

Wiesbaden, 23. Mai 2025

**Konzept zur Anonymisierung des Zensus 2011
zur Verwendung als Public Use File (PUF)
für Integrated Public Use Microdata Series International (IPUMS International)
mit Nachnutzung über den nationalen Datenzugang**

1. Vorbemerkungen

Das Projekt Integrated Public Use Microdata Series International (IPUMS International) wurde mit dem Ziel ins Leben gerufen, internationale Zensus- und Surveydaten in harmonisierter, dokumentierter und anonymisierter Form der internationalen Forschungslandschaft zur Verfügung zu stellen. IPUMS international ist eine Zusammenarbeit zwischen der Universität Minnesota, nationaler Statistikämter weltweit, internationaler Datenarchive und anderen internationalen Organisationen. Das Projekt wird vom Minnesota Population Center der Universität Minnesota koordiniert und von der U.S. National Science Foundation und dem National Institute of Child Health and Human Development gefördert.

Aktuell umfasst das Datenangebot von IPUMS International 656 Zensen und Surveys aus 104 Ländern.¹ Für Deutschland wurden bisher 4 absolut anonymisierte Mikrodatsätze bereitgestellt: Die Volkszählungen von 1970 und 1987 der alten Bundesrepublik sowie die Volkszählungen von 1971 und 1981 der ehemaligen Deutschen Demokratischen Republik. Darüber hinaus sind fünf absolut anonymisierte Mikrodatsätze aus den Mikrozensen der Bundesrepublik aus den Jahren 1973, 1982, 1987, 1991 und 2001 über das Projekt Integrated European Population Microdata (in Kooperation mit IPUMS International) verfügbar.

Um das Datenangebot für international vergleichende Forschung mit Bezug zu Deutschland zu verbessern und auch im außereuropäischen Vergleich zu ermöglichen, sollen Daten aus dem Zensus 2011 in IPUMS International integriert werden. Ein entsprechendes Memorandum of Understanding wurde am 9. Oktober 2024 vom

¹ <https://international.ipums.org/international/>

Statistischen Bundesamt und IPUMS International – vertreten durch die Universität Minnesota – unterzeichnet.

Das vorliegende Konzept befasst sich mit der Anonymisierung der Mikrodaten aus dem Zensus 2011. Als Erhebung der amtlichen Statistik unterliegt der Zensus im Hinblick auf seine Nutzung gesetzlichen Bestimmungen. Das Gesetz über die Statistik für Bundeszwecke (kurz: Bundesstatistikgesetz – BStatG) gestattet nach § 16, Absatz 1 die Weitergabe von Einzelangaben nur dann, wenn die Einzelangaben dem Befragten nicht zuzuordnen sind. In diesem Fall wird von absoluter Anonymität gesprochen.

Ausgehend von den Anonymisierungsmethoden, welche bei der Erstellung von PUFs auf Basis der Volkszählung 1987 (BRD) sowie verschiedener Mikrozensen angewandt wurden, werden im Folgenden die Anonymisierungsmaßnahmen beschrieben, die zur absoluten Anonymität der Einzeldaten im PUF auf Basis des Zensus 2011 führen. Diese orientieren sich an den Empfehlungen von Südfeld (1987)², die bei der Übermittlung absolut anonymisierter Einzelangaben durch das Statistische Bundesamt zu prüfen sind:

- „Der zu übermittelnde Datenbestand soll nur noch eine Stichprobe aus der jeweiligen Statistik sein. (...)“
- Der zu übermittelnde Datenbestand soll ein bestimmtes Mindestalter aufweisen. In der Regel sollen die Angaben durch eine neue Erhebung bereits überholt sein. (...)“
- Die Anordnung der Datensätze im zu übermittelnden Datenbestand muß systemfrei sein. (...)“
- Direkte Identifikationsmerkmale (Namen und Anschrift, Telefonnummer, Personen und sonstige Kennnummern) dürfen im Datenbestand nicht enthalten sein. (...)“
- Regionalangaben sollen nur als Typisierungsangaben im Datenbestand belassen werden, allerdings nur insoweit, als dadurch keine Regionaltypen mit weniger als 500 000 Einwohnern identifiziert werden. (...)“

² Südfeld, E.: Anonymisierungsstandards und generelle Abwicklungsregelungen für Anforderungen nach anonymisierten Einzelangaben im Statistischen Bundesamt. In: „Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik – Bedingungen und Möglichkeiten“, Schriftenreihe Forum der Bundesstatistik, Band 5, S.146 – 156, Wiesbaden 1987.

- Jede Ausprägung eines einzelnen Merkmals soll mindestens fünffach besetzt sein. (...)
- Sensible Merkmale (z. B. Daten zur Gesundheit, zur Einkommens- und Vermögenslage u. a. m.) sollen allenfalls klassifiziert übermittelt werden. (...)
- Merkmale, über die sehr einfach Zusatzinformationen zu erhalten sind (z. B. Geburtsdatum u. a. m.), sollen ebenfalls nur klassifiziert übermittelt werden. (...)
- Die Kombination sensibler Merkmale und von Merkmalen, über die sehr einfach Zusatzinformationen zu erhalten sind, sollen mindestens dreifach besetzt sein.“

2. Basismaterial

Grundlage für die Erstellung des PUFs bilden die Einzeldaten des Zensus 2011. Der Zensus 2011 wurde registergestützt durchgeführt. Informationen aus vorhandenen Verwaltungsregistern wurden mit Informationen aus Vollerhebungen und Stichprobenerhebungen ergänzt bzw. korrigiert. Stichtag des Zensus war der 9. Mai 2011. Die Einzeldaten des Zensus 2011 bestehen aus ca. 80,2 Mio. Personendatensätzen aus den Registern. Die Haushaltsstichprobe umfasst ca. 7,5 Mio. Personendatensätze, die in die Hochrechnung eingehen. Hinzukommen Informationen zu ca. 18,9 Mio. Gebäuden und ca. 40,5 Mio. Wohnungen. Der PUF enthält Daten aus der Haushaltsstichprobe, die um Informationen aus den Personendatensätzen sowie den Daten zu Gebäuden und Wohnungen ergänzt werden. Personen stellen die statistische Einheit dar.

3. Anonymisierungsmaßnahmen

Die folgenden Anonymisierungsmaßnahmen, durchgeführt am Basismaterial, führen zur absoluten Anonymität des PUFs. Nach Anwendung der Maßnahmen ist eine Zuordnung von Einzelangaben zu Merkmalsträgern im PUF nach menschlichem Ermessen ausgeschlossen.

a) Entfernen direkter Identifikationsmerkmale

Personenbezogene Angaben, wie Name und Anschrift, Telefonnummer usw. sind für die Erstellung eines absolut anonymisierten Datenfiles aus dem Zensus nicht von Bedeutung, da diese Merkmale bereits im Aufbereitungsprozess aus dem Datenmaterial entfernt wurden. Zusätzlich wurden Angaben zum Geburtsdatum aus dem Datensatz entfernt.

b) Entfernen von Variablen

In den absolut anonymisierten Datensatz wird als einzige Regionalinformation der Wohnort vergrößert nach drei Bundeslandgruppen aufgenommen (Variable „GEBIET“, siehe Abschnitt zu Merkmalsvergrößerung). Darüber hinaus werden keine weiteren Regional- und Anschriftsinformationen zu Wohn- und Arbeitsort aufgenommen. Die entfernten Variablen umfassen Name, Regionalschlüssel, Einwohnerzahl sowie Fläche von Gemeinde, Verwaltungsverband/ Samtgemeinde, Landkreis, Regierungsbezirk und Bundesland. Auch alle Gitterzelleninformationen werden aus dem Datensatz entfernt.

c) Substichprobenziehung

Eine weitere Methode zur Verringerung des Deanonymisierungsrisikos besteht in der Ziehung und Weitergabe einer Substichprobe des vorliegenden Mikrodatenfiles. Datenfiles, die aus Stichproben bestehen, weisen ein wesentlich geringeres Deanonymisierungsrisiko auf als Vollerhebungen. Durch den Auswahlatz in der Haushaltsstichprobe des Zensus von 9,6% der Gesamtbevölkerung sind diese Datenbestände bereits sehr gut vor Reidentifikationsversuchen geschützt. Die maximale Erfolgswahrscheinlichkeit, eine in der Stichprobe enthaltene Person zu reidentifizieren, entspricht dabei dem Auswahlatz der Stichprobe. Um das Reidentifikationspotenzial weiter zu reduzieren wird eine zufällige Substichprobe gezogen. Angesichts der weiteren Anonymisierungsmaßnahmen ist eine 50%-Substichprobe aus dem Originalmaterial der Haushaltsstichprobe zur Erstellung eines absolut anonymisierten Datenfiles als ausreichend anzusehen. Die Substichprobe umfasst somit ca. 4,8% der Gesamtbevölkerung und ist damit in etwa vergleichbar mit der 5%-Stichprobe, die auf Basis der BRD-Volkszählung 1987 bereits an IPUMS International geliefert wurde. Die Substichprobenziehung erfolgt geschichtet nach den drei Gebietseinheiten (Nord, Ost, Süd) (siehe Abschnitt zu Merkmalsvergrößerung).

d) Systemfreie Sortierung

Als eine weitere Anonymisierungsmaßnahme wurde der Datensatz systemfrei (d.h. nach einem nicht nachvollziehbaren System) sortiert und anschließend die Variablen Gebäude, Haushalt, Wohnung sowie Person mit einer eindeutigen systemfreien Nummerierung versehen.

e) Merkmalsvergrößerung

Ziel von Merkmalsvergrößerungen ist es, die Fallzahlen schwach besetzter Merkmalsausprägungen durch Zusammenfassung einzelner Ausprägungen zu erhöhen. Durch schwach besetzte Zellen ist es möglich Einzelfälle zu identifizieren und gezielt Zusatzwissen zu beschaffen, so dass ein einfacher Abgleich der Informationen möglich wird. Werden Ausprägungen zusammengefasst, verringert sich das Reidentifikationsrisiko, da die Wahrscheinlichkeit von einzigartigen Ausprägungskombinationen sinkt. Mit einer Vergrößerung ist ein Verlust an Informationen verbunden, der die Erstellung eines absolut anonymen Datenmaterials ermöglicht. Dieser Informationsverlust ist umso größer, je gröber die Einteilung der Kategorien vorgenommen wird.

Analog zur absoluten Anonymisierung des Mikrozensus wird zunächst der Wohnort in eine dreigeteilte Gebietsgliederung Deutschlands (Nord, Ost und Süd) vergrößert. Die Gebietseinheit Nord umfasst die Bundesländer Schleswig-Holstein, Hamburg, Niedersachsen, Bremen und Nordrhein-Westfalen. Die Gebietseinheit Ost umfasst die Bundesländer Berlin, Brandenburg, Mecklenburg-Vorpommern, Sachsen, Sachsen-Anhalt und Thüringen. Die Gebietseinheit Süd umfasst die Bundesländer Hessen, Rheinland-Pfalz, Baden-Württemberg, Bayern und Saarland. Die drei Gebietseinheiten umfassen jeweils mindestens 15,7 Mio. Einwohner.

Der Wohnort nach den drei Gebietseinheiten stellt die einzige Regionalinformation dar, die im PUF enthalten ist. Insbesondere wird auf das Merkmal Gemeindegrößenklasse verzichtet, um die absolute Anonymisierung bei einem gleichzeitig vertretbaren Variablenumfang zu gewährleisten.

Für alle weiteren Merkmalsausprägungen gilt eine Mindestbesetzung von 10 000 Einwohnern je Gebietseinheit (Nord, Ost, Süd) in der Grundgesamtheit, d.h. es müssen nach Hochrechnung mindestens 10 000 Fälle je Gebietseinheit gegeben sein. Für alle Variablen, die diesem Kriterium nicht entsprechen, wurden ggf. vorhandene gruppierte Variablen geprüft. Diese gruppierten Variablen, z. B. Baujahr (Gebäude) in 10-Jahres-Schritten, gehen in den PUF ein, sofern sie den Anonymisierungskriterien entsprechen. Darüber hinaus wurde für folgende Variablen eine Merkmalsvergrößerung durchgeführt, um die beschriebene Mindestbesetzung zu gewährleisten:

- Alter
- Geburtsregion (EU-Definition)

- Familienstand
- Zuzugsjahr
- Herkunftsregion bei Zuzug aus dem Ausland
- Herkunftsregion der Mutter bei Zuzug aus dem Ausland
- Herkunftsregion des Vaters bei Zuzug aus dem Ausland
- Größe des privaten Haushalts
- Typ des privaten Haushalts (nach Kindern)

Alle Variablen, die auch nach Vergrößerung nicht obigen Anonymisierungskriterien entsprechen, wurden aus dem Datensatz entfernt.

4. Merkmalsliste

Untenstehende Liste umfasst alle Merkmale, die im PUF enthalten sind. Bereits im Basismaterial gruppierte Merkmale und für den PUF vergrößerte Merkmale sind entsprechend gekennzeichnet. Nähere Informationen zu allen enthaltenen Merkmalen und Merkmalsausprägungen finden sich im Schlüsselverzeichnis (IPUMS_Zensus2011_Variables.xlsx). Insgesamt gehen 43 von 344 Merkmalen aus dem Basismaterial in den PUF ein.

<i>Merkmale aus Haushaltsstichprobe</i>	
GEBIET (v)	Region (Nord, Ost, Süd)
GESCHLECHT	Geschlecht
ALTER (v)	Alter
STAATSANGE_GRP (g)	Staatsangehörigkeit
RELIGION_AUSF	Religion
FAMSTAND_KURZ (v)	Familienstand
MIGRATION_KURZ (g)	Migrationshintergrund
MHGLAND_KURZ (g)	Migrationshintergrund (Regionen)
ZUZUG	Zuwanderung/Zuzug nach 1955 in das heutige Gebiet der Bundesrepublik Deutschland
ZUZUG_JAHR_GRP (v)	Zuzugsjahr
ZUZUG_STAAT_GRP (v)	Herkunftsregion bei Zuzug aus dem Ausland

ZUZUG_MUTTER	Zuwanderung/Zuzug der Mutter
ZUZUG_VATER	Zuwanderung/Zuzug des Vaters
ZUZUG_STAAT_MUTTER_GRP (v)	Herkunftsregion der Mutter bei Zuzug aus dem Ausland
ZUZUG_STAAT_VATER_GRP (v)	Herkunftsregion des Vaters bei Zuzug aus dem Ausland
SCHUELER	Schulbesuch in der Woche vom 9. bis 15. Mai 2011
SCHUELER_SCHULFORM	Schulform
BILDUNG_ISCED	ISCED-Bildung
ERWERBSTAT_AUSF	Erwerbsstatus
BERUF_ISCO_HAUPTGR	Erwerbstätige nach Berufshauptgruppen ISCO-08
BERUF_KLDB_HAUPTGR	Beruf (Hauptgruppen KldB 2010 - 2-Steller)
WZ_UNTERBER_STP	Erwerbstätige nach Wirtschaftszweig
ARBEITSORT_AUSP	Arbeitsort nach Auspendlern
EU_CAS	Erwerbsstatus (EU-Definition)
EU_HST	Stellung im Haushalt (EU-Definition)
EU_POB_GRP (v)	Geburtsregion (EU-Definition)
EU_ROY (g)	Üblicher Aufenthaltsort ein Jahr vor der Zählung (EU-Definition)
EU_SIE	Stellung im Beruf (EU-Definition)
<i>Merkmale aus Registern</i>	
HH_GROESSE_PERSONEN_KURZ (v)	Größe des privaten Haushalts
HH_TYP_FAMILIE	Typ des privaten Haushalts (nach Familien)
HH_TYP_LEBENSFORMEN	Typ des privaten Haushalts (nach Lebensform)
HH_TYP_SENIORENSTATUS	Typ des privaten Haushalts (nach Senioren)
HH_TYP_KINDER_KURZ (v)	Typ des privaten Haushalts (nach Kindern)
HH_TYP_MIGRATIONSHINTERGRUND	Typ des privaten Haushalts (nach Migrationshintergrund)
<i>Merkmale aus Gebäude- und Wohnungszählung</i>	
BAUJAHR_10JA (g)	Baujahr (Jahrzehnte)

EIGENTUM_GEB	Eigentumsform des Gebäudes
GEBTYPBAUWEISE	Gebäudetyp-Bauweise
HEIZTYP	Heizungsart
ZAHLOWHNUNGEN_5KL (g)	Zahl der Wohnungen im Gebäude (5 Klassen)
AUSSTATTUNG	Ausstattung der Wohnung
WOHNEIGENTUM	Eigentumsverhältnisse der Wohnung
RAUMANZAHL_KLASS (g)	Zahl der Räume (7 Klassen)
WOHNFLAECHE_10S (g)	Fläche der Wohnung (10 m ² -Intervalle)
<i>Hochrechnungsfaktoren</i>	
HRF_sub	Hochrechnungsfaktor Personen ³

Zeichenerklärung: g=bereits gruppiertes Merkmal; v=Für PUF vergrößertes Merkmal

³ Der Hochrechnungsfaktor HRF_sub entsteht durch Anpassung des Hochrechnungsfaktors HRF_ZUSATZ (Haushaltsstichprobe) an die für den PUF gezogene Unterstichprobe. Durch die Hochrechnung der zugespielten Merkmale aus Registern sowie Gebäude- und Wohnungszählung anhand des Hochrechnungsfaktors aus der Haushaltsstichprobe können die Häufigkeiten dieser Merkmale von bereits veröffentlichten Ergebnissen abweichen.