Bretnütz, Hella et al.

Article

Aufbereitung und Plausibilisierung der primärstatistischen Erhebungsteile im Zensus 2022

WISTA - Wirtschaft und Statistik

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

Suggested Citation: Bretnütz, Hella et al. (2024) : Aufbereitung und Plausibilisierung der primärstatistischen Erhebungsteile im Zensus 2022, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 76, Iss. 6, pp. 51-60

This Version is available at:
https://hdl.handle.net/10419/309499

AUFBEREITUNG UND PLAUSIBILISIERUNG DER PRIMÄRSTATISTISCHEN ERHEBUNGSTEILE IM ZENSUS 2022

Hella Bretnütz, Sabrina Estatico, Sabrina Walther, Fabienne Hüsgen-Brodhäcker, Kerstin Lange, Birgit Kleber, Benjamin Haberzettl

↘ **Schlüsselwörter:** Datenaufbereitung – Plausibilitätskontrollen – Fehlerbereinigung – Imputation – Personenerhebung – Gebäude- und Wohnungszählung – Gemeinschaftsunterkünfte

ZUSAMMENFASSUNG

Fehlerhafte Datensätze können die Ergebnisse einer Statistik verzerren. Um dies zu vermeiden, fanden im Zensus 2022 verschiedene Aufbereitungs-, Plausibilisierungs- und Imputationsschritte statt. Für jeden Erhebungsteil des Zensus wurde dabei das beste Set an Plausibilisierungsmethoden genutzt, um eine valide Datenbasis für Auswertungen zu schaffen. Der Beitrag beschreibt die einzelnen Aufbereitungs- und Plausibilisierungsschritte für die unterschiedlichen Erhebungsteile des Zensus 2022 mit den jeweiligen Besonderheiten.

↘ ***Keywords:*** *data processing – plausibility checks – error elimination – imputation – survey of individuals – census of buildings and housing – collective living quarters*

ABSTRACT

Incorrect data sets may bias statistical results. To avoid this, various processing, editing and imputation steps were carried out in the 2022 Census. The best set of methods for plausibility checking was used for each survey component of the census in order to create a valid data basis for evaluations. This article describes the individual steps of processing and editing the different 2022 Census components and looks at special aspects of each component.

Hella Bretnütz

war beim Zensus 2022 bei der Gebäude- und Wohnungszählung mit der Datenaufbereitung und Imputation betraut.

Sabrina Estatico

war beim Zensus 2022 im Teilprojekt „Personenerhebung (Konzeption und Aufbereitung)“ unter anderem für die Plausibilisierung und Aufbereitung der Erhebungsdaten aus der Personenerhebung verantwortlich.

Sabrina Walther

war in der Personenerhebung des Zensus 2022 tätig und betreute hier unter anderem die Plausibilisierung und Imputation der Daten.

Fabienne Hüsgen-Brodhäcker

war für den Zensus 2022 schwerpunktmäßig für die fachlichen Vorgaben und die Begleitung der technischen Umsetzung der erhebungsteilübergreifenden Plausibilisierung zuständig.

Kerstin Lange

war beim Zensus 2022 mit der Methodik und Umsetzung der Imputation betraut.

Birgit Kleber

betreut im Statistischen Bundesamt für den Zensus die Themen Imputation und Geheimhaltung.

Benjamin Haberzettl

war im Zensus 2022 im Bereich „Melderegister und Sonderbereiche“ für die Umsetzung der Fehlerprüfungen von Melderegister-Dateneingängen und der IDEV-Erhebung an Gemeinschaftsunterkünften zuständig.

Alle Autorinnen und der Autor dieses Artikels sind im Statistischen Bundesamt tätig.

1

Einleitung

Innerhalb einzelner Erhebungsteile und auch zwischen den verschiedenen Erhebungsteilen kam es im Zensus 2022 zu unplausiblen (fehlerhaften) oder fehlenden Werten. Unplausible oder fehlende Werte stellen vor allem bei der Interpretation der Ergebnisse eine Herausforderung dar, denn sie können die Ergebnisse verzerren (Preising und andere, 2021). Geeignete Aufbereitungsschritte sollen diesem Umstand Rechnung tragen und Ergebnisverzerrungen vermeiden. Neben der Aufbereitung der Daten (Kapitel 2) finden Plausibilisierungen und Imputationen in den Datensätzen (Kapitel 3) statt. Je nach Erhebungsteil des Zensus kommen dabei zum Teil unterschiedliche Aufbereitungsschritte zum Einsatz.

Der Zensus 2022 wurde als registergestützte Erhebung durchgeführt. Alle verwendeten Quellen waren hierbei in sich selbst und auch untereinander zu plausibilisieren. Neben dem Referenzdatenbestand, der alle Registerinformationen (insbesondere die Melderegisterdaten) und Primärstatistiken zusammenfasst, schloss der Zensus 2022 drei Primärstatistiken ein:

- die Gebäude- und Wohnungszählung,
- die Haushaltebefragung auf Stichprobenbasis (auch Haushaltsstichprobe genannt) und die Befragung an Wohnheimen (zusammengefasst unter dem Begriff Personenerhebung) sowie
- die Befragung an Gemeinschaftsunterkünften (an Anschriften mit sogenannten Sonderbereichen).

Der Beitrag beschreibt die vorgenommenen Plausibilisierungsarbeiten und nimmt dabei Bezug auf die zeitliche Zuordnung:

- Während der Erhebungsphase:
  Zwischen den Erhebungsteilen fanden erhebungsteilübergreifende Plausibilisierungen (Abschnitt 2.1) statt sowie verschiedene Aufbereitungsschritte und Plausibilisierungen innerhalb der Primärerhebungen (Abschnitt 2.2).
- Nach Abschluss der Erhebungsphase:
  Die Primärerhebungen wurden final plausibilisiert und fehlende oder unplausible Werte mittels eines Nächste-Nachbarn-Ansatzes imputiert (Kapitel 3).

Ein kurzes Fazit beschließt den Artikel.

2

Aufbereitung während der Erhebungsphase

2.1 Referenzdatenbestand

Der Referenzdatenbestand diente als zentrale Datenbank im Zensus 2022 dazu, die Erhebungen zu steuern und zu organisieren. Er enthielt Daten auf Personenebene (Personenbestand) und auf Anschriftenebene (Anschriftenbestand, Steuerungsregister). Die Ergebnisse aus den verschiedenen Erhebungsteilen liefen im Referenzdatenbestand zusammen. Dadurch konnte dieser genutzt werden, um durch automatisierte Abgleiche zwischen den Daten aus den verschiedenen Erhebungsteilen die Konsistenz des Gesamtdatenbestands des Zensus 2022 zu verbessern. Die Abgleiche erfolgten bereits im Aufbereitungsprozess und werden im Folgenden unter dem Begriff „erhebungsteilübergreifende Plausibilisierung“ (EÜPL) zusammengefasst. Zu unterscheiden sind hierbei die sogenannte EÜPL Existenzen während der Erhebungsphase und die Sonder-EÜPL (Abschnitt 3.1) zur Qualitätsverbesserung nach der Erhebungsphase.

Bei der EÜPL Existenzen handelte es sich um einen wichtigen methodischen Bestandteil des Zensus 2022 im Hinblick auf die korrekte Ermittlung der Einwohnerzahl. Die EÜPL Existenzen war relevant für Anschriften, die im Zuge der Personenerhebung erfasst wurden. Sie sollte Unplausibilitäten zwischen den Angaben aus den Lieferungen der Melderegister (MR) und dem Ergebnis der primärstatistischen Erhebungen zur Existenzfeststellung auf Personenebene aufdecken sowie bei Bedarf bereinigen. Abweichungen zwischen Melderegister und einer Existenzfeststellung auf Personenebene an einer Anschrift sollten überprüft werden, um mögliche Fehler bei der Erfassung der Personen oder der Abgrenzung der Anschrift im Steuerungsregister auszuschließen. Den wesentlichen Input für die EÜPL Existenzen bildeten die gemeldeten Personen laut der aktuell vorliegenden MR-Lieferung. Zudem wurde das Ergebnis der Existenzfeststellung aus der Personenerhebung in Form von Strukturmerkmalen auf Anschriftenebene (Anzahl der vorläufigen Karteileichen, vorläufigen Fehlbestände

und vorläufigen paarig existenten Personen)[1] herangezogen. Die EÜPL Existenzen wurde an einer Anschrift im Referenzdatenbestand automatisch gestartet, sobald die Integration des Rücklaufs aus der Personenerhebung abgeschlossen war.

Um potenziell unplausible Anschriften zu ermitteln, wurden verschiedene Prüfkategorien definiert, in die die Strukturmerkmale und bei Bedarf zusätzliche Informationen zum Beispiel aus der Gebäude- und Wohnungszählung eingingen. Beispielsweise sollten mit der Prüfkategorie „Deutlich mehr existente Personen, als laut MR zu erwarten“ Anschriften aufgedeckt werden, bei denen in der Erhebung möglicherweise nicht nur die jeweilige MR-Anschrift, sondern auch ein Nachbargebäude fälschlicherweise mit erhoben wurde. Die Prüfkategorie „Keine paarigen Datensätze und nur Karteileichen an einer Anschrift“ sollte dagegen Anschriften aufdecken, bei denen gar keine Existenzen in der Personenerhebung festgestellt wurden, weil die Anschrift gar nicht erhoben wurde, sondern stattdessen der Anschriftenbefund „Baulücke“ oder „komplett gewerblich genutzte Anschrift“ (fälschlicherweise) vergeben wurde. Um festzulegen, ab wann eine Anschrift als unplausibel gilt, wurden bundeseinheitliche Schwellenwerte definiert. Dabei wurden die Prüfbedingungen und Schwellenwerte leicht variiert, je nachdem, ob es sich um eine Stichprobenanschrift oder eine Anschrift mit Sonderbereichen (Wohnheime, Gemeinschaftsunterkünfte) handelte. Wurde bei einer Prüfung der Schwellenwert über- beziehungsweise unterschritten, konnten Mitarbeitende der Statistischen Ämter der Länder manuelle Prüfungen vornehmen. Diese Prüfungen führten entweder zur Auflösung der Unplausibilität oder gegebenenfalls zu einer erneuten Begehung der Anschrift.

2.2 Gebäude- und Wohnungszählung

Die Gebäude- und Wohnungszählung wurde als Vollerhebung aller Gebäude mit Wohnraum und darin befindlicher Wohnungen sowie bewohnten Unterkünften durchgeführt. Zu jeder zählungsrelevanten Einheit (Gebäude und/oder Wohnung) wurden die betreffenden Auskunftspflichtigen (in der Regel Eigentümerin, Eigentümer oder Verwaltung) über Verwaltungsdatenquellen ermittelt und angeschrieben. Dabei wurden zunächst Zugangsdaten für den Online-Fragebogen versendet und kein Papierfragebogen. Insgesamt betrug die Online-Quote bei der Gebäude- und Wohnungszählung 82 %. Einer der Vorteile der Online-Datenerhebung ist die mögliche Ad-hoc-Plausibilisierung bereits während der Eingabe der Daten durch die Auskunftspflichtigen (Freier/Mosel, 2019).

Sowohl die online erhobenen Daten als auch die Daten aus der Beleglesung von Papierfragebogen wurden in standardisierter Form in die automatisierte Datenaufbereitung der Gebäude- und Wohnungszählung übermittelt. Dieser Prozess erfolgte kontinuierlich während der Datenerhebungsphase. Sobald die Daten aller Auskunftspflichtigen eines Gebäudes vorlagen, wurden die Daten für dieses Gebäude einschließlich aller dazugehörigen Wohnungen maschinell aufbereitet. Am Ende der Erhebungsphase erfolgte die maschinelle Aufbereitung ein letztes Mal auch für die Daten all derjenigen Gebäude, zu denen nicht alle erwarteten Antworten eingegangen waren.

Eine besondere Herausforderung der Gebäude- und Wohnungszählung waren Gebäude mit mehreren Eigentümerinnen und Eigentümern, und zwar insbesondere dann, wenn es sich um Gebäude mit Eigentumswohnungen gehandelt hat. Haben diese Eigentümerinnen beziehungsweise Eigentümer widersprüchliche Angaben zu den Gebäudemerkmalen[2] gemacht, war bei der Aufbereitung zu entscheiden, welche Angabe zu übernehmen war. Die von der jeweiligen Eigentümerin oder dem jeweiligen Eigentümer berichteten Angaben zu den einzelnen Wohnungen wurden immer als Ganzes übernommen. Um mit unterschiedlichen Angaben bei den Gebäudemerkmalen umzugehen, gab es eine Hierarchie nach Art der Eigentümer, zum Beispiel hatten Wohnungsunternehmen Vorrang gegenüber Privatpersonen. Gab es unterschiedliche Angaben in einer Gruppe von Eigentümerinnen und Eigentümern, wurde die meistgenannte Angabe übernommen.

1 Unter Karteileichen sind Personen zu verstehen, die laut Melderegister in einer Gemeinde leben, zum Stichtag der Erhebung aber nicht mehr an der im Melderegister geführten Anschrift wohnen. Personen, die an einer bestimmten Anschrift wohnen, jedoch nicht im Melderegister mit dieser Anschrift geführt werden, werden als Fehlbestände bezeichnet. Paarig existente Personen wohnen an der im Melderegister geführten Anschrift.

2 Gebäudemerkmale sind „Zahl der Wohnungen im Gebäude“, „Gebäudeart“ und „Gebäudetyp“, „Eigentumsverhältnisse“, „Baujahr“, „Heizung“ und „Energieträger“.

Bei widersprüchlichen Angaben zum Merkmal „Anzahl der Wohnungen im Gebäude“ wurde vor allem die Zahl der gelieferten Wohnungsdatensätze berücksichtigt.

Es kam vor, dass für einzelne Wohnungen eine Datenmeldung mehrfach abgegeben wurde, etwa weil eine Eigentümerin oder ein Eigentümer für ein und dieselbe Wohnung mehrfach gemeldet hatte oder weil mehrere Anschreiben für dieselbe Wohnung ergingen. Diese sogenannten Wohnungsdubletten zu erkennen und auszusortieren, war bei der Aufbereitung ebenfalls wichtig. Ergab die maschinelle Aufbereitung aufgrund der hinterlegten Regeln, dass die Anzahl der Wohnungen im Gebäude geringer war als die Anzahl der gelieferten Wohnungsdatensätze, wurden im nächsten Schritt die Wohnungsdubletten aussortiert. Um Wohnungsdubletten zu identifizieren, erfolgte unter anderem ein Vergleich der Bewohnernamen miteinander. Kamen Bewohnernamen mehrfach vor, wurde nur eine der Wohnungen übernommen.

3

Aufbereitung nach der Erhebungsphase

3.1 Referenzdatenbestand

Neben der erhebungsteilübergreifenden Plausibilisierung (EÜPL) Existenzen (Abschnitt 2.1) fand im Referenzdatenbestand noch eine Sonder-EÜPL nach Abschluss der Erhebungsphase statt, um die Qualität zu verbessern. Diese Überprüfung sollte gezielt noch einmal auffällige Klumpungen von (vorläufigen) Karteileichen im Zuständigkeitsbereich einzelner Erhebungsstellen kontrollieren. Bei einem Verdacht auf fehlerhafte Erhebung einer Anschrift konnten die Statistischen Landesämter diese prüfen und eine sogenannte Befundsetzung auf Anschriftenebene (Befund „Abgrenzungsproblem“ für Stichprobenanschriften beziehungsweise Befund „Anschrift konnte nicht abgeschlossen werden“ für Anschriften mit Sonderbereichen) durch das Statistische Bundesamt veranlassen. Eine nachträgliche Überprüfung und Übermittlung von Existenzen war aufgrund des großen zeitlichen Abstands zum Stichtag nicht sinnvoll. Diese Korrektur durch Befundsetzung (Befundkorrektur) hatte für Stichprobenanschriften eine Aussteuerung aus der Hochrechnung zur Ermittlung von Über-/Untererfassung zur Folge. Für Anschriften mit Sonderbereichen wurden durch die Befundkorrektur Personen an diesen Anschriften gemäß dem Ergebnis der Mehrfachfallprüfung gezählt. Die Befundkorrektur führte dazu, dass sich die Nettostichprobe verkleinerte, wodurch sich der Standardfehler geringfügig erhöhte. Dieser Nachteil wog jedoch geringer als die Berücksichtigung von Anschriften mit festgestellten Ungenauigkeiten bei der Datenerhebung. Im Ergebnis wurden bei diesen Anschriften die Angaben der Melderegister ohne primärstatistische Korrektur für die Einwohnerzahlermittlung übernommen.

3.2 Allgemeine Aspekte bei Primärstatistiken

Ziel der Plausibilisierung und Imputation der Zensusdaten war es, wie bei anderen Primärstatistiken auch, plausible, valide und vollständige Datensätze für die nachgelagerten Aufbereitungsschritte bereitzustellen. Die Kernziele dieser Aufbereitungsphase sind:

1. Identifikation und Kennzeichnung von Fehlern in den Daten

   Die Zensusdaten wurden auf folgende Fehler geprüft und entsprechend gekennzeichnet:

   - Unzulässig fehlende Angabe: Es liegen keine Daten vor, obwohl die Frage hätte beantwortet werden müssen (Item Nonresponse).
   - Filterfehler: Es liegen Daten vor, obwohl die Frage aufgrund der Filterführung nicht hätte beantwortet werden dürfen.
   - Unzulässige Mehrfachangabe: Es wurde mehr als eine Antwortmöglichkeit bei einer Frage ausgewählt, obwohl dies nicht zulässig war.
   - Strukturunplausibilität: Die Angabe entspricht nicht dem zulässigen Format oder liegt außerhalb des zulässigen Wertebereichs.
   - Interunplausibilität: Die Angabe ist widersprüchlich zu einer anderen Angabe dieser Befragungseinheit.

2. Fehlerbereinigungen

   Zur Behebung wurden zwei Verfahren verwendet:

   › die deterministische Imputation (Abschnitte 3.3 und 3.4) und

   › das Hot-Deck-Nächste-Nachbarn-Verfahren (Abschnitt 3.6).

Während im Zensus 2011 der primäre Erhebungsweg die Befragung durch Interviewerinnen und Interviewer beziehungsweise die Befragung mit Papierfragebogen war, setzte der Zensus 2022 auf die Online-First-Strategie (Freier/Mosel, 2019; Gaedke und andere, 2024). Das erwies sich nicht zuletzt hinsichtlich der Datenqualität als ausgesprochen zielführend, da einige Plausibilitätsprüfungen bereits im Online-Erhebungssystem implementiert werden und erste Fehlerprüfungen und -bereinigungen schon bei der Datenerhebung erfolgen konnten.

3.3 Plausibilisierung der Gebäude- und Wohnungszählung

Die Plausibilisierung der Gebäude- und Wohnungszählung wurde im ersten Schritt als maschinelles Verfahren durchgeführt. Die zu plausibilisierende Einheit war das Gebäude einschließlich aller darin befindlichen Wohnungen. Stattgefunden hat die Plausibilisierung, sobald die Gebäudedaten eines Gebäudes aufbereitet worden waren.

Geprüft wurde auf Vollständigkeit, unzulässige Mehrfachcodierungen, Strukturplausibilität und logische Zusammenhänge zwischen den Merkmalen (Interplausibilität). Für jede Unplausibilität wurde ein Fehlerschlüssel gesetzt. In einigen Fällen konnte eine deterministische Fehlerkorrektur erfolgen. Wenn zum Beispiel bei einer Wohnung die Wohnfläche gefehlt hat, aber eine Raumanzahl angegeben war, wurde geprüft, ob es im Gebäude eine andere Wohnung mit der gleichen Anzahl der Räume gab. War dies der Fall, wurde die Fläche dieser Wohnung übernommen. Alle deterministischen Korrekturen wurden ebenfalls durch Fehlerschlüssel gekennzeichnet.

Mit Fehlerschlüsseln gekennzeichnete Datensätze konnten von den Mitarbeiterinnen und Mitarbeitern der Statistischen Landesämter bei Bedarf im zweiten Schritt manuell geprüft und, wenn möglich, korrigiert werden. Besonders wichtig war dies bei dem Fehlerschlüssel, der eine Abweichung vom Gebäudemerkmal „Anzahl der Wohnungen im Gebäude“ und den tatsächlich vorhandenen Wohnungsdatensätzen aufzeigte.

Zusätzlich gab es eine Prüfung der erhebungsteilübergreifenden Plausibilität zwischen der Gebäude- und Wohnungszählung und dem Melderegister beziehungsweise der Personenerhebung. Diese fand beispielsweise statt, wenn an einer Anschrift mehr bewohnte Wohnungen in der Gebäude- und Wohnungszählung erhoben worden waren als Haushalte von der Personenerhebung. Bei der erhebungsteilübergreifenden Plausibilität wurden dabei lediglich sogenannte Prüffälle erstellt, die dazu dienten, auf mögliche Fehler hinzuweisen und gegebenenfalls einzelne Korrekturen vorzunehmen.

3.4 Plausibilisierung der Personenerhebung

Die Plausibilisierung (PL) der Daten aus der Personenerhebung wurde als maschinelles Verfahren konzipiert. Im ersten Schritt galt es zu berücksichtigen, dass einige Personenangaben aus bis zu drei Quellen vorliegen konnten:

1. Sogenannte elektronische Erhebungsliste der Personenerhebung, das heißt Kerndaten zur Person, die im Zuge der Existenzfeststellung durch die Interviewerinnen und Interviewer persönlich aufgenommen wurden,
2. Fragebogen der Personenerhebung und/oder
3. Melderegister.

Aus dem vorliegenden Datenmaterial wurde zunächst die für die weitere Datenaufbereitung herangezogene Datenquelle bestimmt. Lagen für Merkmale mehrere Datenquellen vor, kam folgende Vorfahrtsregel zur Anwendung: Melderegisterangaben[3] vor Angaben aus den Interviews (elektronische Erhebungsliste) vor Antworten aus dem Fragebogen.

3 Gemäß § 11 Absatz 4 Zensusgesetz 2022 umfasst die Feststellung nach Absatz 1 (Haushaltebefragung auf Stichprobenbasis) nicht die Berichtigung der aus den Melderegistern übernommenen Daten zur Person. Die Melderegisterdaten waren somit – sofern vorhanden und plausibel – vorrangig zu verwenden. Eine Vorfahrt der Angaben aus der Erhebungsliste folgte hingegen auf Basis einer fachlichen Bewertung.

Des Weiteren wurden in der Personenerhebung unterschiedliche Fragebogen eingesetzt, die sich hinsichtlich des Merkmalskranzes und des Fragebogenaufbaus unterschieden. Die Kurzbefragungen (Haushalte und Wohnheime) dienten ausschließlich dazu, Merkmale zur Ermittlung der Einwohnerzahlen zu erheben. Sie unterschieden sich dahingehend voneinander, dass die Befragung an Wohnheimen über die Haushaltebefragung weitergehende Informationen wie etwa zum Geburtsnamen und zum Geburtsstaat einer Person erfasste. Mit den Zusatzbefragungen wurden weitere Merkmale – etwa zur Bildung und zur Erwerbstätigkeit – erhoben, die nicht (ausreichend) aus Verwaltungsregistern gewonnen werden konnten. Die meisten Fragen zur Bildung und Erwerbstätigkeit richteten sich nur an Personen im Alter von 15 Jahren und älter. Die Haushaltebefragung und die Befragung an Wohnheimen unterschieden sich hierbei nicht.

Um dies bei den Fehlerprüfungen und -bereinigungen entsprechend berücksichtigen zu können, wurde das vorliegende Datenmaterial in die nachfolgenden Teilbestände (PL-Klassen) unterteilt:

1. Kurzbefragung an Haushalten
2. Kurzbefragung an Wohnheimen
3. Haushaltebefragung und Befragung an Wohnheimen für Personen unter 15 Jahren (Zusatzbefragung)
4. Haushaltebefragung und Befragung an Wohnheimen für Personen im Alter von 15 Jahren und älter (Zusatzbefragung)

Anschließend erfolgten nach PL-Klassen getrennt die Prüfungen der Daten auf Vollständigkeit, korrekte Beachtung der Filterführung, auf unzulässige Mehrfachangaben, auf Strukturplausibilität sowie auf Interplausibilität. Identifizierte Fehler wurden gekennzeichnet und gegebenenfalls bereits deterministisch bereinigt. Die deterministische Bereinigung wurde nur durchgeführt, wenn eine eindeutige Beziehung zwischen einem unplausiblen beziehungsweise fehlenden Merkmal und einem oder mehreren plausiblen Merkmalen vorlag. Wenn beispielsweise der Familienstand einer Person laut Melderegister „ledig“ ist, die Person im Fragebogen aber angegeben hat, dass sie mit einer Partnerin beziehungsweise einem Partner zusammenwohnt und mit dieser/diesem verheiratet ist, liegt eine Interunplausibilität zwischen den vorliegenden Informationen vor. In solchen Fällen wurde unter Berücksichtigung der Vorfahrtsregel die Angabe aus dem Fragebogen an die Angaben aus dem Melderegister angepasst und deterministisch dahingehend bereinigt, dass die Person zwar mit einer Partnerin beziehungsweise einem Partner zusammenwohnt, jedoch nicht mit dieser/diesem verheiratet ist.

Zur Fehlerdokumentation waren in dieser Aufbereitungsphase zwei Schritte vorgesehen: die bereits beschriebene Kennzeichnung der Fehler und das Anlegen sogenannter Qualitätskennzeichen. Die Kennzeichnung der Fehler lässt Rückschlüsse auf die Art des Fehlers zu, also ob beispielsweise eine Filterung missachtet wurde. Die Qualitätskennzeichen geben an, ob ein Merkmal plausibel ist, durch deterministische Imputation bereits verändert wurde oder unplausibel ist und durch das Hot-Deck-Verfahren (Abschnitt 3.6) korrigiert werden muss (Hentschke und andere, 2024). Bei plausiblen Merkmalen kann dem Qualitätskennzeichen zudem entnommen werden, aus welcher Quelle der plausible Wert ursprünglich stammt. Dies war wichtig, da Melderegisterangaben aufgrund von § 11 Absatz 4 Zensusgesetz 2022 stets unverändert bleiben mussten. Somit durften nur Werte, die ursprünglich aus dem Fragebogen oder der elektronischen Erhebungsliste stammten, mithilfe deterministischer Imputation oder Hot-Deck-Verfahren im Nachhinein nochmals verändert werden.

3.5 Plausibilisierung der Gemeinschaftsunterkünfte

Die Plausibilisierung der Personendaten aus den Gemeinschaftsunterkünften erfolgte analog zur Vorgehensweise der Personenerhebung. Als Besonderheiten sind die nachstehenden Punkte hervorzuheben:

- Personendaten konnten aus bis zu zwei Datenquellen vorliegen (Melderegister und elektronische Erhebungsliste).
- Für die Gemeinschaftsunterkünfte gab es einen reduzierten Fragebogen, da die Einrichtungsleitungen stellvertretend für die Bewohnerinnen und Bewohner auskunftspflichtig waren. Daher war es für die Aufbereitung nicht notwendig, das vorliegende Datenmaterial wie in der Personenerhebung in Teilbestände aufzuteilen.

› Die Gemeinschaftsunterkünfte bildeten eine eigene PL-Klasse, da der Umfang der erhobenen Merkmale geringer war.

› Aufgrund des Befragungsaufbaus wurden lediglich vier Fehlerprüfungen vorgenommen. Die Prüfung auf falsche Filterführung entfiel, da die Befragungsstruktur kein Überspringen von Fragen vorsah.

3.6 Imputationssoftware CANCEIS

Im Anschluss an die beschriebenen Plausibilisierungsschritte und die deterministische Imputation erfolgte für alle primärstatistischen Erhebungsteile separat die sogenannte Hot-Deck-Imputation. Diese wurde mittels der Software CANCEIS[4] durchgeführt, die einen Nächste-Nachbarn-Ansatz verwendet. Die Grundidee ist, alle fehlenden und fehlerhaften Werte eines Datensatzes (= Empfänger) mit den beobachteten Werten eines vollständigen und plausiblen Datensatzes (= Spender) aus derselben Erhebung zu ersetzen. Ziel dieses Imputationsverfahrens ist, einen vollständigen und plausiblen Datenbestand auszugeben.

CANCEIS wurde im Zensus 2022 für alle primärstatistischen Erhebungsteile (Gebäude- und Wohnungszählung, Personenerhebungen, Befragung an Anschriften mit Sonderbereichen) eingesetzt. Die Software kam bereits bei der Gebäude- und Wohnungszählung im Zensus 2011 zum Einsatz (Grunwald/Krause, 2014). Innerhalb der Erhebungsteile wurden die Daten vor der spenderbasierten Imputation analog zum Vorgehen bei der Plausibilisierung in verschiedene Imputationsklassen aufgeteilt, die die Daten in homogene Untergruppen unterteilten und einen einheitlichen Merkmalskranz umfassten (Abschnitte 3.4 und 3.5). Die Imputation fand innerhalb jeder Imputationsklasse separat statt, sodass als Spender immer nur ein Datensatz aus derselben Imputationsklasse wie der des Empfängers verwendet wurde.

Für jeden Empfänger wurde genau ein Spender verwendet. Das sollte sicherstellen, dass der imputierte Datensatz nicht nur plausibel wurde, sondern auch in der Erhebung selbst beobachtete Kombinationen von Merkmalswerten enthielt. Es wurde ein Spender gesucht, der dem Empfänger in allen im Zensus erhobenen Merkmalen ähnlich war. Dem zugrunde liegt die Annahme, dass Zusammenhänge zwischen allen Bereichen und Merkmalen einer Erhebung bestehen. Ein und derselbe Spender kann für mehrere Empfänger spenden. Um die Werte des Spenders aber nicht zu häufig zu duplizieren, wurde für jeden Erhebungsteil eine Höchstgrenze festgelegt, wie häufig ein plausibler Datensatz maximal spenden durfte.

Die in CANCEIS implementierte Methodik bietet gegenüber anderen Verfahren den Vorteil, dass sie Plausibilitätsregeln bei der Imputation direkt berücksichtigen kann. Das heißt bei Inkonsistenzen zwischen Merkmalen ist nicht zunächst eine Identifizierung derjenigen Felder vorzunehmen, die imputiert werden müssen, und anschließend eine Imputation durchzuführen. Vielmehr entscheidet CANCEIS bei Inkonsistenzen zwischen Merkmalen anhand der Daten und der definierten und CANCEIS bereitgestellten Fehlerbeschreibungen, welcher Wert ersetzt werden soll (beziehungsweise welche Werte ersetzt werden sollen). Dabei wird der Grundsatz verfolgt, dass nur potenziell fehlerhafte Werte korrigiert und möglichst wenige der beobachteten Werte angepasst werden (Bankier, 2012).

Für die Imputation unterteilt CANCEIS einen Datenbestand zunächst in Empfänger- und Spenderdatensätze. Bei den Empfängern handelt es sich um all jene Datensätze, die mindestens eine unter Abschnitt 3.2 Nr. 1 beschriebene Fehlerart aufweisen. Die anderen, vollständig plausiblen Datensätze bilden den Pool der Spenderdatensätze. Anschließend wird für jeden Empfängerdatensatz ein Spenderdatensatz gesucht, der dem Empfänger in den vorhandenen Angaben möglichst ähnlich ist. Um die Ähnlichkeit zu bestimmen, wird die mathematische Distanz zwischen Empfänger und Spender mittels einer Distanzfunktion berechnet, die die in der Erhebung vorkommenden Merkmale berücksichtigt, die auch beim Empfänger erhoben werden. Von den Spendern mit den geringsten Distanzen zum Empfänger wird einer ausgewählt, dessen Werte für die zu ersetzenden Werte des Empfängers eingesetzt werden, sodass der imputierte Empfänger vervollständigt wird und alle Prüfregeln einhält.

Um alle Imputationen auch im späteren Prozessverlauf nachvollziehen zu können, wurden beim Imputationsprozess, ergänzend zu den Qualitätskennzeichen der

4 Die Software CANCEIS (Canadian Census Edit and Imputation System) ersetzt fehlende oder unplausible Angaben der Empfängerdatensätze mit beobachteten Werten von Spenderdatensätzen aus derselben Erhebung.

Plausibilisierung (Abschnitt 3.4), weitere Qualitätskennzeichen angelegt. Sie gaben für jeden Wert an, ob dieser durch die spenderbasierte Imputation eingesetzt oder verändert wurde oder nicht.

3.7 Imputation der Gebäude- und Wohnungszählung

Die spenderbasierte Imputation mit CANCEIS wurde bei der Gebäude- und Wohnungszählung für zwei Arten von Antwortausfällen eingesetzt:

Unit Nonresponse: Kompletter Antwortausfall auf Gebäudeebene, es liegen keinerlei Daten zum Gebäude oder den darin befindlichen Wohnungen vor.

Item Nonresponse: Einzelne fehlende oder fehlerhafte Werte bei Gebäude- oder Wohnungsmerkmalen sowie fehlende Angaben zu ganzen Wohnungen, die sich innerhalb eines Gebäudes befinden, zu dem Rückmeldungen zu weiteren Wohnungen vorliegen.

Die im Datenmaterial vorhandenen vollständigen und plausiblen Gebäude können dabei sowohl bei der Unit- als auch bei der Item-Nonresponse-Imputation als Spender dienen.

Beim Unit Nonresponse lagen weder zum Gebäude noch zu den Wohnungen Erhebungsdaten vor, hier konnte die spenderbasierte Imputation nur unter Zuhilfenahme von Zusatzinformationen sinnvoll durchgeführt werden. Hierzu wurden Angaben aus vorliegenden Registern zur räumlichen Lage der Gebäude sowie Informationen aus externen Quellen zur Zahl der Haushalte und Personen an der Gebäudeanschrift in die Spendersuche einbezogen. Dieses Vorgehen stellte sicher, dass das Spendergebäude in derselben Region lag und die korrekte Gebäudegröße beziehungsweise Zahl an bewohnten Wohnungen aufwies. Für das zu imputierende Gebäude wurden dann alle Angaben zu dem Gebäude und den darin enthaltenen Wohnungen vom Spendergebäude übernommen.

Bei der Imputation des Item Nonresponse erfolgte die Imputation anhand zweier verschiedener Vorgehensweisen je nach Größe des Empfängergebäudes. Bei Gebäuden mit bis zu 15 Wohnungen wurden die Gebäude entsprechend der Anzahl der Wohnungen im Gebäude in 15 Imputationsklassen aufgeteilt. Um die Größenstruktur der Gebäude zu berücksichtigen, kamen dann nur Spendergebäude infrage, die eine identische Anzahl Wohnungen aufwiesen. Dieses Vorgehen stellte sicher, dass beispielsweise Einfamilienhäuser auch nur Spender für andere Einfamilienhäuser sein konnten. Bei Gebäuden mit mehr als 15 Wohnungen konnte dieses Verfahren nicht beibehalten werden, da bei sehr großen Gebäuden gar keine oder nur sehr wenige (Spender-) Gebäude zur Verfügung standen. Diese Gebäude wurden in einer gemeinsamen Imputationsklasse für CANCEIS bereitgestellt. Dabei erfolgte die Imputation schrittweise, wobei zunächst die Wohnungen mit ihren Wohnungsmerkmalen und anschließend die Gebäudemerkmale imputiert wurden. Die Imputation der Wohnungen erfolgte bei beiden Vorgehensweisen unter Berücksichtigung der Anzahl der Haushalte an der Anschrift.

Einen Einblick in die Quoten der Unit- und Item-Nonresponse-Imputation bei der Gebäude- und Wohnungszählung des Zensus 2022 bietet der Aufsatz von Hentschke und anderen (2024).

3.8 Imputation der Personenerhebung

Bei der Personenerhebung fand ausschließlich eine Item-Nonresponse-Imputation statt, das heißt es wurden einzelne fehlende oder fehlerhafte Merkmale (zum Beispiel zur Klassenstufe) von Personen ergänzt. Eine Unit-Nonresponse-Imputation fand bei der Personenerhebung nicht statt. Fehlten komplette Rückmeldungen von Auskunftspflichtigen, wurde dies bei der Hochrechnung der Ergebnisse berücksichtigt.

Dabei galt es, die im Zensus 2022 definierte Vorfahrtsregel zu berücksichtigen, dass Personen, die sowohl im Melderegister als auch der primärstatistischen Erhebung enthalten sind, bei demografischen Angaben immer die Informationen aus dem Melderegister erhalten. So wurden die Angaben aus den Melderegistern, die in der Aufbereitungsphase bereits plausibilisiert wurden (Abschnitt 3.4), durch die spenderbasierte Imputation nicht mehr verändert.

Bei der Personenerhebung unterteilten sich die Imputationsklassen nach dem Merkmalskranz der Befragung und dem Alter der Befragten (PL-Klassen). Für die Hauptbefragung wurden die unter 15-Jährigen und die Personen, die 15 Jahre oder älter sind, getrennt voneinander imputiert, da den ab 15-Jährigen unter anderem auch

Fragen zur Berufstätigkeit gestellt wurden. Um die Spendersuche zu beschleunigen, wurden die Datensätze nach dem Alter sortiert in CANCEIS eingegeben, sodass Datensätze mit ähnlichem Alter zuerst als potenzielle Spender untersucht wurden.

Personen an Gemeinschaftsunterkünften wurden gemeinsam in einer Imputationsklasse plausibilisiert und imputiert.

4

Fazit

Mit der übergreifenden Plausibilisierung der verschiedenen Teilprojekte sowie den Plausibilisierungsroutinen innerhalb der elektronischen Erhebungsphase war es möglich, Erfassungs- und Abgrenzungsfehler bereits früh im Prozess der Aufbereitung des Zensus 2022 zu identifizieren und angemessene Maßnahmen zur Korrektur zu ergreifen. Die eingesetzten Plausibilisierungs- und Imputationsverfahren bei den primärstatistischen Erhebungsteilen konnten die nach Abschluss der Erhebungsphase noch vorhandenen fehlerhaften und fehlenden Werte identifizieren und korrigieren. Als Resultat der durchgeführten Aufbereitungsschritte steht ein stimmiger Zensusdatenbestand für Auswertungen zur Verfügung.

LITERATURVERZEICHNIS

Bankier, Mike. *Imputing Numerical and Qualitative Variables Simultaneously*. Social Survey Methods Division, Statistics Canada (internes Dokument). 2012.

Freier, Benjamin/Mosel, Juliane. *Online First als Leitgedanke für effiziente Primärerhebungen beim Zensus 2021*. In: WISTA Wirtschaft und Statistik. Sonderheft Zensus 2021. Wiesbaden 2019, Seite 46 ff.

Gaedke, Annika/Pfahl, Miriam/Strohalm, Anna. *Die Online-First-Strategie und die Online-Fragebogen im Zensus 2022*. In: WISTA Wirtschaft und Statistik. Ausgabe 6/2024, Seite 92 ff.

Grunwald, Sven/Krause, Anja. *Umgang mit fehlenden Angaben in der Gebäude- und Wohnungszählung 2011*. In: Wirtschaft und Statistik. Ausgabe 8/2014, Seite 437 ff.

Hentschke, Janine/Tobies, Cara-Aileen/Weber, Susanne/Claus, Dennis. *Kern-Qualitätskennzahlen und Zielwerte im Zensus 2022*. In: WISTA Wirtschaft und Statistik. Ausgabe 6/2024, Seite 106 ff.

Preising, Marcel/Lange, Kerstin/Dumpert, Florian. *Imputation zur maschinellen Behandlung fehlender und unplausibler Werte in der amtlichen Statistik*. In: WISTA Wirtschaft und Statistik. Ausgabe 5/2021, Seite 40 ff.

RECHTSGRUNDLAGEN

Gesetz zur Durchführung des Zensus im Jahr 2022 (Zensusgesetz 2022 – ZensG 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 Dezember 2024
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-24006-4, ISSN 1619-2907