Wittmaack, Moritz

Article

Mehrfachfallprüfung im Zensus 2022 – die neue Strategie zur automatisierten Identifikation und Bewertung von Mehrfachfällen

WISTA - Wirtschaft und Statistik

Provided in Cooperation with:
Statistisches Bundesamt (Destatis), Wiesbaden

Suggested Citation: Wittmaack, Moritz (2025) : Mehrfachfallprüfung im Zensus 2022 – die neue Strategie zur automatisierten Identifikation und Bewertung von Mehrfachfällen, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 77, Iss. 1, pp. 127-143

This Version is available at:
https://hdl.handle.net/10419/313335

MEHRFACHFALLPRÜFUNG IM ZENSUS 2022 – DIE NEUE STRATEGIE ZUR AUTOMATISIERTEN IDENTIFIKATION UND BEWERTUNG VON MEHRFACHFÄLLEN

Moritz Wittmaack

Schlüsselwörter: Record Linkage – Datendeduplikation – Melderegister – amtliche Einwohnerzahl – Personenbestand

ZUSAMMENFASSUNG

Die kombinierte Mehrfachfallprüfung der Melderegisterdaten und der Fehlbestände der primärstatistischen Erhebungen ist ein vom Statistischen Bundesamt entwickeltes vollständig maschinelles Verfahren, um die Qualität der Zensusergebnisse zu sichern: Es bringt potenzielle Mehrfachfälle – mehrere dieselbe Person repräsentierende Datensätze – aus dem Personenbestand des Zensus 2022 zusammen, prüft diese anhand fester Regeln auf Zusammengehörigkeit und stellt die melderechtliche Zulässigkeit der vorliegenden Mehrfachmeldungen sicher. Besonders herausfordernd war hierbei das Fehlen eines registerübergreifenden, persistenten Personenidentifikators. Der Artikel erläutert die Hintergründe, die zur neuen Mehrfachfallprüfung geführt haben, und stellt den Ablauf des Verfahrens vor.

Keywords: *record linkage – data deduplication – population register – official population – population stock*

ABSTRACT

To guarantee the quality of the census results, the Federal Statistical Office developed a fully automated method to check for multiple entries in a data pool that combines population register data and missing data entries identified in primary data surveys. The method gathers potential multiple entries – i.e. two or more data records representing the same person – from the population stock of the 2022 Census, checks whether the entries refer to the same entity using a fixed set of rules, and ensures that the multiple entries identified are permitted under Germany's registration law. The lack of a persistent personal identifier across all registers proved to be a particular challenge in this context. This article explains the background to the introduction of the new multiple entry check and describes the individual steps of this new process.

Moritz Wittmaack

studierte Soziologie und empirische Sozialforschung in Bielefeld und Köln. Von Dezember 2019 bis Mai 2023 war er als Wissenschaftlicher Mitarbeiter im Statistischen Bundesamt mit der Konzeption sowie der Umsetzung der Mehrfachfallprüfung des Zensus 2022 betraut. Seit Juni 2023 ist er Referent im Referat „Lohn- und Einkommensteuer“.

Maßgeblich an der Entwicklung der Mehrfachfallprüfung als vollständig maschinelles Verfahren beteiligt waren Bernd Michel (†) und Paul Konstantin Schmidtke. Philipp Hadeball und Thomas Franke unterstützten die Umsetzung des entwickelten Verfahrens mittels SAS Version 9.4.

1

Einleitung

Die Mehrfachfallprüfung ist ein Prozess zur Qualitätssicherung des Personenbestands des Zensus 2022, auf dessen Grundlage die Einwohnerzahl Deutschlands abgeleitet wurde (Bretschi und andere, 2024). Für die Integration der verschiedenen Datenquellen des Personenbestands stand kein registerübergreifender, persistenter Personenidentifikator zur Verfügung. Daher konnte aufgrund der dezentralen Melderegisterführung nicht davon ausgegangen werden, dass jede zum Zensusstichtag in Deutschland lebende Person mit genau einem einwohnerzahlrelevanten Melderegisterdatensatz im Personenbestand enthalten sein würde.

Um die Qualität der Zensusergebnisse zu sichern, sah § 21 Absatz 1 Zensusgesetz 2022 deshalb vor, dass das Statistische Bundesamt anhand der von den Meldebehörden gelieferten Melderegisterdaten prüft, ob Personen mit mehr als einer alleinigen Wohnung oder Hauptwohnung oder nur mit einer oder mehreren Nebenwohnungen gemeldet sind. Bei nicht melderechtskonformen mehrfachen Meldungen einer Person war zu entscheiden, welcher Melderegisterdatensatz relevant für die Ermittlung der Einwohnerzahl ist (§ 21 Absatz 2 Zensusgesetz 2022). Außerdem war für Personen, die an Sonderbereichsanschriften wohnten, dort aber nicht gemeldet waren, der Wohnungsstatus festzustellen (§ 21 Absatz 3 Zensusgesetz 2022). Ob diese Personen zusätzlich unter gegebenenfalls abweichenden Angaben an derselben oder einer anderen Anschrift in Deutschland gemeldet waren, klärte die Mehrfachfallprüfung.

Das dazu entwickelte fehlertolerante Verfahren bringt alle dieselbe Person repräsentierenden Datensätze zusammen, prüft auf Zusammengehörigkeit und bestimmt letztlich, welcher der vorhandenen Datensätze zur Ermittlung der Einwohnerzahl herangezogen wird. Anstatt wie im Zensus 2011 strittige Fälle händisch zu beurteilen und den Wohnsitz mithilfe einer postalischen Befragung festzustellen (Diehl, 2012), wurde im Zensus 2022 ein vollständig maschinelles Verfahren angewandt.

Kapitel 2 beschreibt den Aufbau des Personenbestands, Kapitel 3 die sich aus der Art der Datenintegration ergebenden Datenfehler. Hiernach stellt Kapitel 4 das entwickelte Verfahren entlang seiner Anwendung im Zensus 2022 vor, die Kapitel 5 und 6 befassen sich mit der Suche nach prozessproduzierten beziehungsweise dauerhaften Mehrfachfällen. Anschließend untersucht Kapitel 7 die Wirkung der Mehrfachfallprüfung auf den Personenbestand. Ein kurzes Fazit beschließt den Artikel.

2

Generierung des stichtagsrelevanten Personenbestands

Entsprechend internationaler Vorgaben erfolgt die Feststellung der amtlichen Einwohnerzahl in Deutschland alle zehn Jahre durch einen Zensus. Wie der vorangegangene Zensus 2011 wurde der Zensus 2022 registergestützt durchgeführt (Dittrich und andere, 2022). Ein Personenbestand in Form eines zentralen Registers aller in Deutschland wohnhaften Personen existiert nicht und war für den Zensus 2022 zu erstellen. Datenquelle hierfür waren die Melderegisterdaten der etwa 11 000 Gemeinden Deutschlands, welche in rund 5 500 Melderegister führende Stellen organisiert sind.

Die Daten des Melderegisterdatenabzugs zum Zensusstichtag (15. Mai 2022; MRZ1) stellten die Grundlage des Personenbestands dar. Drei Monate nach dem Stichtag erfolgte ein weiterer Melderegisterdatenabzug (MRZ2), der die Datengrundlage erweiterte. Dies diente dazu, Änderungen in den Melderegistern nachzuziehen, die sich vor oder am Zensusstichtag ereignet hatten, aber noch nicht im Melderegisterdatenbestand am Zensusstichtag enthalten waren. Beispiele dafür sind Veränderungen des Personenstands, Geburten, Todesfälle oder die Meldung eines Umzugs. Die Daten des MRZ2 wurden anschriftenbasiert in den mit MRZ1 befüllten Personenbestand integriert. Hierfür wurde auf personenidentifizierende Merkmale zurückgegriffen, weil kein registerübergreifender und persistenter Personenidentifikator zur Verfügung stand. Bei erfolgreicher Anbindung an einen vorhandenen Datensatz wurde dieser mit den MRZ2-Informationen aktualisiert. Konnte der MRZ2-Personendatensatz nicht angebunden werden, wurde er neu aufgenommen.

Neben den Melderegistern wurden auch die Daten der beiden primärstatistischen Erhebungen des Zensus 2022 anschriftenbasiert integriert. Hierbei handelt es

sich einmal um die Sonderbereichserhebung.[1] Die Vollerhebung der Sonderbereichsanschriften dient dazu, die Melderegister um Unter- und Übererfassungen zu korrigieren (Boragk und andere, 2024). Eine Untererfassung bedeutet, dass eine Person an einer Anschrift wohnt, aber dort nicht gemeldet ist (Fehlbestand). Eine Übererfassung liegt vor, wenn eine Person an einer Anschrift gemeldet, dort aber nicht wohnhaft ist (Karteileiche).

Die zweite primärstatistische Erhebung ist die Haushaltsstichprobe. Auch diese Haushaltebefragung bei einer Stichprobe der sonstigen Wohnanschriften bereinigt zum einen die Melderegister um Unter- und Übererfassungen. Zum anderen wurden mit einer Unterstichprobe der ausgewählten Wohnanschriften sowie bei einem Teil der Wohnheime zusätzliche zensusrelevante Informationen erhoben, welche nicht in den Melderegistern enthalten waren (Klink/Lorentz, 2022).

Bei erfolgreicher Anbindung der primärstatistischen Personendaten an einen bestehenden Melderegisterdatensatz wurde dieser um die primärstatistischen Merkmale ergänzt. Hierbei konnte die Existenz der durch den Melderegisterdatensatz repräsentierten Person bestätigt werden, wenn die Person angetroffen wurde und die Anbindung erfolgreich war („gemeldet und wohnhaft“). Wurde die Person bei erfolgreicher Anbindung nicht angetroffen, konnte deren Existenz nicht bestätigt werden (Karteileiche). War kein Melderegisterdatensatz zu finden, um die erhobenen primärstatistischen Merkmale an der Anschrift anzubinden, wurde ein neuer Personendatensatz angelegt (Fehlbestand).

Eine weitere Herausforderung bei der Generierung des stichtagsrelevanten Personenbestands war, dass die Melderegister aufgrund der dezentralen Pflege melderechtswidrige Konstellationen von Meldungen derselben Person enthalten. Die Meldedaten werden zwar seit 2007 gemäß dem XMeld-Schema gehalten und mit einem Protokollstandard anlassbezogen ausgetauscht.[2] Allerdings können etwa fehlerhafte Eingaben, Fehlnutzungen von Feldern oder Datenverluste bei Datenbankmigrationen zu fehlerhaften Einträgen führen. Auch ist nicht auszuschließen, dass Karteileichen in MRZ1 und MRZ2 enthalten waren, welche auf die Zeit vor dem standardisierten Datenaustausch im Meldewesen zurückzuführen sind. Daher waren alle Melderegisterdatensätze zu einer Person zusammenzubringen und es war zu entscheiden, welcher von ihnen zur Ermittlung der Einwohnerzahl herangezogen werden sollte.

3 Arten von Mehrfachfällen

Durch die sukzessive Datenintegration entstanden Mehrfachfälle, beispielsweise weil sich Namensbestandteile zwischen den Datenlieferungen verändert haben. Der Personenbestand war um diese prozessproduzierten Mehrfachfälle zu bereinigen, bevor nach den Mehrfachfällen innerhalb der Melderegister gesucht werden konnte. Bei Mehrfachfällen an einer Anschrift (technische Mehrfachfälle) wurde die Integration der Datensätze zu derselben Person nachgeholt. Bei Mehrfachfällen an unterschiedlichen Anschriften, welche auf Umzüge im Zeitraum zwischen MRZ1 und MRZ2 zurückzuführen sind (temporäre Mehrfachfälle), wurde der jeweils veraltete Melderegisterdatensatz inaktiv gesetzt. ↘ Grafik 1 auf Seite 130

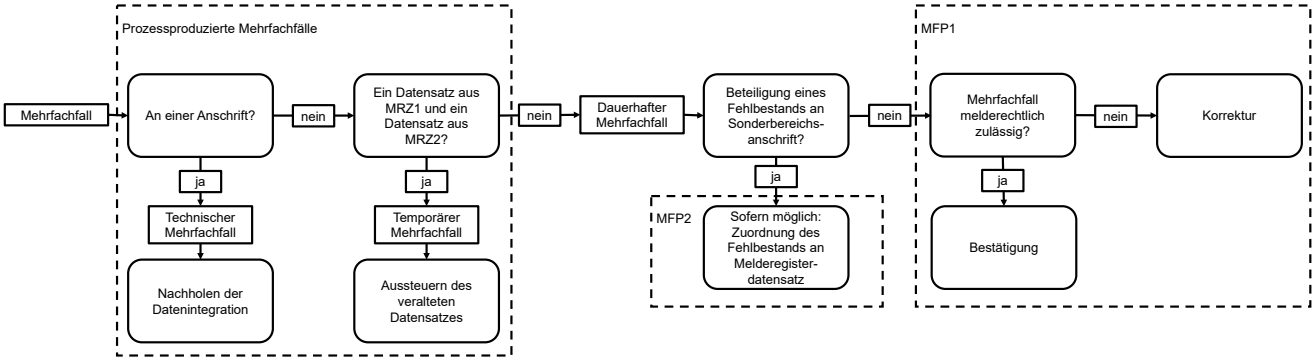
Allerdings sind Mehrfachfälle im Personenbestand nicht zwangsläufig auf Fehler innerhalb der Melderegister oder auf Fehler bei der Erstellung des Personenbestands zurückzuführen. In Abgrenzung zu den im Datenintegrationsprozess entstehenden Mehrfachfällen werden diese im Folgenden als dauerhafte Mehrfachfälle bezeichnet. Sie lassen sich noch weiter differenzieren:

- Mehrfache Meldungen des Wohnsitzes einer Person in den Melderegistern einer oder mehrerer Gemeinden sind gemäß § 21 Bundesmeldegesetz legitim, sofern zwischen Haupt- und Nebenwohnung unterschieden wird. Hier galt es, den Wohnungsstatus melderechtlich zulässiger Mehrfachfälle durch die Mehrfachfallprüfung zu bestätigen.
- Unzulässige dauerhafte Mehrfachfälle liegen bei mehreren einwohnerzahlrelevanten Melderegisterdatensätzen (alleinige Wohnung und Hauptwohnung) einer Person vor. Hier war genau ein Melderegisterdatensatz zu bestimmen, welcher die Person repräsentiert.

1 Sonderbereiche bezeichnen Anschriften, die sich aufgrund des Meldeverhaltens ihrer Bewohnenden von gewöhnlichen Anschriften unterscheiden, etwa aufgrund einer besonders hohen Fluktuation (zum Beispiel Studierendenunterkünfte) oder abweichenden Meldeverpflichtungen (zum Beispiel Kasernen).

2 Zur Einführung des standardisierten Datenaustauschs im Meldewesen siehe Tramer (2007) und für nähere Erläuterungen zu XMeld siehe KoSIT (2021a; 2021b).

Grafik 1
Arten der Mehrfachfälle im Zensus 2022 – Entstehung und Handlungsbedarfe



MRZ: Melderegisterauszug; MFP: Mehrfachfallprüfung

Die weiteren einwohnerzahlrelevanten Melderegisterdatensätze dieser Person wurden genau wie eine Nebenwohnung ohne Anbindung an einen einwohnerzahlrelevanten Datensatz als nicht existent ausgezeichnet. Diese Suche und Verarbeitung entsprechender dauerhafter Mehrfachfälle heißt nachfolgend Mehrfachfallprüfung 1 (MFP1).

› Für bestimmte Sonderbereichsanschriften besteht nach dem Bundesmeldegesetz eine Ausnahme von der Meldepflicht. Hier wurde geprüft, ob die Fehlbestände der Sonderbereichserhebung tatsächlich Untererfassungen der Melderegister aufdeckten oder ob die durch den Fehlbestand der Sonderbereichserhebung repräsentierte Person auch in den Melderegistern enthalten war. In einem an die Mehrfachfallprüfung anschließenden Prozess wurde dann entschieden, wo diese Person gezählt wurde (Bretschi und andere, 2024). Diese Suche wird nachfolgend MFP2 genannt.

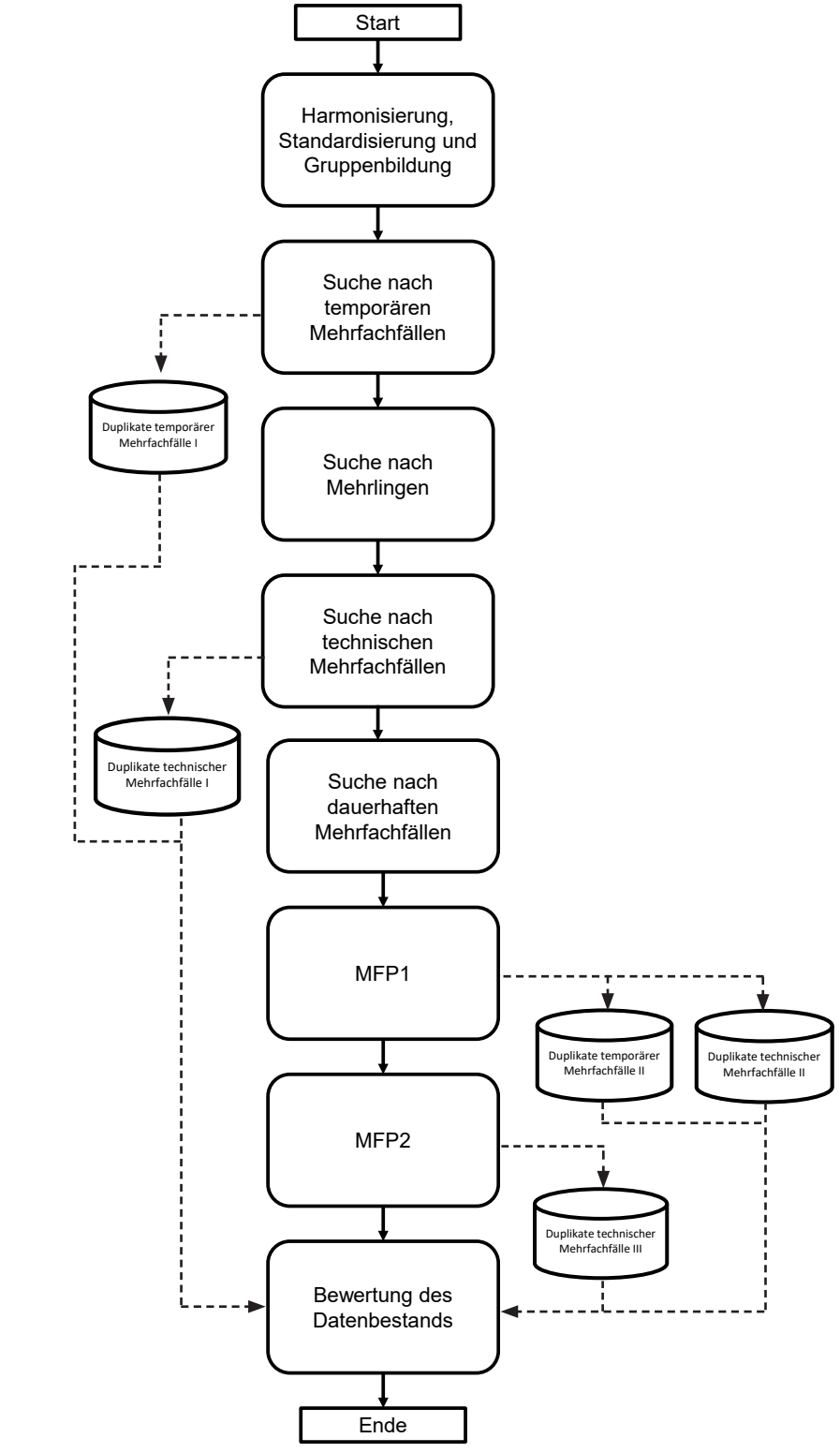
4

Ablauf der Mehrfachfallprüfung

Der Suche nach Mehrfachfällen lagen die Annahmen zugrunde, dass sich bestimmte Merkmale einer Person im Lebensverlauf nicht oder nur selten ändern und dass diese Änderungen mit den Informationen aus den Melderegistern oder den primärstatistischen Erhebungen nachvollzogen werden können. Hierzu wurden die Merkmale Geschlecht, Geburtsdatum, Geburtsort und die vorhandenen Namensinformationen als „Quasi-Identifikatoren“ (Schnell, 2020, hier: Seite 151 f.) genutzt: Über gleiche oder zulässig ähnliche Merkmalsausprägungen sollten Datensätze identifiziert werden, welche mit an Sicherheit grenzender Wahrscheinlichkeit dieselbe Person im Personenbestand repräsentierten.

Die Mehrfachfallprüfung folgte dem in der Fachliteratur üblichen Vorgehen aus Harmonisierung und Standardisierung, Blocken, Vergleichen und Entscheidungsfindung (beispielsweise Christen, 2012, hier: Seite 23 f.; Schnell, 2016, hier: Seite 665 f.). Die manuelle Bewertung strittiger Mehrfachfälle im Zensus 2011 wurde gegen ein vollständig maschinelles Verfahren ausgetauscht. Hierbei ersetzte die weiter unten beschriebene Analyseeinheit „Kette“ die bis dato zentrale Analyseeinheit „Dublette“ für das Verarbeiten und Bewerten der Mehrfachfälle. ↘ Grafik 2

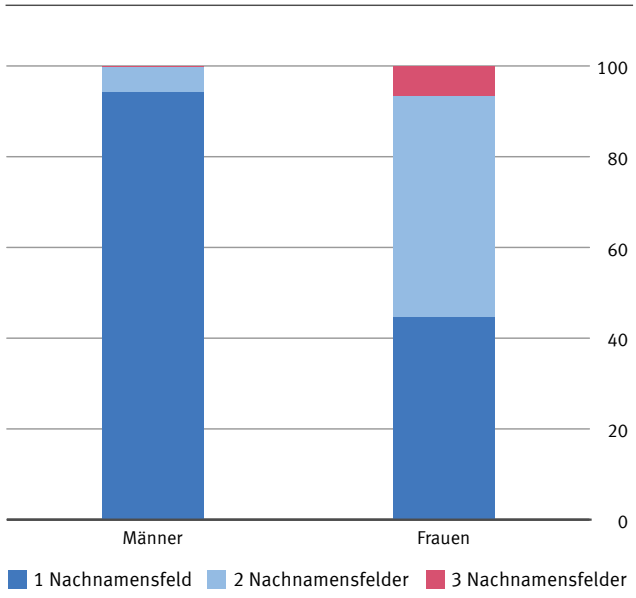
Grafik 2
Schematische Darstellung der Mehrfachfallprüfung des Zensus 2022



4.1 Harmonisierung und Standardisierung

Da Nachnamen sich insbesondere mit der Eheschließung ändern können, wurden neben dem Vornamen und dem aktuellen Familiennamen sofern vorhanden auch der Geburtsname sowie der Familienname vor erster Änderung auf Ähnlichkeit abgeglichen. Dass in Deutschland deutliche Unterschiede zwischen Männern und Frauen in Hinblick auf Nachnamenswechsel im Lebensverlauf existieren, zeigt ↘ Grafik 3.

Grafik 3
Befüllte Nachnamensfelder vor Standardisierung
in %



Die Angaben zum Geschlecht und zum Geburtsdatum wurden so übernommen, wie sie in den jeweiligen Ausgangsfeldern enthalten waren.[3] Für die fehleranfälligen Freitextfelder zu Namen und Geburtsort wurden folgende Standardisierungsschritte vorgenommen:

1. Alle Buchstaben kleinschreiben
2. Diakritika (Akzentzeichen), Umlaute sowie ß in ihre Stammbuchstaben überführen
3. Bindestriche durch Leerzeichen ersetzen
4. Sonderzeichen, Hochkommata und mehrfache sowie führende Leerzeichen entfernen
5. Klammern und deren Inhalte entfernen

Da gleiche Ausprägungen innerhalb der drei Nachnamensvarianten keinen informativen Mehrwert bringen, wurde der standardisierte Geburtsname nur behalten, sofern er vom aktuellen Familiennamen abwich. Der Familienname vor erster Änderung wurde nur dann behalten, wenn er wiederum weder dem standardisierten aktuellen Familiennamen noch dem standardisierten Geburtsnamen glich. Die Standardisierung des Geburtsortes wurde noch ausgeweitet: Lag im Ausgangsfeld eine Befüllung vor, welche auf Unkenntnis des Geburtsortsnamens hindeutete, beispielsweise „nicht bekannt“, dann blieb das standardisierte Geburtsortsfeld leer. Die sich aus der Standardisierung ergebenden Felder werden im Folgenden mit dem tiefgestellten Zusatz „standardisiert“ gekennzeichnet.

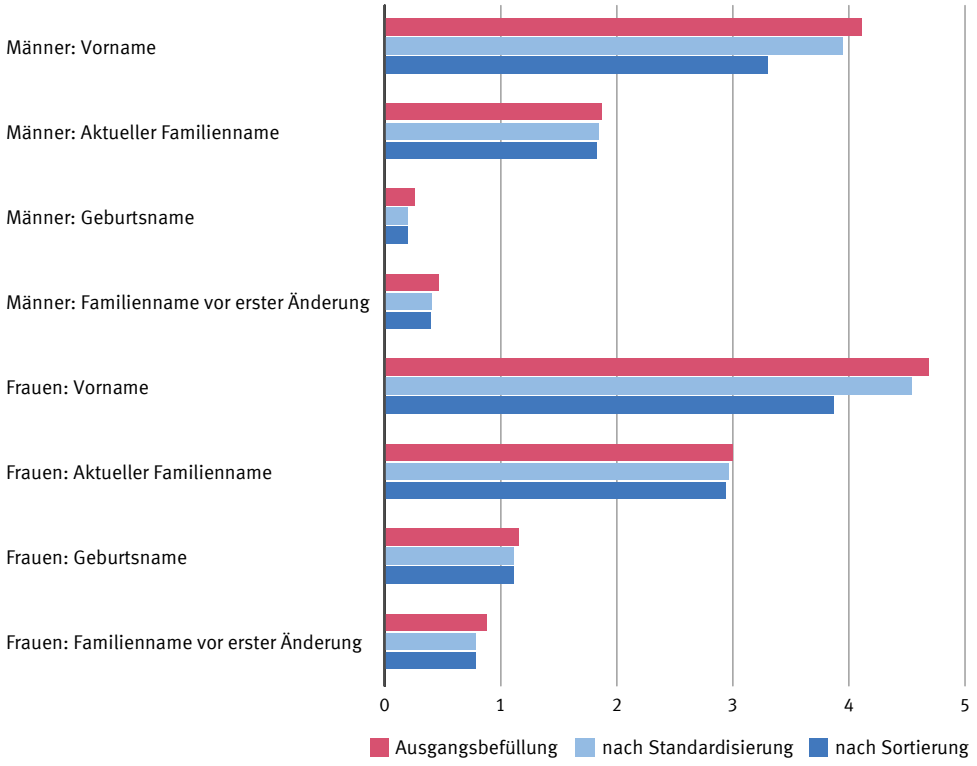
Zudem wurde für jedes der vier Namensfelder jeweils ein weiteres Feld angelegt, in welchen die in den jeweiligen standardisierten Namensfeldern enthaltenen Namen alphabetisch sortiert wurden.[4] Dies reduziert das Risiko, dass Personen mit Doppelnamen aufgrund unterschiedlicher Reihenfolgen der Namenseinträge nicht gefunden werden. Wie bei den standardisierten Nachnamensfeldern wurden die sortierten Geburts- und Familiennamensfelder vor erster Änderung nur bei den zuvor beschriebenen Abweichungen behalten. Die sich aus der Sortierung ergebenden Felder werden im Folgenden mit dem tiefgestellten Zusatz „sortiert“ gekennzeichnet.

Zuletzt wurde das standardisierte Vornamensfeld jeweils mit den drei Nachnamensvarianten kombiniert und alphabetisch sortiert, sodass drei weitere, als Vollnamen bezeichnete Felder entstanden. Auch hier wurden die Vollnamensfelder nur dann behalten, wenn sie sich voneinander unterschieden. Dies kompensiert unter anderem Unterschiede in der Übertragung von einzelnen Namen in die jeweiligen Felder für Personen aus Kulturkreisen, in welchen nicht in Vor- und Familiennamen aufgeteilt wird, oder einfache Vertauschungen in den Namenseingabefeldern. Die Wirkung der beschriebenen Aufbereitungsschritte stellt ↘ Grafik 4 dar.

3 Die Angaben zum Geschlecht enthalten die Ausprägungen „männlich“, „weiblich“, „divers“ und „ohne Angabe“. Aufgrund der geringen Besetzungen von „divers“ und „ohne Angabe“ werden die nachfolgenden Grafiken nur nach Männern und Frauen aufgeschlüsselt.

4 Die Standardisierung und Sortierung wurde auch für die Vornamen der ersten zehn Kinder und die Vor- und aktuellen Familiennamen der Ehe-/Lebenspartner sowie der gesetzlichen Vertreter angewandt.

Grafik 4
Unterschiedliche Ausprägungen der Namensfelder
Mill.



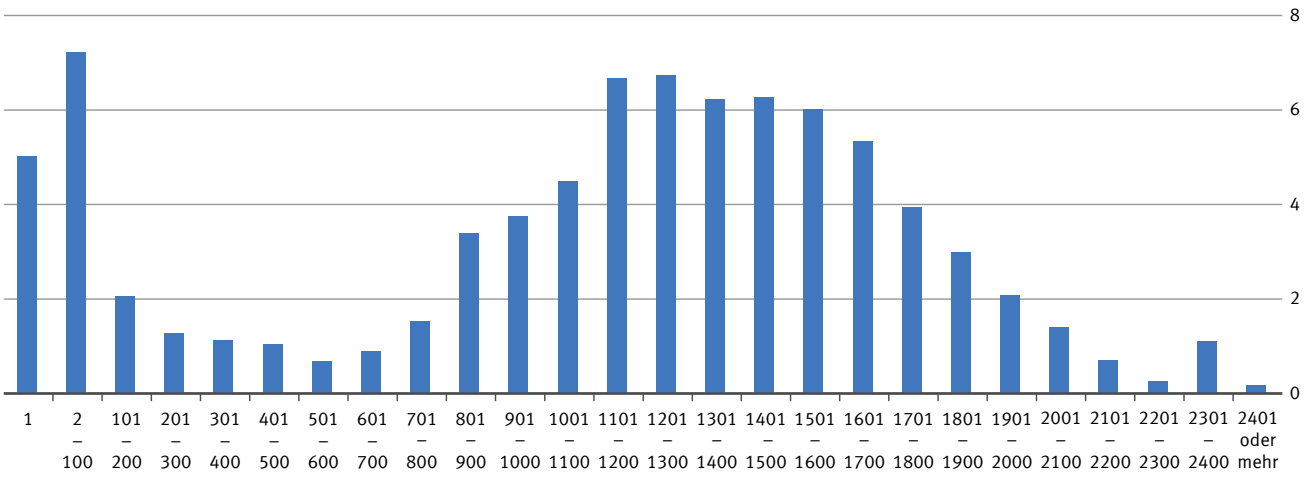
Auch die Felder zur aktuellen Anschrift und zur Zuzugsanschrift, welche nur für Melderegisterdatensätze zur Verfügung stand, wurden vergleichbar zu den Namensfeldern standardisiert. Die unterschiedlichen Einträge im Straßenfeld der aktuellen und der Zuzugsanschrift wurden jeweils auf Basis des im Zensus 2011 entwickelten Straßenthesaurus (Schöneich, 2012) standardisiert. Des Weiteren wurde für alle Melderegisterdatensätze ein Datumsfeld regelbasiert aus den verschiedenen Meldungsdaten zur Wohnung generiert. Dieses diente in der Mehrfachfallprüfung als Proxy für den tatsächlichen Wohnbeginn an einer Anschrift mit dem aktuellen Wohnungsstatus (Universaldatum).

4.2 Gruppenbildung

Die Felder zum Geschlecht und zum Geburtsdatum grenzten distinkte Suchräume für die paarweisen Abgleiche ab (Blocking). Datensätze, welche in diesen beiden Feldern identische Ausprägungen aufwiesen, gehörten zu derselben Obergruppe, innerhalb der die Namensabgleiche stattfanden. Insgesamt gab es 81 566 Obergruppen, sie enthielten im Durchschnitt 1 094 Datensätze. ↘ Grafik 5

Um die Anzahl der paarweisen Abgleiche innerhalb der Obergruppen noch weiter zu reduzieren, wurde der $\text{Geburtsort}_{\text{standardisiert}}$ als weitere, gleitende Einschränkung des Suchraums herangezogen. Innerhalb jeder Obergruppe musste sich der erste Buchstabe des $\text{Geburtsort}_{\text{standardisiert}}$ gleichen oder mindestens einer der abzugleichenden Datensätze wies einen leeren $\text{Geburtsort}_{\text{standardisiert}}$ auf. Dies reduzierte die durchgeführten Abgleiche um 93 %: Während ohne Berücksichtigung des ersten Buchstabens des $\text{Geburtsort}_{\text{standardisiert}}$ rund 65,3 Milliarden paarweise Abglei-

Grafik 5
Obergruppen nach gruppierter Anzahl enthaltener Datensätze
in 1 000



che durchzuführen gewesen wären, waren es so lediglich rund 4,61 Milliarden.

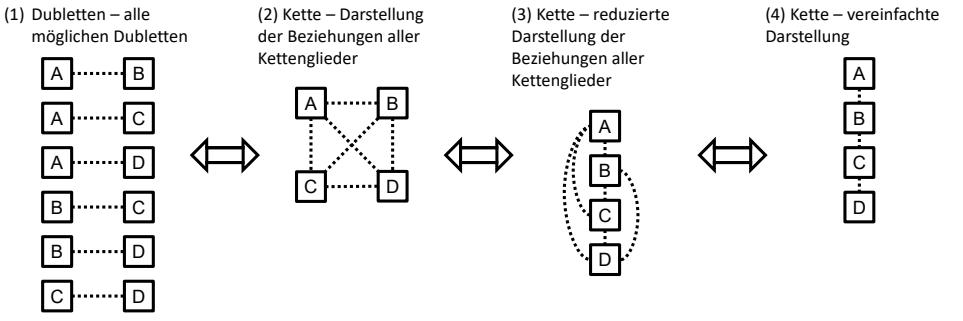
4.3 Vergleichen

Um die Ähnlichkeit der Namensinformationen zu bestimmen, wurde wie im Zensus 2011 auf die Jaro-Winkler-Stringmetrik (JW) zurückgegriffen. Hierbei handelt es sich um eine Metrik, bei welcher die Ähnlichkeit zweier Strings (a und b) auf der Basis der Reihenfolge identischer Zeichen innerhalb eines von der Länge der zu vergleichenden Strings abhängigen zulässigen Suchradius durch einen Wert im Wertebereich null bis eins ausgedrückt wird (Winkler, 1990). Für die Anwendung in der Mehrfachfallprüfung wurden die von Winkler (1990) vorgeschlagenen Grundeinstellungen genutzt. Allerdings wurden auch namentrennende Leerzeichen innerhalb des zulässigen Positionsradius als gemeinsame Zeichen gewertet.

4.4 Dubletten und Ketten

Wenn die miteinander verglichenen Datensätze (Elemente) einen bestimmten Schwellenwert erreichten oder überschritten ($JW(a,b) \geq \alpha$), wurden sie als initial paarig betrachtet und als sogenannte Dubletten weiterverarbeitet. Dubletten dienen als Zwischenergebnis der Identifizierung von Mehrfachfällen und als Voraussetzung für die Verkettung sowie für die anschließende detaillierte Prüfung der als potenzielle Mehrfachfälle

Grafik 6
Zusammenhang der Konzepte Dublette und Kette



derselben Person ausgemachten Datensätze. Ketten entstehen aus einer oder mehreren Dubletten mit mindestens einem gemeinsamen Element. Aus den beiden Dubletten $D(A,B)$ und $D(B,C)$ kann die dreigliedrige Kette $K(A,B,C)$ abgeleitet werden. In sprachlicher Abgrenzung zur Dublette werden die distinkten Elemente einer Kette als Kettenglieder bezeichnet. Ketten können auch in die maximal mögliche Anzahl sie konstituierender Dubletten zurückgewandelt werden. ↘ Grafik 6

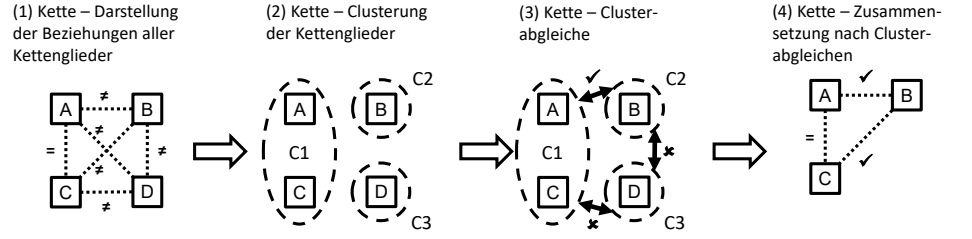
Ein wesentlicher Vorteil der Kette im Vergleich zur Dublette besteht für die Datenverarbeitung darin, dass alle derselben Kette angehörigen Datensätze in einem Long-Format abgelegt und mit einer gemeinsamen Identifikationsnummer für die jeweilige Kette versehen werden können.

4.5 Maschinelle Klärung strittiger Fälle: Kettentransitivität und Repräsentanz

Um sicherzustellen, dass alle Kettenglieder einer Kette dieselbe Person repräsentieren, muss jede Kette vollständig transitiv sein. Aufgrund der Verwendung des Schwellenwerts α und dem Vorgehen bezüglich des Geburtsorts konnten zwei Fälle vorkommen: Datensätze (A,C) wurden aufgrund der initialen Paarigkeitsbestimmung mit einem weiteren Datensatz (B) zu Kettengliedern derselben Kette verknüpft, welche im direkten Abgleich nicht paarig bewertet wurden ($\alpha \leq JW(a_A,b_B) < 1, \alpha \leq JW(a_B,b_C) < 1$ und $JW(a_A,b_C) < \alpha$). Oder es wurden mitunter auch Datensätze zu Dubletten, deren $\text{Geburtsorte}_{\text{standardisiert}}$ sich ab dem zweiten Buchstaben unterschieden. Derart „strittige“ Ketten wurden näher untersucht. Hierzu wurden die Kettenglieder strittiger Ketten auf Basis der Ausprägungen der standardisierten Namensfelder sowie dem $\text{Geburtsort}_{\text{standardisiert}}$ in distinkte Cluster unterteilt. Im Anschluss wurden die Kettenglieder verschiedener Cluster derselben Kette mithilfe weiterer Informationen aus den Melderegisterfeldern auf Gemeinsamkeiten abgeglichen (Clusterabgleich). Der Vorteil dieser nachgelagerten finalen Paarigkeitsprüfung bestand darin, dass die Anzahl der aufwendigen Abgleiche auf die initial verketteten Datensätze, welche Namens- beziehungsweise Geburtsortsdiskrepanzen aufwiesen, begrenzt wurde. Lagen keine Informationen vor, welche die Diskrepanz zwischen den Clustern innerhalb einer Kette überbrückten, so wurde die strittige Kette entlang der untereinander unpaarigen Cluster aufgetrennt. ↘ Grafik 7

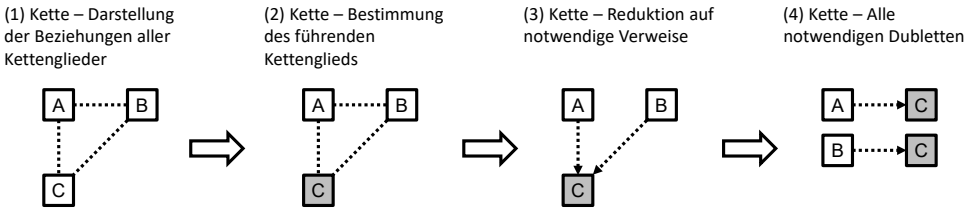
Jede Person darf nur durch einen einwohnerzahlrelevanten Datensatz im Personenbestand repräsentiert werden. Um dies zu erreichen, wurde für jede Kette ein führendes Kettenglied bestimmt. Alle weiteren Kettenglieder wurden als Duplikate bezeichnet und Datensätze, die nicht mit mindestens einem weiteren Datensatz verkettet werden konnten, als Unikate. Das führende Kettenglied wurde auf Basis fester Regeln bestimmt, welche sich darin unterschieden, ob es sich um prozessproduzierte Mehrfachfälle, Mehrfachfälle der Melderegister oder Mehrfachfälle mit Fehlbeständen der Sonderbereichserhebung handelte. Das führende Kettenglied ermöglichte, jede Kette in die Anzahl der zur Kettenkonstitution notwendigen Dubletten zu überführen, indem jedes Duplikat einzig auf das führende Kettenglied der Kette verweist. Dies war wichtig, damit die an die Mehrfachfallprüfung anschließenden Prozesse, welche aufgrund ihrer IT-Infrastruktur teilweise keine zeilenübergreifenden Operationen zuließen, die Ergebnisse der Mehrfachfallprüfung verarbeiten konnten. ↘ Grafik 8

Grafik 7
Schematische Darstellung zum Vorgehen des Clusterabgleichs zur Sicherstellung von Kettentransitivität



Erläuterungen: = identische Ausprägung; ≠ unterschiedliche Ausprägung; ✓Gemeinsamkeiten vorhanden; × keine Gemeinsamkeiten vorhanden.

Grafik 8
Schematische Darstellung zum Vorgehen der Bestimmung des führenden Kettenglieds



Anmerkung: Der grau hinterlegte Datensatz wurde als führendes Kettenglied bestimmt.

5

Suche nach prozessproduzierten Mehrfachfällen

Zunächst wurde der Personenbestand um die prozessproduzierten Mehrfachfälle bereinigt. Für die Suche nach temporären Mehrfachfällen wurden die ausschließlich in MRZ1 gelieferten mit den ausschließlich in MRZ2 gelieferten Datensätzen innerhalb der Untergruppen abgeglichen. Diese Suche zielte darauf ab, triviale temporäre Mehrfachfälle zu finden. Daher wurde die Gleichheit der standardisierten Vornamen sowie mindestens einer der standardisierten Nachnamensvarianten zur Paarigkeitsbestimmung genutzt. Die nicht trivialen temporären Mehrfachfälle wurden bei der späteren Suche nach dauerhaften Mehrfachfällen gefunden. Im Anschluss wurden die Dubletten entlang übereinstimmender Elemente verknüpft. Sofern mehr als zwei Datensätze aus MRZ2 in der entstandenen Kette enthalten waren, wurde der gemäß Universaldatum jüngste Datensatz als führendes Kettenglied ausgewählt und alle anderen MRZ2-Datensätze aus der Kette ausgesteuert. Alle Duplikate wurden mit einem Verweis zum entsprechenden führenden Kettenglied versehen, in eine Tabelle geschrieben und aus den nachfolgenden Prozessen der Mehrfachfallprüfung ausgesteuert.

Für technische Mehrfachfälle sollte bei der Integration der als Duplikate markierten Datensätze in die führenden Kettenglieder verhindert werden, dass im Elternhaus zusammenlebende gleichgeschlechtliche Mehrlinge mit ähnlich klingenden Vornamen oder mit gleichen Vornamen in unterschiedlicher Reihenfolge versehentlich zusammengefasst wurden. Daher wurden zum einen unter allen an einer Anschrift befindlichen Datensätzen derselben Untergruppe nach Dubletten mit identischen standardisierten Nachnamensvarianten, aber unterschiedlichen standardisierten Vornamen gesucht, sofern sie zum Stichtag das 27. Lebensjahr noch nicht vollendet hatten.[5] Zum anderen wurden gleichgeschlechtliche Mehrlinge in den Melderegisterdatensätzen der Eltern identifiziert. Sofern diese Mehrlinge im Personenbestand gefunden werden konnten, wurden auch sie zu Mehrlingsketten verknüpft in einer Tabelle gesammelt.

Im Anschluss wurde an allen Anschriften mit mindestens zwei Datensätzen von alleinigen Wohnungen und Hauptwohnungen nach technischen Mehrfachfällen gesucht. Aufgrund des beschränkten Suchraums konnte auch eine Abweichung von einem Jahr im Geburtsdatum als weitere gleitende Einschränkung des Suchraums genutzt werden. Da in erster Linie triviale technische Mehrfachfälle gefunden werden sollten, wurde Übereinstimmung in mindestens einer Vollnamensvariante als Paarigkeitskriterium gewählt. Die paarigen Datensätze wurden verkettet und auf enthaltene Mehrlinge gemäß der zuvor erstellten Mehrlingsketten kontrolliert. Innerhalb der mehrlingsfreien Ketten wurden anschließend die führenden Kettenglieder bestimmt. An Anschriften der Sonderbereichserhebung sowie der Haushaltsstichprobe wurde dafür der Existenzbefund gemäß Erhebung und an den sonstigen Anschriften das jüngste Universaldatum als Kriterium herangezogen. Anschließend wurde die Integration der Duplikate in die führenden Kettenglieder nachgeholt. Die Duplikate wurden dann in eine Tabelle geschrieben und aus den nachfolgenden Prozessen der Mehrfachfallprüfung ausgesteuert.

5 Die Anzahl der im Elternhaus lebenden Kinder nimmt mit dem Alter ab: Während etwa drei Viertel der 25-Jährigen in Deutschland im Jahr 2022 das Elternhaus bereits verlassen hatten, waren es bei den 30-Jährigen bereits über 90 % (Statistisches Bundesamt, 2023). Um auch möglichen systematischen, kulturell bedingten Verzögerungen der Haushaltsgründung von Menschen mit Migrationshistorie gerecht zu werden, wurden auf das relativ zeitkonstante Durchschnittsalter von etwa 24 Jahren für das Verlassen des Elternhauses in Deutschland noch drei Jahre aufgeschlagen.

6

Suche nach dauerhaften Mehrfachfällen

Nach Aussonderung der trivialen Duplikate temporärer und technischer Mehrfachfälle wurde innerhalb der beschriebenen Untergruppen[6] für alle abzugleichenden Datensätze der maximale $JW(a,b)$ über die neun Vergleichskombinationen der drei Vollnamen von Datensatz 1 (ν_i) und Datensatz 2 (ν_j) berechnet (β), wobei die Vollnamen sowohl von links nach rechts $(JW(\overrightarrow{\nu_i},\overrightarrow{\nu_j}))$ als auch von rechts nach links $(JW(\overleftarrow{\nu_i},\overleftarrow{\nu_j}))$ abgeglichen wurden:

$$\beta = MAX\left(JW(\overrightarrow{\nu_i},\overrightarrow{\nu_j}),JW(\overleftarrow{\nu_i},\overleftarrow{\nu_j})\right)$$

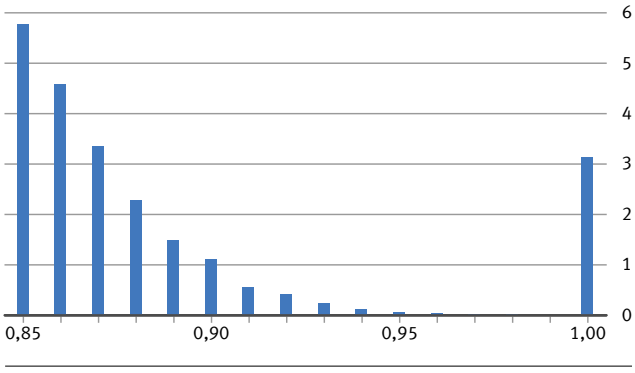
wobei

ν_i = Vollname i von Datensatz 1 mit i =1,2,3

ν_j = Vollname j von Datensatz 2 mit j =1,2,3.

Als Grenzwert für die initiale Bewertung der Paarigkeit der verglichenen Datensätze wurde 0,85 gewählt. Dieser Grenzwert ergab sich aus der Abwägung, einerseits möglichst wenige Datensätze als fälschlicherweise nicht paarig zu klassifizieren. Hierauf zielten auch die Abgleiche in und gegen Leserichtung ab, um nicht erfasste Zweitnamen zu kompensieren, welche durch die Sortierung der Namen innerhalb der Vollnamen einen Einfluss auf den JW-Wert haben konnten.[7] Auch die Wahl des maximalen JW-Werts aller neun Vollnamensabgleiche beider Vergleichsrichtungen sollte die Gefahr minimieren, vorzeitig Datensätze als unpaarig zu klassifizieren, welche sich mithilfe weiterer Informationen noch als Mehrfachfall derselben Person entpuppen konnten. Andererseits führte jede Absenkung des Schwellenwerts zu einem starken Anstieg der entstehenden Dubletten und damit zu einem Anstieg der notwendigen Rechenoperationen der folgenden Schritte. Als Zwischenergebnis der Suche nach dauerhaften Mehrfachfällen ergab sich das in ↘ Grafik 9 dargestellte maximale Dublettenset mit 23,216 Millionen Dubletten.

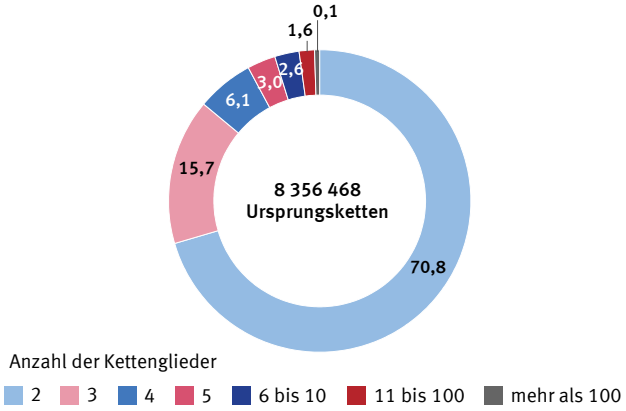
Grafik 9
Dauerhafte Dubletten nach maximaler Jaro-Winkler-Distanz über alle Abgleichskombinationen der Vollnamen
Mill.



Die Werte für die kaum sichtbaren Dubletten zwischen den Jaro-Winkler-Werten zwischen 0,90 und 1,00 sind: 562 000 (0,91); 420 000 (0,92); 233 000 (0,93); 116 000 (0,94); 59 000 (0,95); 30 000 (0,96); 19 000 (0,97); 22 000 (0,98); 6 000 (0,99)

Aus der Verkettung der dauerhaften Dubletten entstanden 8,357 Millionen sogenannte Ursprungsketten mit einer durchschnittlichen Länge von drei Kettengliedern, welche die Basis der nachfolgenden Schritte bildeten. ↘ Grafik 10

Grafik 10
Ursprungsketten nach Anzahl der Kettenglieder
in %



6 Für Fehlbestände der Haushaltsstichprobe bestand noch die Zusatzeinschränkung, dass sie nur innerhalb der Anschrift mit Melderegisterdaten abgeglichen wurden. Hintergrund dafür war, tatsächliche Aufdeckungen von Untererfassungen innerhalb der Melderegister von prozessproduzierten Fehlbeständen der Haushaltsstichprobe abzugrenzen, welche das Ergebnis einer nicht erfolgten Zusammenführung mit dem Melderegisterpendant der Person waren.

7 Zum Beispiel bei ν_i „judith-weiss“ und ν_j „anita-judith-weiss“ ergibt $JW(\overrightarrow{\nu_i},\overrightarrow{\nu_j}) = 0{,}74$ und $JW(\overleftarrow{\nu_i},\overleftarrow{\nu_j}) = 0{,}93$.

6.1 Suche nach Mehrfachfällen in den Melderegistern (MFP1)

Aus den Ursprungsketten wurden die miteinander verketteten Melderegisterdatensätze herausgezogen und als initiale MFP1-Ketten abgelegt. Innerhalb jeder als „strittig" markierten MFP1-Kette wurde jedes Kettenglied eines Clusters mit allen Kettengliedern der anderen Cluster paarweise abgeglichen. Sofern eine der folgenden Bedingungen für mindestens einen der entstehenden paarweisen Abgleiche zutraf, galt die Diskrepanz zwischen den Clustern als überbrückt. Alle in diesen Clustern enthaltenen Datensätze galten dann als dieselbe Person repräsentierend.

1) Dieselbe aktuelle oder letzte Lebens- beziehungsweise Ehepartnerschaft konnte auf Basis der Lebenspartnerschafts- beziehungsweise Eheschließungsdaten oder der -auflösungsdaten in Kombination mit vier von fünf hinterlegten Angaben zum oder zur aktuellen oder letzten Lebens- oder Ehepartner/-partnerin ($\text{Vorname}_{\text{sortiert}}$, $\text{Familienname}_{\text{sortiert}}$, $\text{Geburtsname}_{\text{sortiert}}$, Geschlecht, Geburtsdatum) ausgemacht werden.
2) Mindestens eines der hinterlegten Kinder stimmte in Bezug auf zwei von drei Merkmalen überein ($\text{Vorname}_{\text{sortiert}}$, Geschlecht, Geburtsdatum).
3) Mindestens einer der hinterlegten gesetzlichen Vertreter stimmte in Bezug auf drei von vier Merkmalen überein ($\text{Vorname}_{\text{sortiert}}$, $\text{Nachname}_{\text{sortiert}}$, Geschlecht, Geburtsdatum).
4) Mindestens eine aktuelle Anschrift stimmte überein.
5) Mindestens eine aktuelle Anschrift stimmte mit mindestens einer Zuzugsanschrift überein.
6) Mindestens eine Zuzugsanschrift stimmte überein.

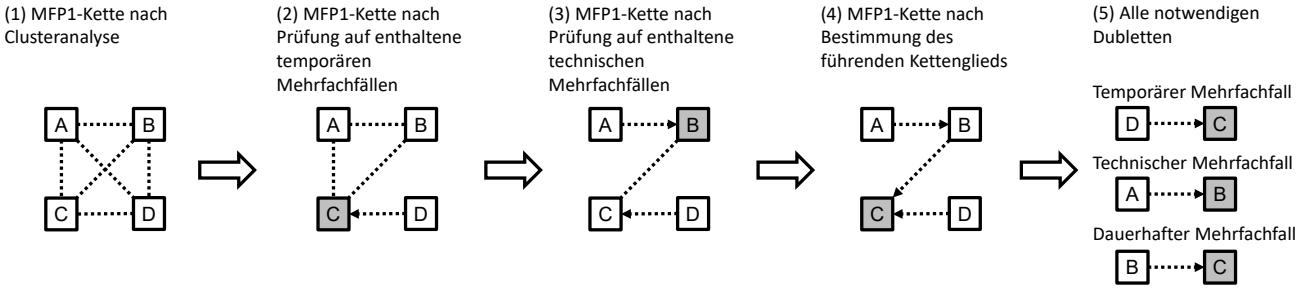
Die Bedingungen 3) bis 6) wurden allerdings nur als die Diskrepanz überbrückend gewertet, wenn einer der folgenden detaillierten Namensabgleiche zutraf:

a) Der $\text{Vorname}_{\text{sortiert}}$ und mindestens einer der drei sortierten Nachnamensvarianten (Aktueller $\text{Familienname}_{\text{sortiert}}$, $\text{Geburtsname}_{\text{sortiert}}$ und $\text{Familienname vor erster Änderung}_{\text{sortiert}}$) stimmten überein.
b) Mindestens einer der Vollnamen stimmte überein.
c) Bei gleichem $\text{Vornamen}_{\text{sortiert}}$ betrug der maximale JW-Wert mindestens eines Abgleichs der sortierten Nachnamensvarianten mindestens 0,9.
d) Bei einer unterschiedlichen Anzahl von Namen im $\text{Vornamen}_{\text{sortiert}}$ stimmte mindestens einer der drei sortierten Nachnamensvarianten überein und alle in dem weniger Worte beinhaltenden $\text{Vornamen}_{\text{sortiert}}$ enthaltenen Namen wurde mit einem JW-Wert von mindestens 0,95 in dem $\text{Vornamen}_{\text{sortiert}}$ des Vergleichsdatensatzes gefunden.
e) Bei mindestens einer gleichen sortierten Nachnamensvariante betrug der JW-Wert $\text{Vorname}_{\text{sortiert}}$ mindestens 0,9.
f) Bei einer unterschiedlichen Anzahl von Namen in den sortierten Nachnamensvarianten stimmte der $\text{Vorname}_{\text{sortiert}}$ überein und alle in der weniger Worte beinhaltenden sortierten Nachnamensvariante enthaltenen Namen wurden mit einem JW-Wert von mindestens 0,95 in einer der sortierten Nachnamensvarianten des Vergleichsdatensatzes gefunden.
g) Bei einer unterschiedlichen Anzahl von Namen in den Vollnamensfeldern wurden alle in dem weniger Worte beinhaltenden Vollnamensfeld enthaltenen Namen mit einem JW-Wert von mindestens 0,95 in einem Vollnamensfeld des Vergleichsdatensatzes gefunden.

Wenn keine Diskrepanz überbrückenden Übereinstimmungen zwischen Clustern gefunden wurden, dann wurden die Ketten entlang der nicht übereinstimmenden Cluster aufgetrennt. So entstanden aus der initialen MFP1-Kette teilweise neue Ketten und teilweise Unikate. Alle nach den Clusterabgleichen bestehenden MFP1-Ketten wurden auf verkettete Mehrlinge geprüft und sofern notwendig entlang dieser Mehrlinge aufgetrennt. Innerhalb der bestehenden Ketten wurden erneut prozessproduzierte Mehrfachfälle gesucht und die zuvor beschriebenen Schritte zum Umgang mit prozessproduzierten Mehrfachfällen vollzogen. Als letztes wurde innerhalb jeder MFP1-Kette das einwohnerzahlrelevante Kettenglied mit dem jüngsten Universaldatum zum führenden Kettenglied bestimmt. Bei reinen Nebenwohnungsketten wurde ersatzweise die Nebenwohnung mit dem jüngsten Universaldatum ausgewählt. Alle Duplikate erhielten wieder den Verweis zum führenden Kettenglied. ↘ **Grafik 11** stellt das beschriebene Vorgehen schematisch dar.

Grafik 11

Schematische Darstellung der sukzessiven Kontrolle der MFP1-Ketten auf enthaltene temporäre und technische Mehrfachfälle und der Bestimmung des Führenden Kettenglieds der MFP1-Kette



Anmerkung: Der grau hinterlegte Datensatz wurde in dem jeweiligen Abschnitt als führendes Kettenglied bestimmt.
MFP1: Mehrfachfälle in den Melderegistern.

6.2 Suche nach Mehrfachfällen mit Fehlbeständen der Sonderbereichserhebung (MFP2)

Für die MFP2 wurden alle nach MFP1 nicht als temporäre oder technische Duplikate ausgesteuerten Kettenglieder der Ursprungsketten, welche mindestens einen Fehlbestand der Sonderbereichserhebung enthielten, als initiale MFP2-Ketten abgelegt. Innerhalb dieser MFP2-Ketten wurden die Fehlbestände der Sonderbereichserhebung mit allen als führende Kettenglieder hervorgegangenen Datensätzen sowie den Unikaten[8] paarweise abgeglichen. Hierbei wurde ein Fehlbestand der Sonderbereichserhebung nur dann einem Melderegisterdatensatz zugeordnet, um eine finale MFP2-Kette zu bilden, wenn die in MFP1 angewandten Bedingungen der Namensinformationsabgleiche zutrafen. Sofern aus diesen Namensinformationsabgleichen mehrere Melderegisterdatensätze hervorgingen, welchen derselbe Fehlbestand der Sonderbereichserhebung zugeordnet werden konnte, wurden zusätzlich die Informationen zum Familienstand, zur Staatsangehörigkeit und zum Geburtsstaat zur finalen Zuordnung herangezogen. Innerhalb der finalen MFP2-Ketten wurde nochmals nach technischen Mehrfachfällen gesucht und, sofern vorhanden, das Prozedere zum Umgang mit technischen Mehrfachfällen wiederholt. Als führendes Kettenglied wurde der jeweils enthaltene Fehlbestand der Sonderbereichserhebung ausgewählt und die Duplikate der finalen MFP2-Kette erhielten den Verweis zu diesem.

8 Dies umfasst sowohl die im Lauf der MFP1 aus der initialen MFP1-Kette herausgelösten Datensätze als auch jene, welche die einzigen Melderegisterdatensätze innerhalb der Ursprungskette ausmachten und daher nicht in die MFP1 eingingen.

7 Wirkung der Mehrfachfallprüfung auf den stichtagsrelevanten Personenbestand

Nach dem Durchlauf der MFP2 wurden die prozessproduzierten Mehrfachfälle konsolidiert: Hatte sich das führende Kettenglied eines temporären Mehrfachfalls im Zuge der sukzessiven Abfolge der Schritte der Mehrfachfallprüfung als Duplikat eines technischen Mehrfachfalls entpuppt, dann wurden die Verweise der zugehörigen Duplikate auf das entsprechende führende Kettenglied dieses technischen Mehrfachfalls umgewidmet. Hiernach wurde der Personenbestand konsolidiert, wobei alle Duplikate der prozessproduzierten Mehrfachfälle als nicht existent gemäß der Melderegisterkonsolidierung ausgezeichnet wurden ($Existenz_{MRKONS}$). Die übrigen Melderegisterdatensätze erhielten abhängig vom initialen Wohnungsstatus aus dem Melderegister ($Wohnungsstatus_{MR}$) sowie der Anzahl der in der finalen MFP1-Kette enthaltenen alleinigen Wohnungen, Hauptwohnungen und Nebenwohnungen eine Existenzsetzung nach MFP1 ($Existenz_{MFP1}$) und einen Wohnungsstatus nach MFP1 ($Wohnungsstatus_{MFP1}$). Die Bewertung der Datensätze gemäß Mehrfachfallprüfung zeigt ↘ Übersicht 1.

Übersicht 1

Vergabe der Existenz und des Wohnungsstatus gemäß MFP1

Art des Mehrfachfalls	$Wohnungsstatus_{MR}$	$Existenz_{MRKONS}$	Anzahl weiterer MFP1-Kettenglieder mit $Wohnungsstatus_{MR}$ AW/HW	Anzahl weiterer MFP1-Kettenglieder mit $Wohnungsstatus_{MR}$ NW	$Existenz_{MFP1}$	$Wohnungsstatus_{MFP1}$
Duplikat eines temporären Mehrfachfalls	AW/HW	n				
Duplikat eines technischen Mehrfachfalls	AW/HW	n				
Unikat nach MFP1	AW	e	0	0	e	AW
	HW	e	0	0	e	AW
	NW	e	0	0	n	
Duplikat einer finalen MFP1-Kette	AW/HW	e	≤ 1	≤ 0	n	
	NW	e	0	≤ 1	n	
	NW	e	≤ 1	≤ 0	e	NW
Führendes Kettenglied einer finalen MFP1-Kette	AW	e	≤ 1	0	e	AW
	HW	e	≤ 1	0	e	AW
	AW	e	≤ 0	≤ 1	e	HW
	HW	e	≤ 0	≤ 1	e	HW
	NW	e	0	≤ 1	n	

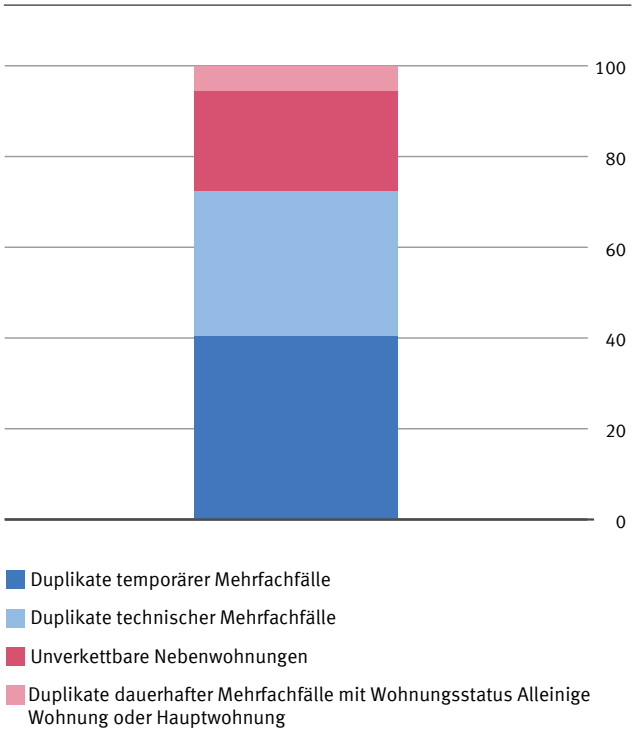
MFP1: Mehrfachfälle in den Melderegistern; MR: Melderegister; AW: Alleinige Wohnung; HW: Hauptwohnung; NW: Nebenwohnung; n: nicht existent; e: existent

Die Ergebnisse dieser (Nicht-)Existenzsetzung und Wohnungsstatusvergabe gemäß Mehrfachfallprüfung zeigen die Grafiken 12 und 13: Den Großteil der rund 692 700 als nicht existent markierten Datensätze machten die prozessproduzierten Mehrfachfälle an verschiedenen Anschriften mit rund 279 900 (40,4 %) sowie an einer Anschrift mit rund 222 500 (32,1 %) aus. Weitere 152 700 (22 %) Nicht-Existenzsetzungen entfielen auf Nebenwohnungen, welche keinem einwohnerzahlrelevanten Melderegisterdatensatz zugeordnet werden konnten. Die überzähligen einwohnerzahlrelevanten Melderegisterdatensätze beliefen sich auf rund 37 600 (5,4 %). Dass trotz der im Vergleich zum Zensus 2011 erweiterten Methode verhältnismäßig wenige melderegisterimmanente, melderechtlich unzulässige einwohnerzahlrelevante Mehrfachfälle gefunden wurden, ist mutmaßlich auf den automatisierten Datenaustausch der melderegisterführenden Stellen zurückzuführen. ↘ Grafik 12

Grafik 12

Grund für die Nicht-Existenzsetzung gemäß Mehrfachfallprüfung

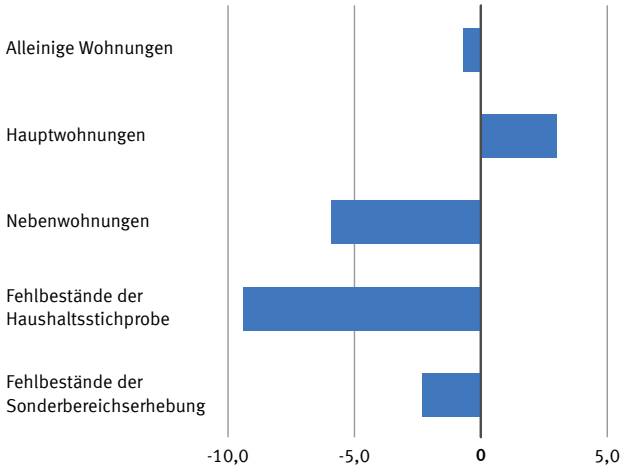
in %



Innerhalb des Personenbestands reduzierte die Mehrfachfallprüfung die Anzahl der alleinigen Wohnungen um 0,7 %. Allerdings war diese Abnahme nicht nur auf Nichtexistenzsetzungen zurückzuführen: In einigen Fällen konnten eine oder mehrere Nebenwohnungen einer originären alleinigen Wohnung zugeordnet werden, sodass der Wohnungsstatus auf Hauptwohnung umge-

ändert wurde. Hierdurch ist die Zunahme der Zahl der Hauptwohnungen nach Mehrfachfallprüfung um 3,0 % im Prä-Post-Vergleich zum initialen Personenbestand zu erklären. Die Abnahme der Zahl der Nebenwohnungen um 5,9 % ist vorwiegend auf die Nicht-Existenzsetzung jener Nebenwohnungen zurückzuführen, für welche keine einwohnerzahlrelevanten Datensätze gefunden wurden. Hierbei handelte es sich vermutlich um inaktive Nebenwohnsitze, wie ein Vergleich der durchschnittlichen Universaldaten zeigte: Während das Universaldatum der durch die Mehrfachfallprüfung in ihrer Existenz bestätigten Nebenwohnungen im Durchschnitt auf März 2011 datierte, betrug das durchschnittliche Universaldatum der nicht existent gesetzten Nebenwohnungen April 1992. Mithilfe der erweiterten Methode der Mehrfachfallprüfung konnten 9,4 % der initialen Fehlbestände der Haushaltsstichprobe nachträglich mit einer originären Karteileiche aus den Melderegistern an derselben Anschrift zusammengeführt werden. Dies gelang auch für 2,3 % der initialen Fehlbestände der Sonderbereichserhebung. ↘ Grafik 13

Grafik 13
Veränderung der Datensatzart im Personenbestand nach gegenüber vor der Mehrfachfallprüfung
in %



8

Fazit

Die Mehrfachfallprüfung ist eine wichtige qualitätssichernde Maßnahme im Personenbestand des Zensus 2022. Der Perspektivwechsel von der Dublette zur Kette als zentrale Analyseeinheit und der Einsatz der Clusteranalyse strittiger Ketten ermöglichte die Entwicklung eines vollständig maschinellen Verfahrens, um Mehrfachfälle zu suchen und zu bewerten. Im Ergebnis wurden vor allem prozessproduzierte Mehrfachfälle aufgedeckt.

LITERATURVERZEICHNIS

Bretschi, Corinna/Seibel, Steffen/Vorndran, Ingeborg/Meyn, Christoph. *Ermittlung der Einwohnerzahl im Zensus 2022*. In: WISTA Wirtschaft und Statistik. Ausgabe 6/2024, Seite 17 ff.

Christen, Peter. *Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Heidelberg 2012. DOI: 10.1007/978-3-642-31164-2

Diehl, Eva-Maria. *Methoden der Mehrfachfallprüfung im Zensus 2011*. In: Wirtschaft und Statistik. Ausgabe 6/2012, Seite 473 ff.

Dittrich, Stefan/Bretschi, Corinna/Stepien, Halina Danuta/Vorndran, Ingeborg/Michel, Bernd/Kleber, Birgit/Timm, Ulrike/Pfahl, Miriam. *Der Zensus 2022 – mit Online First an der Schwelle zu einem Registerzensus*. In: WISTA Wirtschaft und Statistik. Ausgabe 6/2022, Seite 90 ff.

Boragk, Lisa/Gaedke, Annika/Gemmeke, Charlotte/Meyn, Christoph. *Die Ermittlung des Berichtskreises am Beispiel der Sonderbereiche im Zensus 2022*. In: WISTA Wirtschaft und Statistik. Ausgabe 6/2024, Seite 41 ff.

Dreschmitt, Kai/Pfahl, Miriam. *Die Personenerhebung im Zensus 2022*. In: WISTA Wirtschaft und Statistik. Ausgabe 6/2024, Seite 29 ff.

Klink, Steffen/Lorentz, Kai. *Auswahlplan und Stichprobenhauptziehung für den Zensus 2022*. In: WISTA Wirtschaft und Statistik. Ausgabe 1/2022, Seite 13 ff.

KoSIT (Koordinierungsstelle für IT-Standards). *XMeld*. 2021a [Zugriff am 7. Januar 2025]. Verfügbar unter: www1.osci.de

KoSIT (Koordinierungsstelle für IT-Standards). *Datensatz für das Meldewesen. Einheitlicher Bundes-/Länderteil (DSMeld)*. Bearbeitungsstand: 28. Oktober 2021 (12. Änderung, wirksam ab 1. November 2021). 2021b. [Zugriff am 7. Januar 2025]. Verfügbar unter: www1.osci.de

Schnell, Rainer. *Record Linkage*. In: Wolf, Christof/Joye, Dominique/Smith, Tom W./Fu, Yang-Chih (Herausgeber). The SAGE Handbook of Survey Methodology. London 2016. Seite 662 ff.

Schnell, Rainer. *Record Linkage als zentraler Baustein der Forschung mit Registern und Big Data-Nutzungen*. In: Klumpe, Bettina/Schröder, Jette/Zwick, Markus (Herausgeber). Qualität bei zusammengeführten Daten. Befragungsdaten, administrative Daten, neue digitale Daten: miteinander besser? Wiesbaden 2020. Seite 151 ff. DOI: 10.1007/978-3-658-31009-7_11

Schöneich, Cordula. *Der Straßenthesaurus im Zensus 2011*. In: Wirtschaft und Statistik. Ausgabe 11/2012, Seite 957 ff.

LITERATURVERZEICHNIS

Statistisches Bundesamt. *Junge Menschen verlassen ihr Elternhaus im Schnitt im Alter von 23,8 Jahren*. Zahl der Woche Nr. 36 vom 5. September 2023. [Zugriff am 7. Januar 2025]. Verfügbar unter: www.destatis.de

Tramer, Karl. *Elektronischer Datenaustausch im Meldewesen. Modell für überörtlichen Datenaustausch in Deutschland?* In: VM Verwaltung & Management. Jahrgang 13. Heft 6/2007, Seite 300 ff. DOI: 10.5771/0947-9856-2007-6

Winkler, William E. *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*. 1990. [Zugriff am 7. Januar 2025]. Verfügbar unter: eric.ed.gov

RECHTSGRUNDLAGEN

Bundesmeldegesetz vom 3. Mai 2013 (BGBl. I Seite 1084), das zuletzt durch Artikel 6 des Gesetzes vom 23. Oktober 2024 (BGBl. I Nr. 323) geändert worden ist.

Gesetz zur Durchführung des Zensus im Jahr 2022 (Zensusgesetz 2022) vom 26. November 2019 (BGBl. I Seite 1851), das durch Artikel 2 des Gesetzes vom 3. Dezember 2020 (BGBl. I Seite 2675) geändert worden ist.

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im Februar 2025
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-25001-4, ISSN 1619-2907