1. Allgemeine Hinweise zur Anonymisierung des Remote Access SUF (Berichtsjahre ab 2020)

Das Datenprodukt Remote Access SUF (Remote SUF) ordnet sich hinsichtlich des Analysepotentials und des Anonymisierungsgrades zwischen den bestehenden Produkten Off-Site-SUF und On-Site-Daten ein. Wie beim Off-Site-SUF handelt es sich auch beim Remote SUF um faktisch anonyme Daten. Solche Daten dürfen gemäß den Vorgaben des Bundesstatistikgesetzes grundsätzlich an wissenschaftliche Einrichtungen übermittelt werden (vgl. BStatG § 16 (6)). Faktisch anonym heißt, dass die Einzelangaben nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können.

2. Anonymisierungsmaßnahmen am Datenmaterial

Die Basis für den Remote SUF bilden die formal anonymen Daten des Mikrozensus (On-Site-Daten[1]). Um faktische Anonymität zu erreichen, werden am Remote SUF folgende Maßnahmen ergriffen:

2.1. Entfernung räumlicher Informationen

Die tiefste ausgewiesene räumliche Gliederung der Daten bilden die Anpassungsschichten. Die Anpassungsschichten sind in der Regel kreisscharf und bestehen aus einem oder mehreren meist benachbarten Kreisen oder kreisfreien Städten mit durchschnittlich 500.000 Einwohnern. Deutschlandweit gibt es aktuell 147 Anpassungsschichten.

Die notwendigen Zusammenfassungen unterscheiden sich damit in Abhängigkeit der Einwohnerzahl der Kreise und kreisfreien Städte. Das verdeutlicht ein Vergleich einiger Bundesländer: So gibt es in Nordrhein-Westfalen bspw. viele Kreise oder kreisfreie Städte, die eine eigene Anpassungsschicht bilden. Insgesamt werden aus den 53 Kreisen und kreisfreien Städten 34 Anpassungsschichten gebildet. In Rheinland-Pfalz hingegen verteilen sich die 36 Landkreise und kreisfreien Städte auf nur fünf Anpassungsschichten. Das Saarland bildet eine einzige Anpassungsschicht, das Land Bremen besteht aus zweien (Bremen und Bremerhaven). In den weiteren Stadtstaaten bilden grundsätzlich die Bezirke die Anpassungsschichten.[2] Der Remote SUF bietet damit deutlich mehr regionales Analysepotential als der Off-Site-SUF – der maximal bis zur Ebene der Bundesländer ausgewertet werden kann – und kann teilweise auch als Ersatz für On-Site-Projekte mit regionalem Fokus genutzt werden. Für das Berichtsjahr 2020 werden jedoch keine Auswertungen des Remote SUF unterhalb der Bundeslandebene empfohlen.

Alle räumlichen Informationen, die unterhalb der Anpassungsschichten liegen oder in Kombination mit der Anpassungsschicht Rückschlüsse auf eine darunterliegende Regionalinformation geben, wurden aus den Daten entfernt. Hierzu zählen Regions- und Kreistypen, Gemeindegrößenklassen, Arbeitsmarktregionen, Arbeitsagenturbezirke, Gemeindetyp und die Stadt-Land-Gliederung nach Eurostat.

[1] Grundsätzliche Informationen zu den On-Site-Daten des Mikrozensus finden sich in den Metadatenreports zur Statistik (Teil I) und zum Produkt (d.h. zum jahresspezifischen Datenangebot des Mikrozensus, Teil II): https://www.forschungsdatenzentrum.de/de/haushalte/mikrozensus

[2] In Hamburg sind Harburg und Bergedorf zu einer Anpassungsschicht zusammengefasst.

2.2. Vergröberungen

Zusätzliche inhaltliche Eingriffe in das Datenmaterial bilden einen weiteren Schritt, um faktische Anonymität zu erreichen. Das Vorgehen ist hierbei von zwei Überlegungen geleitet. Zum einen sollen im Datensatz grundsätzlich keine Merkmalsausprägungen vorhanden sein, für welche lediglich Daten von weniger als 3 verschiedenen Personen vorliegen. Bei SILC-Haushaltsvariablen beträgt der Grenzwert 3 Haushalte, bei Variablen mit unklassierten Eurobeträgen 10 Haushalte beziehungsweise Personen (vgl. Abschnitt 4.2). Zum anderen muss in Betracht gezogen werden, dass bestimmte Merkmale in der im Mikrozensus vorliegenden Form das Risiko einer Deanonymisierung stärker als andere erhöhen. Als Beispiele können hier die Angaben zur Nationalität oder auch Informationen zu den Berufen genannt werden.

Bei den Variablen, die Merkmalsausprägungen mit lediglich einem oder zwei Fällen aufweisen, handelt es sich im Wesentlichen um Staatsangehörigkeiten, Herkunftsländer, Berufs-, Hauptfachrichtungs- oder Wirtschaftszweiginformationen sowie Jahresangaben (siehe auch die Spalte „Maßnahme im Remote SUF“ im Schlüsselverzeichnis[3]). In vielen Fällen sind dies auch Merkmale, die ein höheres Deanonymisierungsrisiko mit sich bringen. Diese Merkmale werden daher durch Zusammenfassungen vergröbert, wobei sich das Vorgehen im Wesentlichen am Vorgehen beim bisherigen Off-Site-SUF orientiert. D.h. bei den aufgeführten Variablen wird die gesamte Variable auf hochgerechnet 5.000 Fälle vergröbert, im Falle der Angaben zu Nationalitäten auf 50.000 Fälle hochgerechnet auf die Grundgesamtheit.[4] Bei den Berufen wird auf einen künstlichen Viersteller zurückgegriffen, der auch im Off-Site-SUF verwendet wird und das Anforderungsniveau (5. Stelle der KldB 2010) soweit wie möglich erhält.

In einzelnen Fällen, die nur in den Randbereichen zu Einzelwerten oder sehr geringen Fallzahlen führen, wird ein Bottom- oder Top-Coding angewendet.

3. Unterschiede zwischen Remote SUF und On-Site-Material aufgrund der Anonymisierungsmaßnahmen

Im Schlüsselverzeichnis des Remote SUF ist über die Spalte „Maßnahme im Remote SUF“ grundsätzlich erkennbar, wie mit jeder der aufgeführten On-Site-Variablen verfahren wird. So lässt sich erkennen, ob für ein Projekt der Remote SUF infrage kommt, oder ob ein Rückgriff auf das On-Site-Material notwendig ist. Relevante Unterschiede sind hier überblicksartig zusammengefasst.

[3] Hier herrscht auch über die Erhebungsjahre weitgehende Kontinuität bei den kontinuierlich abgefragten Variablen.

[4] Wie genau im jeweiligen Berichtsjahr bei den Variablen, die vergröbert werden müssen, zusammengefasst wird, lässt sich aus dem Datenhandbuch zum Off-Site-SUF erkennen. Die Datenhandbücher finden sich auf den Seiten der Forschungsdatenzentren (hier das Beispiel für den MZ 2022):
https://www.forschungsdatenzentrum.de/de/10-21242-12211-2022-00-00-3-1-1
Im Schlüsselverzeichnis zum Remote SUF sind die Variablen oder die betroffenen Bereiche blau hervorgehoben.

3.1. Umfang des Datenmaterials

Die regionale Tiefe ist im Vergleich zum On-Site-Material eingeschränkt. Dies betrifft wie oben beschrieben die Variablen mit Regionalinformationen unterhalb der Anpassungsschichten. Einige Variablen sind darüber hinaus nicht im Remote SUF aber im On-Site-Material enthalten:

- 5-Steller der Berufsgattungen
- die Hauptfachrichtungen (7-Steller und Stellen 5-7)
- 2-Steller Meisterausbildung
- Hinterbliebenenrente

Infolge von notwendigen Vergröberungen sind einige Originalvariablen aus den Daten entfernt worden, da sie sonst Vergröberungen aufheben würden. Diese Variablen können grundsätzlich aus den vergröberten Daten selbst generiert werden. Es ist dabei zu beachten, dass sich durch die Vergröberungen in Einzelfällen auch auf den ersten Stellen Abweichungen zu den Originaldaten ergeben.

- 1- bis 3-Steller bei den Berufen
- 1- bis 3-Steller bei den Berufsgattungen

Im On-Site-Material sind teilweise Merkmale zu Haupteinkommensbeziehern oder den Bezugspersonen in eigenständigen Variablen vorhanden. Diese werden im Remote SUF nicht bereitgestellt, können aber aus dem Datenmaterial selbstständig generiert werden.

3.2. Vergröberungen

Quantitativ betrachtet ist nur ein kleiner Teil von wenigen Prozent der im Remote SUF vorhandenen Variablen von Vergröberungen nach dem oben beschriebenen Vorgehen betroffen. Die Vergröberungen konzentrieren sich auf folgende inhaltliche Bereiche:

- Berufe und Berufsgattungen
- Wirtschaftszweige
- Hauptfachrichtungen
- Arbeitsstunden
- Jahresangaben (z.B. Zuzugsjahre, Geburtsjahr)
- Staatsangehörigkeiten und Geburtsländer

4. Neue Programme Information- und Kommunikationstechnologie (IKT) und Survey of Income and Living Conditions (SILC)

4.1. Programm Information- und Kommunikationstechnologie (IKT)

Ab 2021 ist die IKT in die Mikrozensuserhebung integriert.

Analog zum Off-Site-SUF werden ausschließlich Zielvariablen in den Remote SUF aufgenommen. Bei der IKT erfolgt die Vergröberung im Remote SUF grundsätzlich wie im Off-Site-SUF. Es werden bei univariaten Verteilungen einzelne Ausprägungen inhaltlich so vergröbert, dass sie hochgerechnet auf die Grundgesamtheit (hr300pn) nicht weniger als 5 000 Fälle umfassen.

In Haushalten mit mehr als 6 Personen zwischen 16 und 74 Jahren werden alle IKT-Merkmale auf -9 gesetzt.

Für das Jahr 2021 soll keine Auswertung von IKT-Merkmalen unterhalb der Bundesebene stattfinden.

4.2. Programm Survey of Income and Living Conditions (SILC)

Ab 2020 ist die SILC in die Mikrozensuserhebung integriert.

Aus dem Unterprogramm SILC werden ebenfalls nur Zielvariablen in den Remote SUF aufgenommen. Bei Merkmalen aus dem h-File findet die Fallzahlprüfung so statt, dass vergröbert wird, wenn weniger als 3 befragte Haushalte in eine Ausprägung fallen. Merkmale aus dem d-File werden nicht aufgenommen. Diese werden dann auf mindestens 5 000 Fälle hochgerechnet auf die Grundgesamtheit zusammengefasst.

Bei unklassierten Eurobeträgen wird zunächst auf volle 100 gerundet und anschließend auf hochgerechnet mindestens 50 000 Fälle vergröbert; hierbei wird sichergestellt, dass jede Merkmalsausprägung mindestens 10 Personen (bzw. Haushalte) umfasst. Auch bei Variablen zu Staatsangehörigkeiten/Geburtsland wird auf hochgerechnet mindestes 50 000 Fälle zusammengefasst

Auch bei Variablen, die Off-Site nicht zur Verfügung stehen, werden an erforderlicher Stelle vergleichbare Vergröberungsschritte durchgeführt.