBs bilden eine Schicht. Bei einem Wechsel in den Merkmalen AWBRotationsviertel, Land oder Dezile gibt es einen Sortierwechsel (d.h. die Durchnummerierung startet wieder bei 1), sodass auch unvollständige Schichten zugelassen werden. Im Hinblick auf die Analysemöglichkeit nach Bundesländern sowie die anzustrebende Homogenität der AWBs innerhalb einer Schicht, die varianzreduzierend wirkt, werden bei Sortierwechseln von Rotationsviertel, Bundesland und Größenklassen (Dezile) der AWBs unvollständige Schichten mit weniger als zehn AWBs zugelassen. Die letzte Schicht kann ebenfalls weniger als zehn AWBs umfassen.

Zur Ziehung der Substichprobe werden für jede Schicht drei reproduzierbare, ganzzahlige und gleichverteilte Zufallszahlen in einer Spanne von 1-10 gezogen. AWBs, deren Schichtnummer nun einer der gezogenen Zufallszahlen entsprechen, werden gekennzeichnet und im MZ-SUF nicht mehr berücksichtigt. Wurde also für eine Schicht die Zufallszahl 6 gezogen, wird der AWB mit der Schichtnummer 6 markiert und nicht mehr mit in den MZ-SUF aufgenommen (usw.). Der Stichprobenumfang jeder vollständigen Schicht beträgt somit 7. Die Anzahl gezogener Auswahlbezirke in unvollständigen Schichten ist zufällig, die Ziehungswahrscheinlichkeit von 7/10 bleibt aber erhalten. Mit der so gezogenen einfachen Stichprobe ohne Zurücklegen werden gleiche Auswahlwahrscheinlichkeiten für alle AWBs des MZ-SUF erreicht.

### 1.2.2 Identifikatoren

Direkte Identifikatoren und Hilfsmerkmale werden aus dem Datenmaterial entfernt. AWB-Nummer, Haushaltsnummer, Gebäudenummer und Wohnungsnummer werden zudem verfremdet. Die Ordnungsnummern des AWBs, des Haushalts im AWB und der Person im Haushalt sind systemfrei sortiert, sodass anhand der Position des einzelnen Falls im Datenmaterial kein Rückschluss auf einzelne Personen möglich ist.

### 1.2.3 Gesperrte Merkmale

Es gibt eine Reihe von Merkmalen, die Forschenden am Gastwissenschaftsarbeitsplatz (GWAP) einsehen und auswerten können, die aber nicht mit in den MZ-SUF aufgenommen werden. Solche Merkmale werden als „gesperrte Merkmale“ bezeichnet.

Hierzu gehören einige Stichprobenkennzeichen, um persönliche oder regionale Rückschlüsse zu vermeiden. Zudem werden Gebietskennzeichen für nicht-administrative Raumeinheiten nicht für den MZ-SUF freigegeben.

| **Merkmalsbezeichnung** | **Variable** |
|---|---|
| Berichtsmonat | TPBerichtsmonat, ABBReferenzmonat |
| Berichtswoche | TPBerichtswoche, ABBReferenzwocheBefragung |
| Befragungswoche | PBFBefragungswoche |
| Regionalkennungen | GBTKreis, GBTRegierungsbezirk, GBTOrtsteil, GBTGemeinde, AWBEZKENNZEICHENMEHREREGEMEIN, AWBLFDNRANPASSUNGSSCHICHTAKTUE, AWBLFDNRANPASSUNGSSCHICHTSTICH, AWBLFDNRREGSCHICHTAKTUELL, AWBLFDNRREGSCHICHTSTICHPROBE, AWBLFDNRREGREGUNTERGRUPPEAKTUE, AWBLFDNRREGREGUNTERGRUPPESTICH, |

| | |
|---|---|
| | GBTGEMEINDEGROESSENKLASSEZENSU, AWBAnsGroessenklasse, AWBANZZERLEGUNGSTEILGEB |
| Gemeinde der Arbeitsstätte | EC0303P |
| Stichprobenkennzeichen | AWBStichprobenNR, AWBRotationsviertel, AWBKennzDispropAuswahlsaetze, AWBWOCHENKENNZEICHEN13, AWBWOCHENKENNZEICHEN14, AWBAKTUALISIERUNGSJAHR, AWBTeilschichtkennzeichen, AWBUnterschichtkennzeichen, AWBNRGebaeudeteil, AWBRVVQUARTALSKENNZEICHEN, AWBAUSWAHLTEIL, AWBANZWOHNKUMANSCHRIFTEN, AWBWOHNBERECHBEVOELKERUNG, AWBZONENNUMMER, ABBNULLBEZIRK |
| Raumeinheit (nicht-administrativ) | Stadt_Land_Glied_EU, BIK_Stadtregion_neu, Regionstyp_grund, Regionstyp_diff, Kreistyp, Arbeitsmarktregion, Raumordnungsregion, Planungsregion, Arbeitsagenturbezirk, Bundestagswahlkreis_von, Bundestagswahlkreis_bis, Gemeindetyp, Verdichtungsraum, EUFoerdergebiet, LUZ, Verflechtungsbereich_Zentr |
| Nettoeinkommen (im letzten Monat): Betrag | DG0101P |
| Haushaltsnettoeinkommen (im letzten Monat): Betrag, nur CAPI/CAWI | DG0300P |
| Merkmale aus Erhebungsteil SILC | SILC-Merkmale |
| Typisierte Merkmale | Typisierte Merkmale, die selbstständig erstellt werden können (siehe: Börlin 2020) |

*Tabelle 2: Gesperrte Merkmale*

Die einzelnen Erhebungsmerkmale des Erhebungsteils zu den Einkommen und Lebensbedingungen (Substichprobe SILC) werden nicht in den MZ-SUF aufgenommen. Stattdessen werden ab dem Berichtsjahr 2020 erstmalig SILC-Merkmale der Beobachtungseinheit zu Indizes zusammengefasst und in den MZ-SUF integriert. Neben den Indizes beinhaltet der MZ-SUF einige wenige Merkmale, die zur SILC-Unterstichprobe

gehören, die aber bis einschließlich Berichtsjahr 2019 Teil des Kern-MZ waren. Als Index im MZ-SUF werden Variablen bezeichnet, die sich durch die Berechnung aus den Werten mehrerer Merkmale für ein neu generiertes Konstrukt (z.B. „Wohnumgebung“) ergeben. Das Ziel der Indexbildung besteht darin, die verschiedenen Indikatoren eines Konstruktes zu einer Messgröße zusammenzufassen. Die Ausprägungen der im MZ-SUF integrierten Indizes ergeben sich aus der Summe der Werte der Merkmale. Indizes, die keinen gültigen Wert in den Merkmalen aufweisen, werden als Missing kodiert. Dies ist immer dann der Fall, wenn mindestens eine der zugrunde liegenden Variablen keinen gültigen Wert aufweist.

Folgende SILC-Indizes stehen zur Verfügung:

- Index_Wohnsituation
- Index_Wohnumgebung
- Index_Finanz_Situation
- Index_Sozialkapital
- Index_Lebenssituation

Ab dem Berichtsjahr 2022 sollen darüber hinaus hinzukommen:

- Index_Kinderbetreuung_Std
- Index_Gesundheitszustand

Die zugrundeliegenden Erhebungsmerkmale für diese beiden Indizes weisen in den Berichtsjahren 2020 und 2021 hohe Ausfallquoten auf.

### 1.2.4 Vergröberungen von Merkmalsausprägungen

Das MZ-SUF unterscheidet sich u.a. vom Originalfile des MZ dadurch, dass bestimmte Merkmale im MZ-SUF, bedingt durch die Anonymisierung, in vergröberter Form verfügbar sind. Vergröberungen von Merkmalsausprägungen werden wie folgt angewendet:

1. Keine einzelne Gemeinde mit weniger als 500 000 Einwohnern darf identifizierbar sein.
2. Werden mehrere Gemeinden zu Größenklassen zusammengefasst, muss die Gemeindegrößenklasse in jedem Bundesland mindestens 400 000 Einwohner umfassen.
3. Angaben über Staatsangehörigkeit oder Geburtsland werden so weit aggregiert, dass eine Nationalität oder Gruppe von Nationalitäten mindestens 50 000 Einwohner in der Grundgesamtheit umfassen. Zusammenfassungen erfolgen unter Beachtung der räumlichen Lage bzw. wirtschaftlichen Verflechtungen. Zusammenfassungen innerhalb der EU richten sich nach den Beitrittsjahren der EU und nicht zwingend nach geografischer Nähe. Die Zuordnung erfolgt nach den Kriterien: EU, EFTA, EU-Beitrittskandidaten, sonstige Staaten.

4. Alle weiteren Variablen werden so weit aggregiert, dass eine ausgewiesene Merkmalsausprägung mindestens 5 000 Einwohner in der Grundgesamtheit umfasst. Die Merkmalsvergröberung bei metrisch skalierten Angaben wird über eine Zusammenfassung direkt benachbarter Merkmalswerte vorgenommen. Im Fall von Vergröberungen wird im MZ-SUF i. d. R. die am stärksten besetzte Kategorie ausgewiesen.

## 1.3 Methodik der Verknüpfung

Der Mikrozensus stellt kein bereits verknüpftes Produkt dar, allerdings besteht die Möglichkeit, auf Individualebene die Wellen ab 2020 miteinander zu einem Paneldatensatz zu verknüpfen.[3] Hierfür stehen ab 2020 verkettete Identifikatoren für Auswahlbezirke (idawb), Haushalte (idhh) und Personen (idpers) zur Verfügung. Unterjährige Wiederholungsbefragungen im Rahmen der Arbeitskräfteerhebung (LFS) lassen sich über die vorgesehene Bogenart (AWBAUSWAHLTEIL=4) identifizieren. Sofern aus methodischen Gründen keine Dubletten innerhalb eines Erhebungsjahres gewünscht sind, ist eine Entfernung der Fälle erforderlich. In diesem Fall ist zu beachten, dass die Hochrechnungsfaktoren HR100QQ und HR100JQ sowie die Hochrechnungsfaktoren für die Kernvariablen nicht für eine Hochrechnung auf die Gesamtbevölkerung geeignet sind, da relevante Fälle entfernt werden. Auswirkungen auf die Hochrechnung mit HR100JJ gibt es nicht, da der Hochrechnungsfaktor für LFS-Strukturvariablen bei Widerholungsbefragungen unbelegt ist. In den ebenfalls enthaltenen querschnittsorientierten Identifikatoren (idawbx, idhhx, idpersx) sind die Wiederholungsbefragungen bereits mit einer abweichenden ID versehen. Nähere Informationen zur Panelverknüpfung sind Brocker und Mühlenfeld 2020 sowie Herter-Eschweiler und Schimpl-Neimanns 2018 zu entnehmen. Die Arbeitspapiere beziehen sich auf die Erhebungsjahre 2012 bis 2015. Die Erkenntnisse sind aber auch auf die Jahre 2016 bis 2019 und in Teilen auch auf die Jahre ab 2020 zu übertragen.

[3] Wie im Metadatenreport Teil I näher beschrieben, stellt die Stichprobe des Mikrozensus ein rotierendes Panel dar. Seit dem Erhebungsjahr 2012 werden als Auswahleinheiten für die Substichprobe die Auswahlbezirke innerhalb eines Rotationsviertels herangezogen. Damit wird seit 2012 die Möglichkeit geschaffen, selbstständig mit den MZ-SUFs Paneldatensätze zu erzeugen. Infolge der Erneuerung der gesamten Mikrozensusstichprobe im Jahr 2016 sind Verknüpfungen der Querschnittsdaten ab dem Erhebungszeitpunkt 2012 nur bis einschließlich der Erhebung 2015 möglich. Zudem sind aufgrund der Weiterentwicklung des Mikrozensus ab 2020 Verknüpfungen der Querschnittsdaten ab dem Erhebungszeitpunkt 2016 nur bis einschließlich der Erhebung 2019 möglich.

# 2. Produkt

## 2.1 Hinweise zur Qualität des Mikrozensus 2020

Das Erhebungsjahr 2020 des Mikrozensus stand im Zeichen von zwei besonderen Herausforderungen. Zum einen wurde für den MZ 2020 ein komplett neues IT-System aufgebaut, dessen Einführung von technischen Problemen begleitet war. Zum anderen erschwerte die Corona-Pandemie die Vorbereitung und Umsetzung der Datenerhebung in den Haushalten. Zentrale Einschränkungen, die sich auf die Datenqualität des Mikrozensus auswirken, werden im Folgenden aufgeführt.

Im Jahr 2020 war es nicht möglich, Erhebungsbeauftragte im gewohnten Maße für die Befragungen einzusetzen. Ausgangs- und Kontaktbeschränkungen infolge der Corona-Pandemie verhinderten überwiegend die gewohnte persönliche Vor-Ort-Befragung der Haushalte (CAPI) und erschwerten zusätzlich die Durchführung der (Vor-)Begehungen von Gebäuden, die zur Stichprobenkonkretisierung notwendig sind. Zudem wurde das Mahnwesen in vielen Ländern nicht oder nur teilweise umgesetzt.

Hieraus resultierte eine vom MZ sonst unbekannt hohe Ausfallquote. Im Bundesschnitt gibt es einen Ausfall von rund 35% bei den Endergebnissen. Die Ausfallquoten für die freiwilligen Fragen, z.B. zu den Lebensbedingungen aus der neu integrierten SILC-Unterstichprobe, sind in der Regel noch deutlich höher. Die Antwortausfälle können nicht als zufällig angenommen werden. Zudem sind sie regional, u.a. zwischen den Bundesländern, und zeitlich sehr unterschiedlich verteilt. Informationen zu länderspezifischen Besonderheiten für das Jahr 2020 finden sich auf den Web-Seiten der Statistischen Ämter der Länder zum Mikrozensus.

Die Erhebungssituation 2020 hat zur Folge, dass Auswertungen – insbesondere in Kombination von fachlicher und regionaler Tiefe – nicht durchgängig die vom Mikrozensus sonst gewohnte Qualität aufweisen. Die Ergebnisse ab 2020 sind nur eingeschränkt mit den Ergebnissen der Vorjahre vergleichbar. Darüber hinaus gibt es 2020 eine Reihe von Auffälligkeiten in verschiedenen Themenbereichen zu beachten, z.B. zur Abbildung gleichgeschlechtlicher Paare, des Migrationshintergrundes oder von atypisch Beschäftigten. Hinweise hierzu finden sich auf den Seiten des Statistischen Bundesamts zur Neuregelung des Mikrozensus ab 2020 und in den offiziellen Qualitätsberichten zum Mikrozensus und der Unterstichprobe SILC.

Die amtliche Statistik hat sich vor dem Hintergrund der genannten Schwierigkeiten darauf verständigt, nur stark eingeschränkt Ergebnisse unterhalb der Landesebene zu veröffentlichen. Vor dem Hintergrund der genannten Besonderheiten des Erhebungsjahres sollten keine Veränderungsanalysen zu Vorjahren durchgeführt werden und auch keine kausalen Interpretationen bezüglich möglicher Effekte der Corona-Pandemie abgeleitet

werden. Für wissenschaftliche Projekte werden die Daten in der bisherigen regionalen und inhaltlichen Tiefe bereitgestellt. Dies dient insbesondere dazu, eine Verknüpfung mit externen Merkmalen zu ermöglichen.

Es ist zu beachten:

- Vor dem Hintergrund der skizzierten Einschränkungen werden Veränderungsanalysen zu Vorjahren sowie Analysen unterhalb der Landesebene (NUTS-1) für das Jahr 2020 nicht empfohlen.
- Auch ein Vergleich zwischen den Bundesländern ist aufgrund der teils großen Unterschiede in der Erhebungspraxis für dieses Berichtsjahr nicht sinnvoll.
- Die neue Struktur des Mikrozensus (verschiedene Unterstichproben und damit einhergehende veränderte Hochrechnungsverfahren, unterjährige Rotation einzelner Teile, verändertes Fragenprogramm) und der erstmalige Einsatz von Online-Fragebögen (CAWI) zur Datenerhebung sind bei der Interpretation von Ergebnissen ebenfalls zu berücksichtigen.

## 2.2 Merkmale und Merkmalsbeschreibung

### 2.2.1 Merkmalsdefinitionen

Folgende Klassifikationen finden in dem beschriebenen Produkt Anwendung:

- Klassifikation der Berufe, Ausgabe 2010, überarbeitete Fassung von 2020 (KldB 2010 ü. F.) (3-Steller + Anforderungsniveau):
  https://statistik.arbeitsagentur.de/DE/Navigation/Grundlagen/Klassifikationen/Klassifikation-der-Berufe/KldB2010-Fassung2020/Systematik-Verzeichnisse/Systematik-Verzeichnisse-Nav.html

- Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008) (3-Steller):
  https://www.destatis.de/DE/Methoden/Klassifikationen/Gueter-Wirtschaftsklassifikationen/Downloads/klassifikation-wz-2008-3100100089004.pdf?__blob=publicationFile&v=5

- Bildungsklassifikation International Standard Classification of Education, Ausgabe 2011 (ISCED-2011):
  http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf (englisch)

- Bildungsfelder ISCED Fields of Education and Training (ISCED-F 2013):
  https://www.statistik.at/KDBWeb/kdb_DownloadsAnzeigen.do?KDBtoken=ignore&&AUFRUF=klass&&NAV=DE&&KLASSID=10527&&KLASSNAME=ISCED-F
  (Übersetzung der Klassifikation von Statistik Austria)
  Auch 2020 stehen zu den Hauptfachrichtungsvariablen EP1103PUG4, ER0703PUG4 und ER0713PUG4 auf die bis 2017 verwendeten Zweisteller umgeschlüsselte Variablen (EP1103PA, ER0703PA, ER0713PA) zur Verfügung.

- Staatsangehörigkeits- und Gebietssystematik (außereuropäische Staaten zum Teil zusammengefasst siehe Schlüsselverzeichnisse): https://www.destatis.de/DE/Methoden/Klassifikationen/Staat-Gebietsystematik/staatsangehoerigkeit-gebietsschluessel.html

### 2.2.2 Datensatzbeschreibung

Der Mikrozensus ist eine Haushalts- und Personenstatistik. Die vollständige Liste der im MZ-SUF vorhandenen Variablen, ist dem Datenhandbuch zu entnehmen. Sofern Variablen das Ergebnis einer einzelnen Frage des Fragebogens wiedergeben, sind die zugehörigen Fragenummern und der Fragetext enthalten. Zudem sind im Datenhandbuch Randverteilungen (ungewichtete Fallzahlen) der Variablen, erläuternde Kommentare und die vollständigen Label (das Statistikprogramm SPSS übernimmt nur 120 Zeichen) enthalten.

Aufgrund der 2020 eingeführten Neukonzeption des Mikrozensus, welche die Integration weiterer Haushaltserhebungen umfasst (die Arbeitskräfteerhebung (LFS), die Statistik der Einkommens- und die Lebensbedingungen (SILC) und ab 2021 die Befragung zu Informations- und Kommunikationstechnologien (IKT)), besteht die Mikrozensus-Befragung ab 2020 aus einem verkürzten Kernfragenprogramm, das alle befragten privaten Haushalte beantworten, und weiteren Erhebungsteilen, welche jeweils lediglich eine Unterstichprobe der Haushalte erhält. Das Kernprogramm und die verschiedenen Erhebungsteile werden nicht modular hintereinander erhoben, sondern das resultierende Frageprogramm verzahnt die Inhalte thematisch. Um den Mikrozensusnutzenden einen kompakten Gesamtüberblick über alle enthaltenen Fragen und die Struktur der verschiedenen Fragebögen zu geben, wurde ein Masterfragebogendokument erstellt, welches die Informationen aus allen fünf Fragebögen enthält und angibt, welche der Fragen in welchen Fragebögen enthalten sind. Zudem enthält das Dokument Informationen zu den korrespondierenden Variablennamen, welche im MZ-SUF enthalten sind.

## 2.3 Vergleichbarkeit der Merkmale über die Zeit

Vor dem Hintergrund der in Abschnitt 2.1 skizzierten methodischen Neuerungen und den Einschränkungen in der Qualität des Mikrozensus 2020 wird von Veränderungsanalysen zu Vorjahren generell abgeraten. Alle Neuregelungen zum Mikrozensus 2020 können hier nachgelesen werden.

### 2.3.1 Variablennamen

Eine weitere Neuerung im Erhebungsjahr 2020 ist die veränderte Bezeichnung der Variablen. Variablennamen beginnen nicht mehr mit EF (Eingabefeld) und sind nicht mehr durchnummeriert. Namen von Erhebungsmerkmalen bestehen aus zwei Buchstaben und vier

Ziffern. Danach folgt die Kennung der Erhebungsebene (P: Person, H: Haushalt, L: Lebensform). Gegebenenfalls folgen weitere Untergliederungen mit dem Buchstaben U. Typisierte Merkmale beginnen mit T. Außerdem gibt es einen anderen Umgang mit Haupt- und Nebenwohnsitzhaushalten als bisher, insbesondere im SILC-Befragungsteil: Die Befragungen für die neu integrierten Unterstichproben SILC und ab 2021 IKT erfolgen nur in Haushalten, in denen mindestens eine Person über 16 Jahren ihren Hauptwohnsitz hat. Weitere Informationen hierzu finden sich in Hochgürtel und Weinmann 2020; 93.

Im Datenhandbuch und im Masterfragebogendokument sind bei Variablen, bei denen es eine Kontinuität zu Vorjahren gibt, die früheren Variablennamen aufgeführt. Im Detail können sich Frageformulierungen oder Antwortkategorien unterscheiden. Die Variablen-Zeitpunkte-Matrix in MISSY gibt Hinweise zur zeitlichen Vergleichbarkeit von Variablen des Mikrozensus ab 1973.

### 2.3.2 Missingkodierung

Die Struktur der Missings hat sich im Vergleich zu 2019 ebenfalls verändert:

- Die Missingcodes -7 und -8 finden im neuen Mikrozensus 2020 keine Anwendung mehr, da nur noch realisierte Interviews in den Datensatz miteinbezogen wurden.
- Der Code -3 wird auch dort verwendet, wo die Altersgrenze nicht 15 Jahre, sondern 16 Jahre ist (z.B. im Erhebungsteil SILC).
- Da es keine Jahresüberhänge mehr gibt, wird der Code -6 ab 2020 für Filter aufgrund von Nebenwohnsitzen verwendet. Fragen des SILC-Erhebungsteils (im Fragebogen 5) werden überwiegend nur Personen am Hauptwohnsitz bzw. Haushalten, in denen mindestens eine Person ab 16 Jahren den Hauptwohnsitz hat, gestellt.
- Missingcodes werden hierarchisch vergeben. Weitere Informationen zur Missingskodierung finden sich im Abschnitt 1.1.1.

### 2.3.3 Haushalte und Lebensformen

Vor dem Hintergrund der genannten Qualitätseinschränkungen (Unterkapitel 2.1) gibt es bei der Analyse von (kleinen) Teilpopulationen Schwierigkeiten in der Vergleichbarkeit zu den Vorjahren. So hat sich zum Beispiel die Zahl der gleichgeschlechtlichen Paare im Mikrozensus 2020 im Vergleich mit 2019 verdoppelt, die von verheirateten gleichgeschlechtlichen Paaren gar verdreifacht[4].

[4] Hierfür sind ggf. auch Veränderungen in der Frageformulierung und den Antwortkategorien verantwortlich. Eine inhaltliche Veränderung dieses Ausmaßes erscheint unplausibel.

### 2.3.4 Migrationstypisierungen

Die Zahlen zur Bevölkerung nach Migrationshintergrund aus dem Mikrozensus 2020 sind nur eingeschränkt mit den Vorjahren vergleichbar. Dies gilt insbesondere für einige Teilpopulationen (z. B. als Deutsche Geborene, Eingebürgerte), die umso stärker schwanken, je kleiner diese sind (z. B. Differenzierung nach Geburtsland). Neben den in Unterkapitel 2.1 erläuterten generellen Umstellungen und methodisch-technischen Einschränkungen sind Umstellungen in der Erhebung/Frageformulierung sowie in der Methodik der Typisierung des Migrationshintergrundes[5] hierfür ursächlich. Durch die methodische Weiterentwicklung der Typisierung des Migrationshintergrundes werden die mit deutscher Staatsangehörigkeit Geborenen besser abgebildet.

### 2.3.5 Variablen zur Arbeitssituation

Bei Analysen zum Thema Arbeitsmarkt besteht eine zusätzliche Unsicherheit bei der Bewertung der Ergebnisse, da sich pandemiebedingt die Situation auf dem deutschen Arbeitsmarkt in vielen Bereichen deutlich verändert hat. So können bei den Ergebnissen nur bedingt Aussagen getroffen werden, ob diese auf reale Entwicklungen oder auf die oben beschriebenen Einschränkungen zurückzuführen sind. Mit zunehmender Gliederungstiefe nehmen diese Unsicherheiten zu (z. B. bei Erwerbslosenquoten in tiefer regionaler oder demographischer Gliederung). Ein weiteres Problem gibt es bei der Erfassung von Zeitarbeit insbesondere bei CAWI-Erhebungen. Aus diesem Grund ist die Vergleichbarkeit im Hinblick auf den Mikrozensus 2021 eingeschränkt.

### 2.3.6 Einkommen und Lebensbedingungen (SILC)

Die früher eigenständige Statistik zu Einkommen und Lebensbedingungen („Leben in Europa“, EU-SILC) ist ab 2020 ein Erhebungsteil des Mikrozensus (MZ-SILC). Bis zu 12% der ausgewählten Haushalte werden neben dem Kernprogramm zu Einkommen, Wohnen, Gesundheit und Lebensbedingungen befragt. Durch den Wechsel der Erhebung von einer freiwilligen zu einer in Teilen auskunftspflichtigen Befragung ist ein inhaltlicher Vergleich der Daten des Erhebungsjahres 2020 mit den Vorjahren nicht möglich. Grundsätzlich besteht im Rahmen des Mikrozensus Auskunftpflicht. Allerdings wird aufgrund der fortgesetzten Freiwilligkeit von Fragen zu Lebensbedingungen ein hoher Anteil an fehlenden Werten (Missings) erzeugt.

### 2.3.7 Zusatzprogramm 2020

[5] Personen, die angegeben haben, die deutsche Staatsangehörigkeit als (Spät-)Aussiedler/in erlangt zu haben, aber gleichzeitig vor 1950 nach Deutschland zugewandert sind, werden als Vertriebene und somit als Person ohne Migrationshintergrund umgesetzt. Gleiches gilt auch für die Personen, die zu ihren externen Eltern angegeben haben, dass diese als (Spät-)Aussiedler/in vor 1950 nach Deutschland zugewandert sind. Für alle anderen Personen wird diese Abgrenzung auf Basis des Zuzugsjahres nicht vorgenommen.

Zusätzlich zum Kernprogramm enthält der Mikrozensus vierjährliche Zusatzprogramme. Im Erhebungsjahr 2020 wurde das Zusatzprogramm zu den Pendlereigenschaften von Schüler*innen, Studierenden sowie Erwerbstätigen erhoben.

Pendlerverhalten Erwerbstätige

- EC2100P Gehen bzw. fahren Sie üblicherweise von dieser Wohnung zu Ihrer Arbeitsstätte?
- EC2200P Wie weit ist der Hinweg zu Ihrer Arbeitsstätte, z.B. zum Betriebsgelände, Dienstgebäude?
- EC2300P Wie lange brauchen Sie normalerweise für den Hinweg zu Ihrer Arbeitsstätte?
- EC2400P Welches Verkehrsmittel benutzen Sie normalerweise auf dem Hinweg zu Ihrer Arbeitsstätte?
- EC2700P Nutzen Sie ein weiteres Verkehrsmittel, mit dem Sie eine wesentliche Strecke für den Hinweg zu Ihrer Arbeitsstätte zurücklegen?
- EC2701P Welches weitere Verkehrsmittel nutzen Sie?

Pendlerverhalten Schüler*innen und Studierende

- DC1100P Liegt die (zuletzt) besuchte Schule/Hochschule in der Gemeinde, in der Sie wohnen?
- DC1201P Liegt Ihre Schule/Hochschule in Deutschland?
- DC1300P Gehen oder fahren Sie üblicherweise von dieser Wohnung zu Ihrer Schule/Hochschule?
- DC1400P Wie weit ist der Hinweg zu Ihrer Schule/Hochschule?
- DC1500P Wie lange brauchen Sie normalerweise für den Hinweg zu Ihrer Schule/Hochschule?
- DC1600P Welches Verkehrsmittel benutzen Sie normalerweise auf dem Hinweg zu Ihrer Schule/Hochschule?
- DC1700P Nutzen Sie ein weiteres Verkehrsmittel, mit dem Sie eine wesentliche Strecke auf dem Hinweg zu Ihrer Schule/Hochschule zurücklegen?
- DC1800P Welches weitere Verkehrsmittel nutzen Sie?

### 2.3.8 Adhoc-Modul 2020

Im Rahmen der in den Mikrozensus integrierten Arbeitskräfteerhebung (LFS) werden normalerweise jährlich Ad-hoc-Module durchgeführt, die dazu dienen, detaillierte Informationen zu politisch relevanten Themen der EU zu sammeln, welche nicht Teil des LFS-Standardfrageprogramms sind. Im Jahr 2020 wurde das Ad-hoc-Modul „Arbeitsunfälle und

andere arbeitsbedingte Gesundheitsprobleme“ durchgeführt. Die Fragen des Ad-hoc-Moduls erhält nur ein Teil der Haushalte der LFS-Unterstichprobe. Ihre Beantwortung ist freiwillig.

Gesundheitliche Belastungen bei der Arbeit

- EK0200P Sind Sie gegenwärtig erwerbstätig?
- EE0100P Sind Sie bei Ihrer Arbeit körperlichen Belastungen ausgesetzt, die Ihre Gesundheit schädigen könnten?
- EE0200P Sind Sie bei Ihrer Arbeit seelischen Belastungen ausgesetzt, die Ihr Wohlbefinden beeinträchtigen?

Fragen zu Arbeitsunfällen

- EK0100P Was trifft auf Ihre gegenwärtige Situation zu?
- EK0300P Hatten Sie in den letzten 12 Monaten einen Arbeitsunfall, bei dem Sie sich verletzt haben?
- EK0400P Handelte es sich bei Ihrem letzten Arbeitsunfall um einen Unfall im Straßenverkehr?
- EK0500P Bei welcher Tätigkeit hat sich der letzte Arbeitsunfall ereignet?
- EK0600P Mussten Sie Ihre Erwerbstätigkeit wegen des letzten Arbeitsunfalls zeitweise unterbrechen?
- EK0700P Konnten Sie Ihre Arbeit nach dem letzten Arbeitsunfall mittlerweile wieder aufnehmen?
- EK0800P Wie lange konnten Sie wegen Ihres Arbeitsunfalls keiner Erwerbstätigkeit nachgehen?

Fragen zu arbeitsbedingten Gesundheitsproblemen (keine Arbeitsunfälle)

- EK0900P Hatten Sie in den letzten 12 Monaten Gesundheitsprobleme, die durch Ihre Arbeit verursacht oder verschlimmert wurden?
- EK1000P Welcher der folgenden arbeitsbedingten Beschwerden beeinträchtigt bzw. beeinträchtigte Sie am meisten?
- EK1100P Bei welcher Tätigkeit wurde das Gesundheitsproblem, das Ihre Gesundheit am meisten beeinträchtigt bzw. beeinträchtigte, verursacht oder verschlimmert?
- EK1200P Sind Sie durch das arbeitsbedingte Gesundheitsproblem, das Ihre Gesundheit am meisten beeinträchtigt bzw. beeinträchtigte, bei der Arbeit oder im Privatleben eingeschränkt?
- EK1300P Mussten Sie wegen des arbeitsbedingten Gesundheitsproblems, das Ihre Gesundheit am meisten beeinträchtigt bzw. beeinträchtigte, Ihre Erwerbstätigkeit zeitweise unterbrechen?
- EK1400P Konnten Sie Ihre Arbeit mittlerweile wiederaufnehmen?

- EK1500P Wie lange konnten Sie wegen Ihres arbeitsbedingten Gesundheitsproblems, das Ihre Gesundheit am meisten beeinträchtigt bzw. beeinträchtigte, nicht arbeiten?

## 2.4 Eckwerte relevanter Merkmale und Merkmalskombinationen

### Verteilung der Bevölkerung am Hauptwohnsitz

### Geschlecht

**TPGeschlecht Geschlecht**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | 1 Männlich | 40.635.340 | 49,5 | 49,5 | 49,5 |
| | 2 Weiblich | 41.506.307 | 50,5 | 50,5 | 100,0 |
| | Gesamt | 82.141.647 | 100,0 | 100,0 | |

### Familienstand

**AB0500P Familienstand**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | 1 ledig | 34.406.240 | 41,9 | 41,9 | 41,9 |
| | 2 verheiratet | 37.018.425 | 45,1 | 45,1 | 87,0 |
| | 3 verwitwet | 5.185.998 | 6,3 | 6,3 | 93,3 |
| | 4 geschieden | 5.391.427 | 6,6 | 6,6 | 99,8 |
| | 5 eingetragene Lebenspartnerschaft | 116.374 | ,1 | ,1 | 100,0 |
| | 6 eingetragene Lebenspartnerin/ eingetragener Lebenspartner verstorben | 7.159 | ,0 | ,0 | 100,0 |
| | 7 eingetragene Lebenspartnerschaft aufgehoben | 16.023 | ,0 | ,0 | 100,0 |
| | Gesamt | 82.141.647 | 100,0 | 100,0 | |

## Privathaushalte am Hauptwohnsitz

**Haushaltsmitglieder2 AnzahlHaushaltsmitglieder**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | 1,00 Einpersonenhaushalt | 16.534.185 | 20,1 | 20,1 | 20,1 |
| | 2,00 Zweipersonenhaushalt | 27.556.592 | 33,5 | 33,5 | 53,7 |
| | 3,00 Dreipersonenhaushalt | 14.777.558 | 18,0 | 18,0 | 71,7 |
| | 4,00 Vierpersonenhaushalte | 15.772.144 | 19,2 | 19,2 | 90,9 |
| | 5,00 Fünf- und mehr Personenhaushalte | 7.501.168 | 9,1 | 9,1 | 100,0 |
| | Gesamt | 82.141.647 | 100,0 | 100,0 | |

## Bundesland

**LAND Bundesland**

| | | Häufigkeit | Prozent | Gültige Prozente | Kumulierte Prozente |
|---|---|---|---|---|---|
| Gültig | 1 Schleswig-Holstein | 2.885.431 | 3,5 | 3,5 | 3,5 |
| | 2 Hamburg | 1.838.859 | 2,2 | 2,2 | 5,8 |
| | 3 Niedersachsen | 7.874.850 | 9,6 | 9,6 | 15,3 |
| | 4 Bremen | 672.262 | ,8 | ,8 | 16,2 |
| | 5 Nordrhein-Westfalen | 17.600.448 | 21,4 | 21,4 | 37,6 |
| | 6 Hessen | 6.232.225 | 7,6 | 7,6 | 45,2 |
| | 7 Rheinland-Pfalz | 4.057.044 | 4,9 | 4,9 | 50,1 |
| | 8 Baden-Württemberg | 11.031.098 | 13,4 | 13,4 | 63,5 |
| | 9 Bayern | 13.038.747 | 15,9 | 15,9 | 79,4 |
| | 10 Saarland | 977.969 | 1,2 | 1,2 | 80,6 |
| | 11 Berlin | 3.602.474 | 4,4 | 4,4 | 85,0 |
| | 12 Brandenburg | 2.487.254 | 3,0 | 3,0 | 88,0 |
| | 13 Mecklenburg-Vorpommern | 1.578.764 | 1,9 | 1,9 | 89,9 |
| | 14 Sachsen | 4.025.120 | 4,9 | 4,9 | 94,8 |
| | 15 Sachsen-Anhalt | 2.146.189 | 2,6 | 2,6 | 97,5 |
| | 16 Thüringen | 2.092.913 | 2,5 | 2,5 | 100,0 |
| | Gesamt | 82.141.647 | 100,0 | 100,0 | |

Im Anhang befindet sich die entsprechende Syntax für das Programm SPSS für die oben aufgeführten Tabellen. Die hier abgebildeten Tabellen sind in ähnlicher Form der Fachserie 1 Reihe 2.2 zu entnehmen.

## 2.5 Auswertbare regionale Ebenen

- Bundesebene (NUTS 0)
- Land: Landesebene (NUTS 1)
- Land * Gemeindegrößenklasse

## 2.6 Produktversionen

Die Versionsnummer des Datensatzes ist der Variablen DOI zu entnehmen.

1.0.0 (DOI: 10.21242/12211.2020.00.00.3.1.0)

- Erstveröffentlichung

# 3. Praktische Hinweise

## 3.1 Hinweise zur Geheimhaltung

Bei den MZ-SUF handelt es sich um faktisch anonyme Mikrodaten. Mikrodaten werden als faktisch anonym bezeichnet, wenn eine Deanonymisierung zwar nicht gänzlich ausgeschlossen werden kann, die Angaben jedoch „nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft“ der jeweiligen Merkmalsträgerin beziehungsweise dem jeweiligen Merkmalsträger zugeordnet werden können (§16 Abs. 6 BstatG). Die MZ-SUF bieten im Vergleich zu den On-Site-Zugangswegen ein geringeres Analysepotenzial, sind jedoch so konzipiert, dass sie sich für einen großen Teil der wissenschaftlichen Forschungsvorhaben eignen. Durch die faktische Anonymisierung der Mikrodaten dürfen sie außerhalb der geschützten Infrastruktur der amtlichen Statistik verwendet werden. Voraussetzung hierfür ist, dass die beantragende Institution ihren Sitz in Deutschland hat und dass die bereitgestellten Daten nur in den Räumen der beantragenden wissenschaftlichen Einrichtungen innerhalb Deutschlands genutzt werden. Zudem müssen alle Datennutzenden zur statistischen Geheimhaltung nach § 16 Abs. 7 BStatG verpflichtet werden (https://www.forschungsdatenzentrum.de/de/zugang). Insbesondere sind Handlungen zu unterlassen, die darauf abzielen oder geeignet sind, anonymisierte statistische Einzelangaben zu deanonymisieren.

## 3.2 FAQ

**Wie erhalte ich Zugang zu den MZ-SUF?**

Die als Scientific Use File zur Verfügung stehenden Mikrozensus-Daten können gegen Zahlung einer Bereitstellungsgebühr bei den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder bestellt werden. Diese beträgt 250 Euro pro Statistik und Erhebungsjahr (unter bestimmten Bedingungen werden für Promovierende und Studierende Ermäßigungen gewährt). Auf den Seiten der Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder finden Sie nähere Informationen zum Datenangebot, Datenzugang, Beantragung und Bedingungen.

**Wo erhalte ich Auskunft, wenn ich Fragen zum MZ-SUF habe?**

Ausführliche Informationen und Auswertungshilfen zum MZ-SUF, u. a. Masterfragebogen, Datenhandbuch mit Randauszählungen, Tools zur Umsetzung sozialwissenschaftlicher

Konzepte, Variablen-Zeitpunkte-Matrix, Verknüpfung von MZ-Querschnitterhebungen zu Panels, stehen auf dem Mikrodaten-Informationssystem (MISSY) der GESIS zur Verfügung. Weitere Informationen zum Datenangebot und zum Datenzugang sind zudem auf den Seiten der Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länderr abrufbar.

Bei weiteren Fragen können sich interessierte Personen und Nutzende des Mikrozensus an das Forschungsdatenzentrum der Statistischen Ämter der Länder – Düsseldorf (insbesondere bei Fragen zum Datenzugang und Datenaufbereitung), das Forschungsdatenzentrum des Statistischen Bundesamtes und an das German Microdata Lab (GML) bei GESIS (insbesondere bei inhaltlichen Fragen und Fragen zum Angebot in MISSY) wenden.

**Wie werden in den Mikrozensusdaten Haushalte selektiert?**

Für den Mikrozensus 2020 sind bereits IDs zur Personen- (idpers, idpersx), Haushalts- (idhh, idhhx) und Auswahlbezirkskennung (idawb, idawbx) im Datensatz enthalten. In den IDs der längsschnittorientierten Daten, IDs ohne x am Ende, kann es auf Grund von Wiederholungsbefragungen des LFS-Teils zu Dopplungen der IDs kommen. Dopplungsfrei, d.h. eindeutig sind die IDs mit x am Ende, welche für die querschnittsorientierten Daten genutzt werden können. Die Wiederholungsbefragungen sind bei der Erstellung von Jahresergebnissen einzubeziehen. Dies gilt nicht bezüglich LFS-Strukturvariablen, die nicht mit dem verkürzten Fragebogen 4 erhoben werden. Bei der Verknüpfung über mehrere Erhebungsjahre kann es gewünscht sein, jeden Haushalt nur einmal je Erhebungsjahr zu berücksichtigen. Hierfür wird eine Entfernung der Wiederholungsbefragungen AWBAUSWAHLTEIL==4 empfohlen.

Die Bildung dieser IDs erfolgt für die jeweiligen Einheiten durch die Aneinanderreihung folgender Variablen:

idpers: LAND AWBNummerFremd HHNummerFremd PERNr

idpersx: LAND AWBNummerFremd TPBerichtsquartal HHNummerFremd PERNr

idhh: LAND AWBNummerFremd HHNummerFremd

idhhx: LAND AWBNummerFremd TPBerichtsquartal HHNummerFremd

idawb: LAND AWBNummerFremd

idawbx: LAND AWBNummerFremd TPBerichtsquartal

Leerzeichen werden bei der Bildung der Identifikatoren durch Nullen ersetzt.

**Wie können einer Befragten Person Informationen zur im Haushalt lebenden Mutter und zum im Haushalt lebenden Vater zugeordnet werden?**

Unter Verwendung der Haushaltsidentifikationsnummer (idhh, idhhx) kann über die Merkmale TL0702P und TL0802P die Personennummer der Mutter bzw. des Vaters ermittelt werden.

Der zugehörige Einzeldatensatz der Mutter/des Vaters lässt sich über die Personennummer PERNr finden. Die Informationen können dann satzübergreifend der jeweiligen Referenzperson zugespielt werden.

**Wie ist das Rotationsschema der Haushalte aus der LFS-Substichprobe aufgebaut?**

Um die unterjährigen Veränderungen auf dem Arbeitsmarkt besser analysieren zu können, wurde das Rotationsschema angepasst. Im Gegensatz zum Kern- und SILC-Programm, welche einmal jährlich abgefragt werden, rotiert der LFS-Teil in kürzeren Abständen. Haushalte, die für den LFS-Teil ausgewählt wurden, werden im 2-(2)-2-Schema befragt. Das bedeutet, dass die Haushalte in zwei aufeinanderfolgenden Quartalen befragt werden, anschließend zwei Quartale pausieren und dann wieder zwei Quartale in Folge befragt werden. Eine genauere Beschreibung dazu ist in den Neuregelungen des Mikrozensus[6] zu finden.

**Wann wird welcher Hochrechnungsfaktor verwendet?**

Aufgrund der verschiedenen Substichproben, die seit dem Berichtsjahr 2020 in den Daten vorliegen, stehen auch mehr Hochrechnungsfaktoren als bisher zur Verfügung. Aus der Übersicht in Tabelle 3 lässt sich erkennen, für die Analyse welcher Substichproben welche Hochrechnungsfaktoren vorgesehen sind. Eine detaillierte Beschreibung der Hochrechnungsfaktoren ist im Metadatenreport Teil I Statistik in Kapitel 2.6 zu finden.

- Für die Merkmale aus dem Kernprogramm wird für Hochrechnungen auf das Jahr der Standardhochrechnungsfaktor HR000JJ verwendet. Mit dem Hochrechnungsfaktor HR000QQ können Quartalsergebnisse berechnet werden. Mit TPBerichtsquartal können dafür einzelne Berichtsquartale selektiert werden. Darüber hinaus steht mit HR000JQ der Durchschnitt der vier Quartale zur Verfügung.

[6] https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2019/06/neuregelung-mikrozensus-062019.pdf?__blob=publicationFile

- Strukturmerkmale aus der LFS-Substichprobe (Fragebögen 2 und 3) werden auf das Jahr mit dem Merkmal HR100JJ hochgerechnet. Entsprechend des Kernprogramms stehen auch für die LFS-Merkmale Hochrechnungsfaktoren für die Quartale (HR100QQ) und den Quartalsdurchschnitt (HR100JQ) zur Verfügung. Hierbei ist vor allem zu beachten, dass Haushalte im LFS-Programm mehrfach im Jahr befragt werden (siehe oben). Bei der Auswertung von Merkmalen, die unterjährig wiederholt erhoben werden (Fragebögen 2, 3 UND 4) empfiehlt sich, mit HR100JQ hochzurechnen, sodass nicht nur das Ergebnis der Erstbefragung in die Auswertung eingeht.
- Für die SILC-Substichprobe wird der Hochrechnungsfaktor HR200JJ verwendet.
- Weitere Hochrechnungsfaktoren sind HR100MO für Merkmale des Ad-hoc-Moduls und HR100BH für Menschen mit Behinderung.

*Tabelle 1: Hochrechnungsfaktoren mit Gesamtbevölkerungszahlen der hochgerechneten Jahresergebnisse*

| **Hoch-rechnungs-faktor** | **Stich-probe** | **Summe[7]** | **Anmerkung** |
|---|---|---|---|
| HR000JJ | Kern | 84.358.363 | Zur Verwendung bei Berechnungen mit Jahresergebnissen. Abweichungen zu den anderen Summen folgen aus dem 2-Stufen-Vorgehen zur Bestimmung der Hochrechnungsfaktoren und sind den Komplikationen im Erhebungsablauf geschuldet. |
| HR000JQ | Kern | 82.216.296 | Zur Verwendung bei Berechnungen mit Quartalsdurchschnitt zur Jahresauswertung. |
| HR000QQ | Kern | 328.865.187[8] | Zur Verwendung bei Berechnungen mit Quartalsergebnissen. Werden nicht wie hier die Jahresergebnisse verwendet, sondern nur die Daten eines Quartals summieren sich die hochgerechneten Werte auf den Quartalsdurchschnitt, der bei Verwendung des Hochrechnungsfaktors HR000JQ resultiert. |
| HR100JJ | LFS | 82.206.996 | Zur Verwendung bei Berechnung mit Jahresergebnissen. |
| HR100JQ | LFS | 82.206.996 | Zur Verwendung bei Berechnungen mit Quartals-durchschnitt zur Jahresauswertung. |
| HR100QQ | LFS | 328.827.981[9] | Zur Verwendung bei Berechnungen mit Quartalsergebnissen. Werden nicht wie hier die Jahresergebnisse verwendet, sondern nur die Daten eines Quartals summieren sich die Werte auf 82.206.996. |

[7] Vollmaterial
[8] Da jedes Quartal einzeln berechnet wird, entspricht der über das ganze Jahr summierte Wert dem Vierfachen der Bevölkerungszahl.
[9] Da jedes Quartal einzeln berechnet wird, entspricht der über das ganze Jahr summierte Wert dem Vierfachen der Bevölkerungszahl.

| | | | |
|---|---|---|---|
| HR100MO | LFS | 62.038.285 | Zur Verwendung bei Berechnungen von Variablen des LFS-Ad-Hoc-Moduls. |
| HR100BH | LFS | 10.482.824 | Zur Verwendung bei Berechnungen mit Jahresergebnissen für Menschen mit Behinderungen. |
| HR200JJ | SILC | 82.174.512 | Zur Verwendung bei Berechnungen des SILC-Querschnitts. Die SILC-Erhebung findet in einem begrenzten Zeitraum des Jahres statt, deshalb gibt es keine weiteren Hochrechnungsfaktoren. |

**Wieso unterscheiden sich die hochgerechneten Gesamtfallzahlen zwischen verschiedenen Variablen bzw. zwischen verschiedenen Hochrechnungsfaktoren?**

Die Unterschiede in den hochgerechneten Gesamtfallzahlen kommen aus verschiedenen Gründen zustande. Differenzen zwischen Jahres- und Quartalshochrechnungsfaktoren können sich daraus ergeben, dass in die Jahreshochrechnungsfaktoren tendenziell mehr für die Bevölkerung bekannte Eckwerte (z.B. Altersgruppen, Geschlecht, Staatsangehörigkeit sowie regionale Verteilungen) eingehen, als in die Quartalshochrechnungsfaktoren. Genauere Informationen finden sich in Metadatenreport Teil I Statistik in Kapitel 2.6 sowie in Schmidt und Stein 2021.

Abweichungen der LFS-Stichprobe kommen durch europäische Konsistenzanforderungen zustande. Diese sehen vor, dass Quartals- und Jahresergebnisse der LFS-Unterstichprobe miteinander konsistent sein müssen. Diese Anforderungen wurden gegenüber der nationalen Anforderung (Ergebniskonsistenz zwischen den einzelnen Unterstichproben) präferiert umgesetzt. Um die europäischen Anforderungen zu erfüllen, werden für amtliche Veröffentlichungen die LFS-Strukturmerkmale (sprich LFS-Jahresergebnisse) am MZ-Kern-Quartalsdurchschnitt hochgerechnet. Das MZ-Kern-Jahresergebnis entspricht nicht dem Quartalsdurchschnitt, sondern wurde eigenständig hochgerechnet. Aufgrund der Einschränkungen bei der Erhebung 2020 wird in amtlichen Veröffentlichungen für die MZ-Kern-Jahresergebnisse nicht auf den Quartalsdurchschnitt zur Hochrechnung zurückgegriffen.

**Warum sind einige Variablen weniger stark besetzt als andere?**

Mit Umsetzung der Neuerungen ab dem Mikrozensus 2020 wurden Unterstichproben eingeführt, die zur Folge haben, dass nicht alle Fragen allen zu befragenden Haushalten gestellt werden. Für den Mikrozensus werden 1% der deutschen Bevölkerung befragt. Allen Befragten in Privathaushalten werden unabhängig von der Zugehörigkeit zu einer Unterstichprobe die Fragen des Mikrozensus-Kernprogramms gestellt. Gemäß Mikrozensusgesetz sollen in 45% der Auswahlbezirke Befragungen zur

Arbeitsmarktbeteiligung (LFS) durchgeführt werden. Der Anteil zu Einkommen und Lebensbedingungen (SILC) macht etwa 12% aus. Die realisierten Auswahlsätze können davon abweichen. Die beiden genannten Substichproben sind überschneidungsfrei. Entsprechend geringer als bei Kern-Variablen sind LFS- und SILC-Variablen belegt. Ebenso ist zu beachten, dass einige Fragen freiwillig sind und die Anzahl der Beobachtungen deshalb geringer ausfallen kann. Die Problematik des Mikrozensus 2020, welche in Kapitel 2.1 näher erläutert wird, trägt ebenfalls zu vermehrten Fehlwerten bei.

Der Artikel „Die Neuerungen des Mikrozensus ab 2020“ enthält weitere Informationen zu den Erhebungsteilen. Informationen zu den Unterstichproben, sowie zur Anzahl an freiwilligen und auskunftspflichtigen Fragen je Befragungsprogramm sind Qualitätsbericht zum Mikrozensus 2020 zu entnehmen.

**Warum beträgt die Anzahl der Befragten weniger als 1% (etwa 831 600) der deutschen Bevölkerung?**

Hauptgrund für die geringere Zahl der Befragten Personen und Haushalte im Erhebungsjahr 2020 ist die hohe Ausfallquote. Diese ist auf technische Schwierigkeiten und die pandemische Lage zurückzuführen.

Erläutert wird die Problematik und die Folgen für die Auswertbarkeit des Datenmaterials in Abschnitt 2.1. des hier vorliegenden Reports sowie in „Die Neuerungen des Mikrozensus ab 2020“ und dem Qualitätsbericht zum Mikrozensus 2020.

**Was muss beim Abgleich der Daten mit den Tabellen der Fachserie beachtet werden?**

Um die Haushaltstabellen und Tabellen der Lebensformen der Fachserien nachzubilden, müssen die Daten nach Hauptwohnsitzhaushalten gefiltert werden. In der Fachserie werden nur Haushalte am Hauptwohnsitz betrachtet. Diese können mittels der Variable TH0201H selektiert werden. Gleiches gilt für Ergebnisse von Eurostat zu den Erhebungsteilen SILC und LFS. Bei Tabellen zum Migrationsstatus von Personen wird nicht durchgängig nach Hauptwohnsitzen gefiltert.

Für Tabellen in denen das Alter verwendet wird, wird die Altersvariable TPALTER_1 verwendet.

Für das Geschlecht wird TPGeschlecht verwendet. Personen mit diversem oder ohne Eintrag des Geschlechts in das Personenstandsregister sind in dieser Variablen mit einer Wahrscheinlichkeit von je 50% männlich oder weiblich zugeordnet. Bedingt durch die Ziehung der 70% Substichprobe ergeben sich Abweichungen zwischen den Häufigkeiten der Variablen

des MZ-SUF und den in den Fachserien des Statistischen Bundesamtes veröffentlichten Zahlen bzw. der Original-Mikrozensusdaten. Die meisten Variablen des SUF sollten nur in geringem Maße von den veröffentlichten Daten abweichen. Größere relative Abweichungen können sich bei Merkmalen ergeben, die mit sehr geringen Fallzahlen besetzt sind.

**Wie werden die Systemfiles für die Programmpakete SPSS, SAS und Stata erstellt?**

Die vom GML der GESIS bereitgestellten Setups für das Mikrozensus SUF 2020 dienen zum Einlesen des Rohdatenmaterials und zum Erstellen von Systemfiles für die Programmpakete SPSS, SAS und Stata. Sie beinhalten Programmanweisungen zur Definition von fehlenden Werten sowie zum Versehen der Variablen und ihrer Ausprägungen mit entsprechenden Labels. Die Setups werden im Dateiformat PC, Dos/Windows angeboten. Eine Umsetzung auf das Dateiformat Unix kann zum Beispiel mit Notepad++ oder Textpad++ durchgeführt werden.

*Spezifika der Statistikprogramme*:

- Am Anfang der Setups sind in der **Configuration Section** die vollständigen lokalen Dateinamen (einschließlich Laufwerkskennzeichen und Verzeichnis) zu nennen. Ansonsten sind im Setup keine weiteren Änderungen vorzunehmen.

- Stata: Das Setup steht mit der Zeichencodierung Unicode (UTF-8) bereit und ist mit Stata ab Version 14 ablauffähig. Für die Verwendung mit älteren Versionen kann es z. B. mithilfe von MS-Edit in die Zeichencodierung ANSI bzw. Windows-1252 umgesetzt werden.
- SPSS: Das Setup mit der Zeichencodierung Windows-1252 ist sowohl mit Version 24 als auch mit älteren Versionen ablauffähig, wenn entsprechend unter „Bearbeiten“ | „Optionen“ | „Sprache“ | „Zeichencodierung ...“ | „[x] Schriftsystem der Ländereinstellung ...“ eingestellt ist.
- Missing Values: In SPSS können vorliegende Werte als benutzerdefinierte Missings deklariert werden. Sie werden bei Auszählungen mit entsprechenden Werten und Labels ausgewiesen, zählen aber i. d. R. bei statistischen Modellen nicht als gültige Werte. In SAS und Stata können dagegen vorliegende Werte nur nach Recodierung als fehlende Werte definiert werden. Im SPSS-Setup werden benutzerdefinierte Missings spezifiziert. Diese Definitionen sind zwar auch in den SAS- und Stata-Setups enthalten, im Unterschied zum SPSS-Setup jedoch auskommentiert, da sonst die Originalwerte der Rohdaten (-1, …, -9) im Systemfile durch benutzerdefinierte Missing-Zeichen (.a, …, .h) ersetzt würden. Bei Bedarf können diese im Setup unter dem Kommentar „Definition of user-missing values“

stehenden Programmanweisungen durch Entfernen der Kommentarzeichen (/*, */) am Beginn und Ende des Anweisungsblocks aktiviert werden.

- Voreinstellungen: Um eine reibungsfreie Aufbereitung der Daten zu gewährleisten, empfiehlt es sich, die in den Setups vorgesehenen Voreinstellungen nicht zu verändern.

Ist das entsprechende Systemfile erstellt, kann die einfache Fallzahl n = 477.079 (ohne Gewichtung, ohne Selektion) zur Kontrolle, ob der Rohdatensatz fehlerfrei eingelesen wurde, mit der des erstellten Datensatzes verglichen werden. Zusätzlich können die Eckwerte aus Abschnitt 2.4 zur Prüfung des fehlerfreien Ablaufs des Setups herangezogen werden. Hierbei ist zu beachten, dass sich die Eckwerte auf die hochgerechnete Bevölkerung (HR000JJ) am Hauptwohnsitz (TH0201H=1) beziehen.

## 3.3 Verfügbare Tools

Syntaxen zur Umsetzung der sozialwissenschaftlichen Konzepte ESeG - European Socioeconomic Groups, ESeC - European Socioeconomic Classification, ISEI - Internationaler Sozioökonomischer Index des beruflichen Status und CASMIN-Bildungsklassifikation stehen für die Programmpakete SPSS und Stata auf dem Mikrodaten-Informationssystem (MISSY) der GESIS zur Verfügung. Für das Erhebungsjahr 2020 werden einmalig keine Syntaxen zur Umsetzung von ESeG, ESeC und ISEI angeboten, da die amtliche Statistik aufgrund von Qualitätsbedenken von einer Veröffentlichung der ISCO-Variablen im MZ-SUF 2020 absieht, die als Grundlage für die Operationalisierung der o.a. Tools dienen.

# Anhang

## SPSS-Syntax zu Abschnitt 2.4

```
*Filtern nach Bevölkerung am Hauptwohnsitz*
if (TH0201H=1) Hauptwohnsitz=1.
filter by Hauptwohnsitz.
EXECUTE.

WEIGHT HR000JJ.
EXECUTE.
**********************************************************************************************************
*Häufigkeitstabelle für Geschlecht*

VARIABLE LABELS TPGeschlecht 'Geschlecht'.
VALUE LABELS TPGeschlecht
1 'Männlich'
2 'Weiblich'.
FREQUENCIES TPGeschlecht.
EXECUTE.
**********************************************************************************************************
*Häufigkeitstabelle für Familienstand*

VARIABLE LABELS AB0500P 'Familienstand'.
FREQUENCIES AB0500P.
EXECUTE.
**********************************************************************************************

COMPUTE Haushaltsmitglieder2 = 0.
if (NpersHH =1) Haushaltsmitglieder2=1.
if (NpersHH =2) Haushaltsmitglieder2=2.
if (NpersHH =3) Haushaltsmitglieder2=3.
if (NpersHH =4) Haushaltsmitglieder2=4.
if (NpersHH >4) Haushaltsmitglieder2=5.
FREQUENCIES Haushaltsmitglieder2.
EXECUTE.
```

*Häufigkeitstabelle für Personenhaushalte*

```
VARIABLE LABELS Haushaltsmitglieder2 'AnzahlHaushaltsmitglieder'.
VALUE LABELS
Haushaltsmitglieder2
1 'Einpersonenhaushalt'
2 'Zweipersonenhaushalt'
3 'Dreipersonenhaushalt'
4 'Vierpersonenhaushalte'
5 'Fünf- und mehr Personenhaushalte'.
EXECUTE.
FREQUENCIES Haushaltsmitglieder2.
EXECUTE.
```

**********

*Häufigkeitstabelle für Bundesland*

```
VARIABLE LABELS Land 'Bundesland'.
FREQUENCIES Land.
EXECUTE.
```

Statistische Ämter des Bundes und der Länder,
Metadatenreport – Teil II: Produktspezifische Informationen zur Nutzung des Mikrozensus Scientific Use Files 2020