Dipl.-Mathematiker Sven Grunwald, Dipl.-Geographin Anja Krause

Umgang mit fehlenden Angaben in der Gebäude- und Wohnungszählung 2011

Im Rahmen des Zensus 2011 fand eine Gebäude- und Wohnungszählung (GWZ) statt, bei der Angaben zu 19 Millionen Gebäuden und den zugehörigen Wohnungen erhoben wurden. Vor der Veröffentlichung der Ergebnisse mussten die Angaben auf ihre Plausibilität untersucht werden, das heißt ob sie vollständig und widerspruchsfrei sind. Aufgrund der großen Datenmenge standen die Statistischen Ämter des Bundes und der Länder dabei vor einer besonderen Herausforderung: Wie kann eine derart große Anzahl an Datensätzen überhaupt geprüft werden? (Wie) schafft man es, alle Unplausibilitäten aufzudecken? Woher weiß man eigentlich, wie ein Merkmal korrigiert werden muss? Und wie bewertet man, ob die durchgeführten Korrekturen die Qualität der Daten wirklich verbessert haben?

Der folgende Beitrag beleuchtet fehlerhafte Angaben und Antwortausfälle in der Gebäude- und Wohnungszählung 2011 sowie ihre möglichen Ursachen.

Notwendige Korrekturen wurden unter anderem mithilfe der vom kanadischen Statistikamt entwickelten Software CANCEIS[1] *durchgeführt. Dieses zum ersten Mal in der amtlichen Statistik in Deutschland eingesetzte Programm wird ebenfalls im Folgenden vorgestellt und es wird erläutert, wie die Software in den Datenaufbereitungsprozess der Gebäude- und Wohnungszählung integriert wurde.*

1 Die Gebäude- und Wohnungszählung – ein kurzer Überblick

Wichtigstes Ziel der Gebäude- und Wohnungszählung 2011 war, Informationen über den Gebäude- und Wohnungsbestand sowie über die Wohnverhältnisse der Bevölkerung in Deutschland zu gewinnen. Die Zählung war dabei eine Art „Inventur“, bei der die Datenlage zu Gebäuden und Wohnungen aktualisiert und damit auch eine neue Grundlage für andere Statistiken aus dem Bereich Bauen und Wohnen geschaffen wurde. Die letzten Gebäude- und Wohnungszählungen fanden 1987 im früheren Bundesgebiet im Rahmen der letzten Volkszählung und 1995 in den neuen Ländern und Berlin-Ost als eigene Erhebung zum Gebäude- und Wohnungsbestand statt.

Um die Befragten zu entlasten und Kosten zu sparen, wurde für den Zensus 2011 eine neue Methode entwickelt, bei der überwiegend Informationen aus Verwaltungsregistern genutzt wurden. Da nicht alle benötigten Angaben in Registern vorhanden waren, sollten ergänzend Befragungen durchgeführt werden. Für die Gebäude- und Wohnungszählung sah das Zensusmodell dabei eine Vollerhebung vor, sodass Daten zu allen Gebäuden mit Wohnraum und zu den zugehörigen Wohnungen erhoben wurden. Damit war die Gebäude- und Wohnungszählung die umfangreichste Erhebung im Zensus. Ihre Ergebnisse stehen flächendeckend bis auf Gemeindeebene zur Verfügung und können unter *https://ergebnisse.zensus2011.de* abgerufen werden.

Die Daten für die Gebäude- und Wohnungszählung wurden in der Regel bei den jeweiligen Haus- beziehungsweise Wohnungseigentümerinnen und -eigentümern sowie bei

1 CANadian Census Edit and Imputation System.

Verwaltungen schriftlich mithilfe eines Fragebogens erhoben. Wie bei allen Befragungen im Rahmen des Zensus bestand Auskunftspflicht. Die zählungsrelevanten Gebäude und die für die schriftliche Befragung erforderlichen Anschriften der Auskunftspflichtigen wurden im Vorfeld über Informationen aus der Verwaltung (zum Beispiel mithilfe der Melderegister, Daten der Bundesagentur für Arbeit[2], von Grundsteuerstellen oder von Versorgungs- und Entsorgungsbetrieben[3]) gewonnen.[4] Stichtag war der 9. Mai 2011. Im Rahmen der Gebäude- und Wohnungszählung wurden allen Auskunftspflichtigen Fragebogen zu ihren Gebäuden und Wohnungen zugeschickt. Die Beantwortung konnte postalisch oder online über einen elektronischen Fragebogen erfolgen. Im Januar 2012 waren Meldungen zu etwa 90 % der Gebäude eingegangen. Insgesamt wurden Angaben zu 19 Millionen Gebäuden mit Wohnraum und 40,5 Millionen Wohnungen erhoben.[5]

Übersicht 1

Erhebungsmerkmale der Gebäude- und Wohnungszählung 2011[1]

Gebäude:
Art des Gebäudes
Anzahl der Wohnungen
Gebäudetyp
Baujahr
Eigentumsverhältnisse des Gebäudes
Heizungsart

Wohnung:
Art der Wohnungsnutzung
Wohnfläche
Raumzahl
Eigentumsverhältnisse der Wohnung
Badewanne/Dusche vorhanden
WC vorhanden
Wohnungstyp (Ferien-/Freizeit- oder Diplomatenwohnung)
Anzahl der Bewohner/-innen (Hilfsmerkmal)
Namen von bis zu zwei Bewohnern/Bewohnerinnen (Hilfsmerkmal)

1 Nach § 6 Zensusgesetz 2011.

2 Antwortausfälle und Unplausibilitäten

Wie in allen Statistiken mussten auch in der Gebäude- und Wohnungszählung 2011 die Daten vor der Auswertung auf ihre Vollzähligkeit, Vollständigkeit und Plausibilität geprüft und – falls notwendig – korrigiert und ergänzt werden. Wurden von den Auskunftspflichtigen einzelne Fragen nicht beantwortet, fehlten also in den Daten einzelne Angaben zu einem Gebäude und/oder zu den zugehörigen Wohnungen, bezeichnet man dies als *Item Nonresponse*. In der Gebäude- und Wohnungszählung wurden aber nicht nur Antwortausfälle bei einzelnen Merkmalen oder ganzen Wohnungen, sondern auch unplausible beziehungsweise fehlerhafte Angaben als Item-Nonresponse-Fälle verstanden. Diese Erweiterung erschien sinnvoll, da sowohl fehlerhafte als auch fehlende Merkmale mit den gleichen Verfahren korrigiert beziehungsweise vervollständigt wurden.

Fehlten alle Angaben zu einem Gebäude, so bezeichnet man dies als *Unit Nonresponse*.

2.1 Ursachen für Item Nonresponse

Die Gründe von Item Nonresponse können ganz unterschiedlich sein. So können Fehler und Antwortausfälle zum einen während der Erhebung selbst auftreten und zum Beispiel mit dem Inhalt oder der Struktur der Fragen beziehungsweise des Fragebogens zusammenhängen. Zum anderen können aber auch nach Abschluss der Erhebungsphase während der Digitalisierung und Aufbereitung noch Fehler in die Daten gelangen.

2.1.1 Fehler und Antwortausfälle während der Datenerhebung

Dillman und andere geben einen Überblick über mögliche Einflussfaktoren auf die Item-Nonresponse-Rate.[6] Einige dieser Faktoren spielten auch bei fehlerhaften Angaben und Antwortausfällen in der Gebäude- und Wohnungszählung eine Rolle.

Art der Befragung

In der Gebäude- und Wohnungszählung wurde ein Fragebogen eingesetzt, der von den Auskunftspflichtigen selbst ausgefüllt werden musste. Dabei blieb es den Befragten überlassen, ob sie zusätzliche Erläuterungen zu den Fragen gelesen haben, in welcher Reihenfolge sie den Fragebogen bearbeiteten oder ob sie Fragen übersprungen haben. Darüber hinaus war auch keine Interviewerin und kein Interviewer anwesend, mit deren beziehungsweise dessen Hilfe gegebenenfalls auftretende Verständnisfragen hätten geklärt werden können.[7] Wie stark dies die Antworten bei der Gebäude- und Wohnungszählung beeinflusst hat, lässt sich im Nachhinein nur schwer einschätzen, zumal nur wenige Fragen gestellt wurden.

Zusätzlich wurde in der Gebäude- und Wohnungszählung ein Online-Fragebogen eingesetzt, in den einige Plausibilitätsprüfungen der Angaben integriert wurden. Er hat dazu beigetragen, den Anteil der Fehler und Antwortausfälle zu verringern, da die Auskunftspflichtigen auf ausgelassene Fragen oder inkonsistente Antworten hingewiesen wurden und diese dann entsprechend korrigieren konnten (siehe Abschnitt 4.2).

2 Siehe §§ 4 bis 6 Zensusvorbereitungsgesetz 2011.
3 Siehe § 10 Zensusvorbereitungsgesetz 2011.
4 Zum Aufbau des Anschriften- und Gebäuderegisters, in das diese Angaben eingeflossen sind, siehe auch Ziprik, K.: „Qualitätsaspekte des Anschriften- und Gebäuderegisters im Zensus 2011“ in WiSta 11/2012, Seite 947 ff.
5 Einen guten Überblick über die Vorbereitung und Durchführung der Gebäude- und Wohnungszählung geben Pruschwitz, A./Martschinke, A.: „Die Gebäude- und Wohnungszählung. Vorbereitung und Durchführung der Erhebung im Land Bremen“ in „Zensus 2011 – Vorbereitung und Durchführung im Land Bremen“, Statistische Mitteilungen, Heft 115, Seite 41 ff.
6 Siehe Dillman, D./Eltinge, J./Groves, R./Little, R.: “Survey Nonresponse in Design, Data Collection, and Analysis” in Groves, R./Dillman, D./Eltinge, J./Little, R. (Herausgeber): “Survey Nonresponse”, New York 2001, Seite 3 ff.
7 Siehe Dillman, D./Eltinge, J./Groves, R./Little, R. (Fußnote 6), hier: Seite 13.

Inhalt der Fragen

Im Allgemeinen kann der Inhalt der Fragen – insbesondere, wenn es sich um sensible Themen handelt – zu einer Häufung von fehlenden, falschen oder ungenauen Angaben führen.[8] Bei der Gebäude- und Wohnungszählung 2011 wurde im Vorfeld erwartet, dass die Frage nach (bis zu) zwei Namen von Bewohnern/Bewohnerinnen der Wohnungen häufig nicht beantwortet werden würde. Es wurde vermutet, dass Auskunftspflichtige diese Information – insbesondere, wenn es sich um die Namen ihrer Mieter/-innen handelt – nicht ohne Rücksprache übermitteln wollen. Diese Befürchtung hat sich allerdings nicht bestätigt. Die Bewohnernamen fehlten nur bei etwa 1 % der von Auskunftspflichtigen übermittelten Angaben zu bewohnten Wohnungen. Allerdings konnte bisher nicht ausgewertet werden, wie häufig zu diesem Merkmal (offensichtlich) falsche Angaben (zum Beispiel der Name „Donald Duck“ oder Ähnliches) übermittelt wurden.

Struktur der Fragen

Ein häufiger Einsatz von Filtern in einem Fragebogen, also von Hinweisen, welche Fragen bei einer bestimmten Antwort übersprungen werden können, kann die Wahrscheinlichkeit erhöhen, dass Fragen versehentlich ausgelassen werden, die eigentlich beantwortet werden müssen.[9] In der Gebäude- und Wohnungszählung 2011 traten Unplausibilitäten aufgrund der Filterführung verstärkt beim Merkmal „Eigentumsverhältnisse der Wohnung“ auf. Dieses Merkmal musste nur ausgefüllt werden, wenn es sich bei der Wohnung um eine Eigentumswohnung handelte. Offensichtlich hatten Auskunftspflichtige Probleme, dies richtig zu verstehen, sodass dieses Merkmal relativ oft fälschlicherweise ausgefüllt wurde.

Schwierigkeit der Fragen

Verständnisprobleme bei den Fragen der Gebäude- und Wohnungszählung gab es vor allem bei besonderen Konstellationen der Besitzverhältnisse von Gebäuden oder Wohnungen (zum Beispiel bei Erbbaurecht oder Gebäuden mit Eigentumswohnungen). Zudem lagen einige Informationen bei den Auskunftspflichtigen nicht oder nicht in der gewünschten Form vor. Ein Unterschied bestand zum Beispiel darin, ob Verwaltungen oder Eigentümer/-innen die Angaben übermittelten. Manchen Verwaltungen von Gebäuden mit Eigentumswohnungen lagen nicht alle Informationen zur Größe und Ausstattung der Wohnungen vor, sodass in solchen Fällen zusätzlich die Wohnungseigentümer/-innen kontaktiert werden mussten.

Abgrenzung von Erhebungseinheiten

Dieser Faktor steht ein wenig außerhalb der bisher aufgezählten Fehlerquellen. In der Gebäude- und Wohnungszählung 2011 spielte die fehlerhafte Abgrenzung der Gebäude durch die Auskunftspflichtigen aber eine wichtige Rolle. Insbesondere bei Gebäudeblöcken mit mehreren Eingängen und mehreren separaten Treppenhäusern trat das Problem auf, dass den Auskunftspflichtigen nicht immer klar war, was für die Gebäude- und Wohnungszählung ein Gebäude ist und für welchen Teil deshalb Angaben gemacht werden müssen. So kam es vor, dass für einen solchen Gebäudeblock (mit mehreren Eingängen, also nach Definition der Gebäude- und Wohnungszählung mehrere Gebäude) mehrfach die Daten für den gesamten Block übermittelt wurden. Im Ergebnis wurde dabei Gebäuden eine zu hohe Zahl an Wohnungen zugeschrieben. Diese Art Fehler konnte im Datenaufbereitungsprozess nur mit viel Aufwand identifiziert und korrigiert werden (siehe Abschnitt 3.3.2).

Es bestätigt sich, dass insbesondere der Konzeption des Fragebogens eine besondere Bedeutung zukommt, da Item Nonresponse durch Frageformulierung, Vorgabe der Antwortmöglichkeiten, Erläuterungen und Design des Fragebogens verringert werden kann.[10] Im Vorfeld der Gebäude- und Wohnungszählung 2011 wurde der Fragebogen in einem qualitativen Pretest mit 18 Probandinnen und Probanden überprüft. Im Anschluss wurden aufgrund der Testergebnisse noch einige wichtige Veränderungen am Fragebogen vorgenommen. Durch die geringe Anzahl der getesteten Personen traten jedoch manche Konstellationen (zum Beispiel Erbbaurecht) überhaupt nicht auf und konnten damit auch nicht im Voraus als Problem identifiziert werden.

2.1.2 Fehler und Antwortausfälle während der Digitalisierung und Datenaufbereitung

Datenfehler können auch nach Abschluss der Erhebungsphase bei der Umwandlung der Angaben in ein elektronisches Datenformat (Digitalisierung zum Beispiel durch Beleglesung) oder in der Datenaufbereitungsphase entstehen.[11]

Digitalisierung

Die Papierfragebogen der Gebäude- und Wohnungszählung 2011 wurden über spezielle Scanner erfasst (Beleglesung). Auch hierbei sind Fehler aufgetreten, etwa vereinzelt aufgrund von Verschmutzungen auf dem Fragebogen. Weit bedeutender waren aber Probleme bei der automatischen Handschrifterkennung (Optical Character Recognition – OCR). So wurden teilweise numerische Werte wie die „Anzahl der Wohnungen“, das „Baujahr“ oder die „Wohnfläche“ nicht korrekt erfasst, weil zum Beispiel eine „0“ als „8“ oder als „6“ interpretiert wurde. Aufgefallen sind diese Beleglesefehler vor allem beim Merkmal „Anzahl der Wohnungen“, da dadurch in etlichen Fällen von zu vielen Wohnungen in einem Gebäude ausgegangen wurde. Schrieb zum Beispiel ein Auskunftspflichtiger in das dreistellige Feld für das Merkmal „Anzahl der Wohnungen“ „001“, so konnte es vorkommen, dass dies von der Maschine fälschlicherweise als „601“ interpretiert wurde. Daraufhin wurden 601 Wohnungen angelegt und später vervollständigt (imputiert), weil das Merkmal „Anzahl der Wohnungen“ zentral für die Plausibilitätsprüfung der Gebäudegröße war und keine anderen Merkmale zu deren Überprüfung (zum Beispiel Zahl

8 Siehe Tourangeau, R./Rips, L. J./Rasinski, K.: “The Psychology of Survey Response”, Cambridge 2000, Seite 264 f.

9 Siehe Dillman, D./Eltinge, J./Groves, R./Little, R. (Fußnote 6), hier: Seite 14.

10 Siehe De Leeuw, E. D./Hox, J./Huisman, M.: “Prevention and Treatment of Item Nonresponse” in Journal of Official Statistics, Jahrgang 19, Ausgabe 2, Seite 162.

11 Siehe Graham, J. W.: “Missing Data. Analysis and Design”, New York 2012, Seite 4.

der Stockwerke oder Nummerierung der Wohnungen durch die Auskunftspflichtigen) zur Verfügung standen. Die zu viel generierten Wohnungen mussten später während der Datenaufbereitung in aufwendigen Prozeduren identifiziert und wieder entfernt werden (siehe Abschnitt 3.3.2).

Auch Wohnungen, die auf dem Fragebogen durchgestrichen waren, konnten bei der Beleglesung zu Problemen führen. Auf jedem Fragebogen gab es die Möglichkeit, Angaben für bis zu sechs Wohnungen vorzunehmen. Auskunftspflichtige, die für weniger als sechs Wohnungen im Gebäude berichteten, haben die überzähligen Wohnungen mitunter durchgestrichen. Wurden bei diesen Streichungen „Kästchen“ auf dem Fragebogen getroffen, so interpretierte das Beleglesegerät dies teilweise als Antwort und legte so eine weitere Wohnung im Gebäude an. Auch dieser Fehler musste im Verlauf der Datenaufbereitung wieder beseitigt werden (siehe Abschnitt 3.3.2).

Duplizierung von (unentdeckten) Fehlern durch Korrekturverfahren

Korrekturmethoden selbst sollen eigentlich keine Antwortausfälle oder weitere Fehler erzeugen.[12] Daher ist es wichtig, die Verfahren im Vorfeld intensiv zu testen, um zum Beispiel zu vermeiden, dass unerwartet Inkonsistenzen entstehen. Werden Fehler in den Daten nicht entdeckt, so kann es passieren, dass diese durch die Korrekturverfahren dupliziert werden oder dass bei der Berichtigung ungenaue Werte ermittelt werden. Dies hängt aber von den eingesetzten Verfahren zur Fehlerbeseitigung ab. In der Gebäude- und Wohnungszählung 2011 wurde ein Spenderverfahren verwendet, bei dem fehlerhafte Datensätze durch vorhandene fehlerfreie Datensätze korrigiert werden (siehe Abschnitt 3.3.3). Bei dieser Methode bestand die Gefahr, dass sich Unplausibilitäten vervielfachten, und vereinzelt ist dies auch eingetreten (siehe Abschnitt 4.1).

3 Imputationsverfahren in der Gebäude- und Wohnungszählung 2011

In den vorangegangenen Abschnitten ist deutlich geworden, wie unterschiedlich die Ursachen sind, die zu Item Nonresponse führen können. Angesichts dieser Vielfalt ist verständlich, dass das Erkennen und Korrigieren von Fehlern äußerst schwierig sein kann. Eine der wichtigsten Maßnahmen zur Verbesserung der Datenqualität sollte daher zunächst immer sein, Fehler und Antwortausfälle zu vermeiden. Denn auch gute maschinelle Verfahren zur Datenkorrektur sind nicht in der Lage, die „wirklichen Werte“ korrekt zu ermitteln. Sie können allerdings entscheidend dazu beitragen, die Probleme zu verringern, die durch Item Nonresponse entstehen.[13]

3.1 Was versteht man unter Imputation und warum ist sie sinnvoll?

Als Imputation bezeichnet man das Einsetzen von Werten in unvollständig beobachtete Daten, also die Korrektur von Antwortausfällen. Da es theoretisch auch denkbar ist, unplausible Werte zunächst zu löschen und anschließend mithilfe von statistischen Verfahren zu ergänzen, kann der Begriff der Imputation auch um die Korrektur unplausibler Werte erweitert werden.

Die Imputation von fehlerhaften und fehlenden Merkmalen ist aus mehreren Gründen wichtig. Zum einen hilft sie Informationsverluste zu vermeiden. Zum anderen ist sie notwendig, weil nicht alle Unplausibilitäten und Antwortausfälle neutral sind. Sie können auch systematisch auftreten.[14] Dies veranschaulichen die folgenden drei möglichen Fehlermechanismen, die in Daten vorhanden sein können. Von *Missing Completely At Random* (*MCAR*) spricht man, wenn ein Antwortausfall rein zufällig ist, also von keinem anderen Merkmal abhängt. Als *Missing At Random* (*MAR*) bezeichnet man Fälle von Item Nonresponse, die zwar von einem anderen vorhandenen Merkmal abhängen, aber unabhängig von dem untersuchten Merkmal selbst auftreten. *Missing Not At Random* (*MNAR*) bedeutet, dass das Auftreten von fehlenden oder fehlerhaften Werten sowohl von anderen Merkmalen als auch vom untersuchten Merkmal mit den Antwortausfällen selbst abhängen kann.[15]

Das Ignorieren von systematischen Fehlern und Datenausfällen (also Fällen von MAR und MNAR) kann Verzerrungen (Bias) in den Daten zur Folge haben. Das Nichtbehandeln von Item Nonresponse kann daher, abhängig von dem zugrunde liegenden Fehlermechanismus, zu verzerrten Ergebnissen und damit zu einem ungenauen Bild der Realität und zu falschen Schlussfolgerungen führen.[16]

3.2 Auswahl der Imputationsverfahren

Vor diesem Hintergrund musste entschieden werden, welche Imputationsmethoden in der Gebäude- und Wohnungszählung 2011 zur Korrektur von Unplausibilitäten und Antwortausfällen eingesetzt werden sollten. Diese Entscheidung wurde von einer Reihe von Rahmenbedingungen beeinflusst, die letzten Endes dazu führten, dass drei unterschiedliche Imputationsverfahren verwendet wurden.

Datenmenge

In der Zählung wurden rund 19 Millionen Gebäude erfasst. Bei dieser Datenmenge war frühzeitig klar, dass überwiegend maschinelle Verfahren für die Korrektur von fehlerhaften Datensätzen eingesetzt werden müssen. Manuelle Verfahren sollten daher ursprünglich so weit wie möglich auf einige wenige Fallkonstellationen beschränkt bleiben.

12 Siehe Messingschlager, M.: „Fehlende Werte in den Sozialwissenschaften – Analyse und Korrektur mit Beispielen aus dem ALLBUS“, Bamberg 2012, Seite 18.

13 Siehe Allison, P. D.: “Missing Data”, Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136, Thousand Oaks 2001, Seite 2 f.

14 Siehe Messingschlager, M. (Fußnote 12), hier: Seite 4.

15 Siehe Little, R. J. A./Rubin, D. B.: “Statistical analysis with missing data” (2. Auflage), New York 2002, Seite 11 ff.

16 Siehe Messingschlager, M. (Fußnote 12), hier: Seite 8.

Kontinuierliche und frühest mögliche Aufbereitung der Daten

Bei der Planung der Verfahren für die Datenaufbereitung mussten einige zeitliche Restriktionen beachtet werden, insbesondere weil das Zensusmodell nach Abschluss der eigentlichen Erhebungen noch weitere Schritte zur Qualitätssicherung[17] und Vervollständigung der Ergebnisse vorsah. Vor allem die Arbeiten, in die die Erhebungsstellen eingebunden waren, aber auch gegebenenfalls notwendige Rückfragen bei den Auskunftspflichtigen (siehe Abschnitt 3.3.2) sollten stichtagsnah erfolgen, um die Verhältnisse zum 9. Mai 2011 möglichst genau abbilden zu können. Aus diesen Gründen konnte mit dem Start der Datenaufbereitung nicht gewartet werden, bis alle Daten zu allen Gebäuden eingegangen waren, sondern die Fehlererkennung für jedes Gebäude sollte automatisch starten, sobald alle Angaben zu dem jeweiligen Gebäude vorhanden waren.

Skalierung und Menge der Merkmale

In der Gebäude- und Wohnungszählung 2011 wurden mit sechs Gebäude- und neun Wohnungsangaben relativ wenige Merkmale erhoben. Die Merkmale waren überwiegend nominal skaliert. Vier metrische Merkmale wurden erfasst (Anzahl der Wohnungen, Baujahr, Wohnfläche und Raumzahl).

Zeit und Aufwand für die Umsetzung

Um die neue Zensusmethode nicht mit der Entwicklung einer neuen Imputationsmethodik sowie der zugehörigen Programmierung der Software zu belasten, wurde untersucht, welche bereits bestehenden Software-Pakete in die Gebäude- und Wohnungszählung 2011 integriert und entsprechend angepasst werden könnten. Mangels entsprechender vergleichbarer Daten aus vorangegangenen Erhebungen war es nicht möglich, die Verfahren im Vorfeld mit einer ausreichenden Menge an echten Einzeldaten testen und anpassen zu können.

3.3 Verfahren zum Umgang mit Item Nonresponse

Die Imputationsverfahren mussten darüber hinaus in den Gesamtprozess der Datenaufbereitung integriert werden. Neben der Fehlerkorrektur (Imputation) ist auch die Fehlererkennung ein wichtiger Bestandteil der Datenaufbereitung. Bei der Fehlererkennung erfolgte die Prüfung der Daten aus der Gebäude- und Wohnungszählung auf ihre

- Vollzähligkeit: zu jedem Gebäude musste ein Datensatz vorhanden sein,
- Vollständigkeit: zu allen erforderlichen Merkmalen mussten Angaben vorhanden sein,
- Strukturplausibilität: Wertebereiche beziehungsweise Kodierungen von Merkmalen mussten korrekt sein,
- Interplausibilität: zwischen Merkmalen durften keine logischen Widersprüche bestehen und
- Plausibilität mit dem Melderegister: die Anzahl der Wohnungen im Gebäude wurde auf starke Abweichungen zur Anzahl der gemeldeten Personen geprüft.

Um die Daten unter diesen Gesichtspunkten zu prüfen, wurden insgesamt 109 Plausibilitätsregeln aufgestellt, die alle Datensätze erfüllen mussten. Bei solchen Regeln wird in der amtlichen Statistik zwischen Fehlern und Prüfhinweisen unterschieden. Fehler sind Unplausibilitäten von Merkmalen, die in jedem Fall korrigiert werden müssen. Von Prüfhinweisen spricht man, wenn der vorhandene Wert für ein Merkmal falsch sein könnte, dies aber erst geprüft werden muss, oder wenn bestimmte Ausprägungen oder Werte nur bei wenigen Datensätzen auftreten (Ausreißer). Aus den im Kapitel 2 gemachten Aussagen wird deutlich, dass bei der Gebäude- und Wohnungszählung 2011 häufig Fehler bei den Merkmalen „Eigentumsverhältnisse der Wohnung“ und „Anzahl der Wohnungen im Gebäude“ auftraten. Ein Beispiel für Prüfhinweise waren die möglichen Unplausibilitäten zwischen der Anzahl der Wohnungen in einem Gebäude und der Anzahl der gemeldeten Personen an einer Anschrift, die immer durch die Statistischen Landesämter geprüft werden mussten.

Zur Fehlerkorrektur wurden bei der Gebäude- und Wohnungszählung 2011 insgesamt drei Verfahren eingesetzt:

- deterministische Imputationen
- manuelle Prüfungen und Rückfragen bei den Auskunftspflichtigen
- Imputationen mit einem Hot-Deck-Verfahren nach dem Nearest-Neighbour-Prinzip (mit der kanadischen Software CANCEIS)

3.3.1 Deterministische Imputationen

Die deterministische Imputation von Merkmalen konnte immer dann angewendet werden, wenn eine eindeutige Beziehung zwischen dem unplausiblen beziehungsweise fehlenden Merkmal und einem oder mehreren plausiblen Merkmalen vorlag, wenn also eindeutig war, wie ein Merkmal korrigiert werden musste. Fehlte zum Beispiel in einem Datensatz die Angabe zum Gebäudetyp, aber der Eintrag „Bewohnte Unterkunft“ war beim Merkmal Art des Gebäudes vorhanden, so wurde das Merkmal Gebäudetyp auf „Anderer Gebäudetyp“ gesetzt.

Zu den Verfahren der deterministischen Imputation gehören auch Fixeinsetzungen. Ein Beispiel ist der Umgang mit fehlenden Angaben beim Merkmal „Wohnungstyp“ (also ob es sich um eine Ferien-/Freizeitwohnung beziehungsweise Diplomaten-/Streitkräftewohnung handelt). In diesen Fällen wurde immer „keines von beiden“ eingesetzt.

Allerdings war der Anteil der Fehler, die deterministisch korrigiert werden konnten, relativ gering, da die dafür notwendigen eindeutigen Beziehungen zwischen den erhobenen Merkmalen nur in wenigen Fällen vorlagen. Insbesondere

17 Etwa die Befragung zur Klärung von Unstimmigkeiten (laut § 16 Zensusgesetz 2011), bei der Anschriften mit einer bewohnten Wohnung in Gemeinden unter 10 000 Einwohnern bei Unstimmigkeiten durch die Erhebungsstellen überprüft wurden.

bei logischen Widersprüchen zwischen zwei Merkmalen konnte das fehlerhafte Merkmal nicht immer eindeutig identifiziert werden. Aus diesem Grund kamen für die Korrektur von Merkmalen noch weitere Verfahren zum Einsatz.

3.3.2 Manuelle Prüfungen und Rückfragen bei den Auskunftspflichtigen

Die manuelle Prüfung von Datensätzen sollte ursprünglich aufgrund der großen Datenmenge auf einige wenige Konstellationen beschränkt bleiben. So war in Fällen, in denen aus den Angaben der Auskunftspflichtigen hervorging, dass es sich um ein Gebäude ohne Wohnraum handelte, dort aber Personen gemeldet waren, eine Prüfung durch die Statistischen Landesämter vorgesehen. Sie kontrollierten diese Angaben mit den ihnen nach dem Zensusgesetz 2011 zur Verfügung stehenden Informationen, zum Beispiel durch Ansehen des Bemerkungsfeldes im Fragebogen oder indem die Anschrift von einem Erhebungsbeauftragten aufgesucht wurde. Einige Landesämter nahmen auch noch einmal Kontakt mit den Auskunftspflichtigen auf.

Darüber hinaus kam es zu Rückfragen oder Prüfungen, wenn die Anzahl der Wohnungen in einem Gebäude deutlich von der Anzahl der gemeldeten Personen abwich.[18] Unter anderem wurden so auch die Fehler behoben, die auftraten, wenn Auskunftspflichtige Probleme bei der Abgrenzung der Gebäude hatten oder die bei Fragebogen mit durchgestrichenen Wohnungen entstanden sind (siehe Abschnitte 2.1.1 und 2.1.2).

Hinzu kamen im Verlauf der Datenaufbereitung noch weitere manuelle Prüfungen und Korrekturen, die notwendig wurden, weil nicht alle Ursachen für Unplausibilitäten bereits bei der Vorbereitung der Erhebung bekannt waren. Einige Fehler zeigten sich erst bei der Prüfung der Daten, sodass „Ad-hoc-Korrekturen“ entwickelt werden mussten, die häufig auch manuelle Prüfungen umfassten.

So wurden zum Beispiel die im Abschnitt 2.1.2 beschriebenen Beleglesefehler („001“ als „601“ interpretiert) korrigiert, indem betroffene Gebäude zunächst über Algorithmen identifiziert wurden (Häufung der Wohnungszahlen 61, 81, 601, 801, ...). Anschließend wurden diese Gebäude (teilweise manuell) geprüft und fälschlicherweise imputierte Wohnungen gelöscht.

3.3.3 Imputation mit einem Hot-Deck-Verfahren nach dem Nearest-Neighbour-Prinzip

Als drittes Verfahren zur Korrektur von Item Nonresponse wurde mit der Software CANCEIS ein Hot-Deck-Verfahren eingesetzt, das nach dem Nearest-Neighbour-Prinzip arbeitet. Hierbei handelt es sich um ein von Statistics Canada entwickeltes Plausibilisierungs- und Imputationsprogramm, welches unter dem Betriebssystem Windows installiert werden kann. Dieses Verfahren wird seit 2001 im kanadischen Zensus verwendet.[19] Neben Italien, Brasilien, der Schweiz, Peru, Neuseeland und dem Vereinigten Königreich wurde es nun zum ersten Mal auch in der amtlichen Statistik in Deutschland eingesetzt und wird deshalb im Folgenden genauer vorgestellt.

3.3.3.1 Eigenschaften eines Nearest-Neighbour-Verfahrens

Bei einem Nearest-Neighbour-Verfahren wird die Menge der Datensätze in unplausible und plausible Datensätze unterteilt. Die plausiblen Datensätze werden in diesem Kontext als „Menge der möglichen Spender“ (oder kurz als Spender) bezeichnet. Die fehlerhaften Datensätze nennt man „Empfänger“. Grundgedanke des Verfahrens ist es, die Ausprägung eines Merkmals (oder die Ausprägungen mehrerer Merkmale) eines plausiblen Datensatzes in das entsprechende Merkmal (oder in die entsprechenden Merkmale) eines unplausiblen Datensatzes zu imputieren, sodass dieser anschließend plausibel ist. Aus Datenqualitätsgründen werden beim Nearest-Neighbour-Verfahren für einen konkreten unplausiblen Datensatz die plausiblen Datensätze ausgewählt, die sich am wenigsten von diesem unterscheiden (daher der Begriff „Nächster Nachbar“). Da in der Regel mehrere „Nächste Nachbarn“ als Spenderdatensätze gefunden werden, wird aus diesen geeigneten Datensätzen einer zufällig gezogen. Dieser Datensatz „spendet“ dem unplausiblen Datensatz anschließend Merkmalsausprägungen. Wegen des zufälligen Ziehens aus dem Datenbestand der gleichen Erhebung gehört dieses in CANCEIS implementierte Verfahren zur Gruppe der sogenannten Hot-Deck-Verfahren.[20]

Erforderliche Eigenschaften im Hinblick auf die Datenqualität

Im Hinblick auf die Qualität der Ergebnisse der Plausibilisierung sollte der Imputationsalgorithmus bestimmte Eigenschaften aufweisen.[21] Diese werden im Folgenden aufgeführt und begründet:

1) Ziel eines nach dem Fellegi-Holt Prinzip[22] arbeitenden Imputationsalgorithmus ist es, möglichst wenige Merkmalsausprägungen innerhalb eines fehlerhaften Datensatzes zu verändern. Dabei wird unterstellt, dass ein Auskunftspflichtiger eher wenige Fehler macht anstelle von vielen.

2) Abweichend von diesem Grundprinzip kann es bei der Imputation manchmal von Vorteil sein, wenn mehr Merkmalsausprägungen als minimal möglich geändert werden. So zum Beispiel ein Datensatz, der sowohl durch die Imputation von nur einem Wert als auch durch

18 So durfte das Verhältnis der Anzahl der gemeldeten Personen zu den Wohnungen 6 (bei einer Wohnung im Gebäude) beziehungsweise 8 (bei mehr als einer Wohnung im Gebäude) nicht überschreiten, beziehungsweise umgekehrt das Verhältnis der bewohnten Wohnungen zur Anzahl der gemeldeten Personen nicht größer/gleich 2 sein (bei Gebäuden mit mehr als 3 bewohnten Wohnungen).

19 Siehe hierzu Bankier, M.: “Evolution of Canadian Census E&I Systems – 1976 to 2011”, Working Paper 22, Konferenz Europäischer Statistiker 2009.

20 Kalton, G./Kasprzyk, D.: “Imputing for Missing Survey Responses” in Proceedings of the Survey Research Methods Section, American Statistical Association, Washington D. C. 1982, Seite 22 ff.

21 Siehe hierzu Bankier, M./Poirier, P./Lachance, M./Mason, P.: “A generic implementation of the nearest-neighbour imputation methodology (NIM)” in Proceedings of the Second International Conference on Establishment Surveys, Buffalo 2000, Seite 571 ff.

22 Fellegi, I. P./Holt, D.: “A systematic approach to automatic edit and imputation” in Journal of the American Statistical Association, Jahrgang 71, Seite 17 ff.

die von zwei anderen korrigiert werden kann. Wenn im ersten Fall die Merkmalsausprägung stark verändert würde (zum Beispiel das Baujahr um viele Jahre), im zweiten Fall allerdings nur geringfügige Änderungen bei zwei anderen Merkmalen vollzogen werden müssten, so sollten beide Imputationsaktionen[23] zur Auswahl stehen und eine der beiden zufällig ausgewählt und durchgeführt werden.

3) Eine weitere erstrebenswerte Eigenschaft ist, dass lediglich ein Datensatz als Spender zur Imputation eines fehlerhaften Datensatzes herangezogen wird. Dadurch soll garantiert werden, dass die Imputation nicht nur formal nach den Plausibilisierungsregeln gültige, sondern auch realistische Datensätze erzeugt.

4) Zudem sollte sichergestellt sein, dass ähnliche oder gleich gute Imputationsaktionen, basierend auf den unterschiedlichen vorhandenen möglichen Spenderdatensätzen, auch eine ähnliche beziehungsweise gleiche Wahrscheinlichkeit besitzen, als die auszuführende Imputationsaktion ausgewählt zu werden. So wird vermieden, dass bestimmte Merkmalsausprägungen unverhältnismäßig stark vervielfältigt werden.

5) Ebenfalls ein wichtiger Punkt ist, dass ein Spender nicht zu oft genutzt wird, da dieser sonst einen unangemessenen Einfluss auf die imputierten Daten haben kann. Das könnte zu einer Verzerrung der Verteilung führen.

6) Zusätzlich sollte beachtet werden, dass nur Datensätze als Spender herangezogen werden, die nicht zuvor schon imputiert wurden. Auch dies könnte zu einer Verzerrung der Verteilung führen.

3.3.3.2 Die Imputationssoftware CANCEIS

Komponenten des Programms

Ein wichtiger Aspekt bei der Nutzung eines maschinellen Imputationsverfahrens im Rahmen der Gebäude- und Wohnungszählung 2011 war, dass durch das Verfahren nicht nur fehlende Werte imputiert, sondern auch Unplausibilitäten und Inkonsistenzen zwischen Merkmalen erkannt und korrigiert werden sollten. Zudem musste bei der Imputation selbst darauf geachtet werden, dass durch den jeweils eingesetzten Wert keine weiteren Unplausibilitäten entstehen. Hierzu werden CANCEIS die einzuhaltenden Plausibilitätsregeln in Form sogenannter „*Decision Logic Tables*“ (*DLTs*) übergeben. Mittels des „*DLT Analyzer*“ erkennt die Software redundante Plausibilitätsregeln sowie doppelt gestellte Bedingungen und beseitigt sie, damit später nicht unnötigerweise wiederholt die gleiche Bedingung abgefragt wird. Abschließend werden in diesem Schritt alle nicht redundanten Plausibilitätsregeln in einer gemeinsamen DLT-Datei kombiniert, anhand derer mögliche Imputationen simultan darauf geprüft werden, ob sie zu zulässigen Ergebnissen führen.

Im Anschluss wird durch den als „*Imputation Engine*“ bezeichneten eigentlichen Kern des Software-Pakets jeder Datensatz zunächst auf fehlende oder ungültige Werte geprüft. Hierfür müssen vorab Gültigkeitsbereiche definiert werden, die die zulässigen Werte umfassen. Wird hierbei kein Fehler im entsprechenden Datensatz gefunden, wird anhand der restlichen Plausibilitätsregeln aus den DLTs auf unzulässige Kombinationen von Werten geprüft.

Dieses Vorgehen dient der Unterscheidung der Datensätze in „Spender“ und „Empfänger“. Sofern nicht anders spezifiziert, stoppt die Prüfung eines Datensatzes, sobald durch eine Regel ein Fehler identifiziert wurde. An dieser Stelle ist bereits bekannt, dass dieser Datensatz fehlerhaft ist und nicht mehr als Spender infrage kommt. Die verbleibenden Plausibilitätsregeln werden dann nicht mehr zur Prüfung des Datensatzes verwendet und es wird mit dem nächsten Datensatz fortgefahren.

Abschließend werden fehlende, ungültige und unplausible Werte in den Empfängerdatensätzen durch Werte aus den Spenderdatensätzen ersetzt. Hierbei werden die zuvor beschriebenen erforderlichen Eigenschaften im Hinblick auf die Datenqualität (möglichst wenige Merkmale imputieren, lediglich ein Spender je Datensatz, ähnliche Wahrscheinlichkeiten für ähnlich gute Imputationsaktionen) umgesetzt. Um dabei den ersten beiden Eigenschaften gerecht zu werden, kombiniert CANCEIS diese Ziele und stellt im Vergleich zum Fellegi-Holt-Prinzip die Reihenfolge der beiden Schritte um. Während Fellegi und Holt zunächst die minimale Anzahl an zu ändernden Merkmalen bestimmen und dann passende Spender suchen, um diese Merkmale zu imputieren, ist CANCEIS so programmiert, dass zunächst die „Nächsten Nachbarn“ bestimmt werden und auf Grundlage dieser dann entschieden wird, welche Merkmale geändert werden müssen. Die Suche nach den „Nächsten Nachbarn“ geschieht mittels eines sogenannten Ripple-Search-Verfahrens[24] und durch den Einsatz von Distanzfunktionen.

Distanzberechnung in CANCEIS

In der Software CANCEIS sind Distanzfunktionen implementiert, um entscheiden zu können, wann sich zwei Datensätze ähnlich sind. Hierbei wird jedes einzelne Merkmal zwischen zwei Datensätzen verglichen. Für nominal skalierte Variablen ist zum Beispiel eine 0/1-Funktion programmiert, welche die Anzahl dieser nicht übereinstimmenden qualitativen Variablen beim Vergleich zweier Datensätze zählt. Eine 0/1-Funktion bewertet übereinstimmende Merkmalsausprägungen mit einem Distanzwert von 0 und unterschiedliche Merkmalsausprägungen mit einem Wert von 1.

Liegt eine metrisch skalierte Variable vor, so ist der Betrag der Differenz beider Werte ein guter Indikator für die Entfernung. Die für solche Variablen verwendete Distanzfunk-

23 Werden Daten aus einem Spendersatz in einen fehlerhaften Datensatz eingesetzt, so wird dies als Imputationsaktion bezeichnet. Dabei können einzelne oder alle Werte des fehlerhaften Datensatzes durch Werte des Spenderdatensatzes ersetzt werden. Eine Imputationsaktion wird zulässig genannt, wenn sie garantiert, dass ein ursprünglich fehlerhafter Datensatz anschließend keine Werte mehr aufweist, die durch die Plausibilitätsregeln als Fehler identifiziert werden.

24 Hierbei werden abwechselnd die direkt vor und nach einem fehlerhaften Datensatz gespeicherten plausiblen Datensätze auf deren Eignung als Spender untersucht. Das Ganze läuft in mehreren Suchschichten ab. Eine Vorabsortierung der Datensätze ist sinnvoll, geht man davon aus, dass Gebäude und Wohnungen aus der Nachbarschaft ähnliche Merkmale aufweisen.

tion teilt beispielsweise einem Baujahr, das nahe an einem anderen liegt, einen kleinen Distanzwert (nahe bei 0) zu und weit auseinander liegenden Werten einen entsprechend höheren. Anschließend wird das Ergebnis normiert, sodass auch hier die Skala zwischen 0 und 1 liegt und die Distanzen die gleiche Größenordnung besitzen wie die von nominalskalierten Variablen. Für jede Variable kann individuell eine andere der in CANCEIS implementierten Distanzfunktionen gewählt werden. Die Einzeldistanzen werden zu einer Gesamtdistanz summiert.

Nach den erforderlichen Eigenschaften im Hinblick auf die Ergebnisqualität (siehe Abschnitt 3.3.3.1) ist eine Minimierung der Gesamtdistanz zwischen einem fehlerhaften Datensatz (V_f) und einem möglichen Spender (V_p) von Interesse (Suche nach „Nächstem Nachbarn“). Diese wird als gewichtete Summe der Einzeldistanzen wie folgt definiert:

(1) $D_{fp} = D(V_f, V_p) = \sum_{(i=1\ldots I)} w_i D_i(V_{fi}, V_{pi})$,

wobei D_i die Distanzfunktion der i-ten Variable und w_i die jeweilige Gewichtung darstellt. Dabei ist berücksichtigt, dass für jede Variable eines Datensatzes eine andere Distanzfunktion gewählt werden kann. Die Gewichte können je nach Wichtigkeit der Übereinstimmung hoch oder niedrig gehalten werden.

Die Gesamtdistanz zwischen fehlerhaftem Datensatz und möglichem Spender lässt sich als Summe der Distanz des fehlerhaften zum imputierten Datensatz (V_a) und der Distanz des imputierten Datensatzes zum Spenderdatensatz darstellen: $D_{fp} = D_{fa} + D_{ap}$. Mit einer entsprechenden Gewichtung α aus dem Bereich (0.5, 1] – je nachdem, ob mehr Wert auf die minimale Anzahl an imputierten Merkmalen (α nahe 1) oder auf realistischere imputierte Datensätze (α nahe 0.5) gelegt wird – ergibt sich folgende Formel:

(2) $D_{fpa} = \alpha D_{fa} + (1 - \alpha) D_{ap}$

Diese Gleichung wird in Betracht gezogen, wenn die zulässigen Imputationsaktionen der „Nächsten Nachbarn“ bestimmt werden. So genügen Datensätze, die einen kleinen Wert für D_{fpa} aufweisen, den ersten beiden erstrebenswerten Eigenschaften aus Abschnitt 3.3.3.1. Daraus lässt sich folgern, dass im Suchalgorithmus nur Imputationsaktionen gespeichert werden, die zulässig sind und der Gleichung

(3) $D_{fpa} \leq \gamma min(D_{fpa})$

genügen, wobei $\gamma \geq 1$ gewählt werden kann. Ein $\gamma > 1$ besagt, dass eine Imputationsaktion fast genauso gut ist wie das bisher ermittelte Minimum. Allerdings sollte γ dabei natürlich nicht zu groß gewählt werden (zum Beispiel bedeutet $\gamma = 1.1$, dass D_{fpa} um bis zu 10 % größer als das bisherige Minimum sein darf). Ebenso wie $min(D_{fpa})$ wird γ im Ablauf des Programms immer wieder aktualisiert. Durch diese Aktualisierungen lässt sich gegebenenfalls früher entscheiden, ob eine Imputationsaktion zu verwerfen ist oder nicht. Der Wert von γ wird dabei so angepasst, dass gilt: $\gamma min(D_{fpa}) = max(D_{fpa})$, wobei $max(D_{fpa})$ der oberen Schranke bei einer gefüllten Liste von Imputationsaktionen entspricht. Insgesamt wird stets nur eine fest vorgegebene Anzahl n (zum Beispiel $n = 10$) an besten Imputationsaktionen gespeichert.

Sind schließlich bestimmte Abbruchkriterien erfüllt, wird zu allen Imputationsaktionen auf der Liste der zulässigen Imputationsaktionen ein Größenmaß berechnet:

(4) $R_{fpa} = (min(D_{fpa}) / D_{fpa})^t$,

wobei t in der Regel um den Wert 1 herum gewählt werden sollte. Zunächst besitzt jede potenzielle Imputationsaktion die gleiche Wahrscheinlichkeit, gezogen zu werden. Diese wird mit dem Faktor R_{fpa} multipliziert und anschließend wird eine Imputationsaktion mit einer Wahrscheinlichkeit proportional zu dem sich ergebenden Wert gezogen.[25]

Einlese- und Output-Formate

Die Output-Dateien von CANCEIS sind sehr zahlreich und geben eine Vielzahl der berechneten Werte und Statistiken wider. So kann zum Beispiel dokumentiert werden, welcher Datensatz durch welchen Spender imputiert wurde, welche Distanz dabei zwischen Empfänger und Spender auftrat, welche Merkmale geändert wurden und aus welchem Grund, sowie welche alternativen „Nächsten Nachbarn“ es gegeben hätte. In der Gebäude- und Wohnungszählung 2011 wurden ausgewählte Kennzeichen zur späteren Einschätzung der Qualität der Imputationen dauerhaft abgespeichert.

Des Weiteren wird aufgelistet, welche Datensätze nicht durch CANCEIS imputiert werden konnten und eventuell noch einmal nachgeprüft werden müssen, was allerdings bei der Gebäude- und Wohnungszählung 2011 nicht vorgekommen ist. Anhand von Fehlerdateien (Error Files) lässt sich darüber hinaus schnell erkennen, ob Fehler oder sonstige Unstimmigkeiten aufgetreten sind. Diese werden dann in den Log-Files näher spezifiziert. Daneben gibt es noch zusätzliche Dateien, die angeben, wie viele Spender in Betracht gezogen wurden (somit lässt sich zum Beispiel die geografische Nähe von Spender/fehlerhaftem Datensatz nachvollziehen) und welche Merkmale wie oft imputiert werden mussten. Alle Ausgabedateien werden in „flachem“ Dateiformat (als .txt-Dateien) abgespeichert. Ebenso müssen alle einzulesenden Dateien im .txt-Format vorliegen. Diese Schnittstelle galt es, aus dem bestehenden Aufbereitungssystem der Gebäude- und Wohnungszählung zu bedienen.

3.3.3.3 Einbindung von CANCEIS in die Gebäude- und Wohnungszählung

Für die Integration von CANCEIS in die Datenaufbereitung der Gebäude- und Wohnungszählung mussten zwei grundsätzliche Anforderungen des Programms berücksichtigt werden:

25 Zur Verdeutlichung des Faktors R_{fpa} ein Beispiel: Es wurden drei zulässige Imputationsaktionen zur Plausibilisierung eines Datensatzes gefunden. Dabei seien die Distanzwerte zu den drei Imputationsaktionen: $D_{fpa1} = 4$, $D_{fpa2} = 6$ und $D_{fpa3} = 12$. Daraus ergeben sich mit Gleichung (4) und $t = 1$ folgende Werte: $R_{fpa1} = 1$, $R_{fpa2} = 2/3$ und $R_{fpa3} = 1/3$. Somit folgt dann, dass mit einer Wahrscheinlichkeit von 50 % die erste, zu rund 33,3 % die zweite und in 16,7 % der Fälle die dritte Imputationsaktion gewählt wird. Diese Wahrscheinlichkeiten verhalten sich proportional zu den berechneten R_{fpa}-Werten.

1) CANCEIS wurde in Kanada ursprünglich für die Korrektur von Haushaltsdatensätzen konzipiert. Die Software betrachtet daher einen kompletten Haushalt als eine Einheit und ist so programmiert, dass nur Datensätze gleicher Länge (also mit gleicher Anzahl an Merkmalen) miteinander verglichen und imputiert werden können. Da Gebäudedatensätze je nach Wohnungszahl unterschiedliche Längen aufweisen, hatte dies für die Gebäude- und Wohnungszählung zur Folge, dass die erhobenen Gebäude entsprechend dem Merkmal „Anzahl der Wohnungen“ aufgeteilt und getrennt imputiert werden mussten. Dafür wurden Datensätze mit jeweils identischer Wohnungszahl in einer Datei abgespeichert.

2) Um Datensätze nicht nur erfolgreich, sondern auch mit guter Qualität imputieren zu können, muss in jeder Datei, die durch CANCEIS bearbeitet wird, eine ausreichend große Menge an plausiblen Datensätzen (potenziellen Spendern) vorhanden sein. Aufgrund von Erfahrungen aus Kanada wurde für die Gebäude- und Wohnungszählung in Deutschland festgelegt, dass in jeder Datei der Anteil der Spenderdatensätze mehr als 50 % betragen muss. Außerdem mussten mindestens hundert plausible Datensätze vorhanden sein.

Vorbereitung der Datensätze

Nach Möglichkeit sollte die Imputation von Datensätzen mit CANCEIS auf Gemeindeebene erfolgen. Daher wurden in einem ersten Schritt für jede Gemeinde einzelne Dateien gebildet, die jeweils Gebäude mit der gleichen Wohnungszahl enthielten. Da bestimmte Gebäudegrößen (etwa Gebäude mit genau 123 Wohnungen) nicht besonders häufig auftreten, war schon bei der Konzeption des Verfahrens klar, dass für bestimmte Gebäudegrößenklassen nicht genügend potenzielle Spender in den Dateien vorhanden sein würden (siehe Anforderung 2). Aus der Gebäude- und Wohnungsstichprobe von 1993 war bekannt, dass 99 % der Gebäude maximal 14 Wohnungen umfassen. Daher wurden separate Dateien nur für Gebäude bis maximal 14 Wohnungen gebildet, also 14 Dateien pro Gemeinde. Größere Gebäude (mit einer „seltenen“ Wohnungszahl) wurden mit einer abweichenden Methodik behandelt.

Nach der Zuordnung der Gebäude wurde geprüft, ob in jeder der Dateien genügend potenzielle Spender für eine Imputation vorhanden waren. Wenn dies der Fall war, konnte die Imputation mit CANCEIS gestartet werden. Anderenfalls wurden Dateien mit gleicher Wohnungszahl aus unterschiedlichen Gemeinden so lange zusammengefasst, bis die Bedingungen erfüllt waren.

Vor dem Start der Imputation wurden die Datensätze in den einzelnen Dateien nach Gemeinde, Ortsteil, Straße und Hausnummer sortiert. Es wurde unterstellt, dass Gebäude aus direkter geografischer Nachbarschaft wahrscheinlich eine Ähnlichkeit in der Bauart aufweisen. Dieses Vorgehen diente dazu, die Laufzeit von CANCEIS zu optimieren, weil somit schneller Spender mit niedrigen Distanzen gefunden werden konnten. Eine Obergrenze für die Anzahl an Datensätzen, die in CANCEIS eingelesen werden können, besteht nicht, sodass auch die Daten aus großen Gemeinden wie Hamburg oder Berlin nicht aufgeteilt werden mussten.

Behandlung von Gebäuden mit fehlendem Gebäudemerkmal „Anzahl der Wohnungen“

Voraussetzung für die beschriebene Vorgehensweise war, dass das Merkmal „Anzahl der Wohnungen“ vorhanden war. Fehlte aber gerade dieses Merkmal[26], so musste zunächst die Anzahl der Wohnungen mit CANCEIS imputiert werden.

Zu diesem Zweck wurden von allen Datensätzen in den betroffenen Gemeinden ausschließlich die Gebäudeangaben in einer Datei abgespeichert. Anschließend wurde geprüft, ob die Datei die notwendige Menge plausibler Datensätzen enthielt und ob das Verhältnis von (potenziellen) Spendern zu Empfängern stimmte. War dies nicht der Fall, mussten wieder Zusammenfassungen erfolgen. Für den folgenden Durchlauf von CANCEIS wurden spezielle Regeln für die Plausibilisierung verwendet, die nur das fehlende Gebäudemerkmal „Anzahl der Wohnungen“ als Fehler definierten. Weitere unplausible oder fehlende Gebäudemerkmale wurden zu diesem Zeitpunkt nicht imputiert. Danach wurde in den Datensätzen, in denen das Gebäudemerkmal „Anzahl der Wohnungen“ eingesetzt worden war, die entsprechende Anzahl an Wohnungsdatensätzen ohne Angaben angelegt. Im Anschluss konnten die übrigen fehlenden und unplausiblen Merkmale – wie eingangs beschrieben – imputiert werden.

Imputation von Gebäuden mit einer „seltenen“ Anzahl an Wohnungen

Gebäude mit mehr als 14 Wohnungen sowie Dateien, in denen auch durch Zusammenfassungen nicht genügend potenzielle Spender vorhanden waren, mussten gesondert behandelt werden.

In diesen Fällen wurde das Gebäude in die einzelnen Wohnungen „zerlegt“ und fortan die Wohnung als eigene Einheit betrachtet. Durch das Zerlegen erhöhte sich nicht nur die Anzahl an Datensätzen, sondern auch der Anteil der plausiblen Datensätze. Zuvor war ein Datensatz bereits unplausibel und somit Empfänger, sobald in nur einer Wohnung des Gebäudes eine Unplausibilität auftrat. Nach der Aufteilung des Gebäudes in einzelne Wohnungsdatensätze war lediglich derjenige Datensatz unplausibel, der den Fehler enthielt, während alle anderen Wohnungen als Spender fungieren konnten.

Bei diesem Vorgehen musste beachtet werden, dass nach Imputation der Wohnungsdatensätze die zusammengehörenden Wohnungen wieder zu einem gemeinsamen Gebäude zusammengefügt werden mussten. Dies funktioniert jedoch nur, wenn fehlerhafte Gebäudeangaben nicht für jede Wohnung unterschiedlich verändert wurden, weil sich ansonsten neue Unplausibilitäten ergäben. Da es

26 Fehlte das Merkmal „Anzahl der Wohnungen“, wurde zunächst versucht, dieses deterministisch anhand der vorhandenen Wohnungen zu imputieren. Dies war allerdings nicht möglich, wenn zu dem Gebäude keine Wohnungsdatensätze vorlagen.

aber für die Gebäude- und Wohnungszählung 2011 Plausibilitätsregeln gibt, die eine Kombination von bestimmten Ausprägungen der Gebäude- mit bestimmten Wohnungsmerkmalen verbietet, können Gebäude und Wohnungen nicht vollständig voneinander getrennt behandelt werden. Deshalb wurden zunächst die Angaben zum Gebäude mit CANCEIS geprüft und gegebenenfalls imputiert. Im Anschluss spielte man diese Angaben an jeden einzelnen zugehörigen Wohnungsdatensatz und setzte den Status für die Gebäudemerkmale auf „nicht imputierbar“. Somit konnten in einem weiteren CANCEIS-Durchlauf bei diesen Gebäuden nur noch die Wohnungsmerkmale verändert werden und die Datensätze ließen sich im Anschluss ohne Komplikationen zu einem plausiblen Gebäude zusammenfügen. Durch eine Erhöhung der Gewichte für die Gebäudeangaben stieg gemäß Gleichung (1) die Wahrscheinlichkeit, dass der Spender für einen unplausiblen Wohnungsdatensatz aus dem gleichen Gebäude kam.

In Abschnitt 3.2 wurde bereits darauf eingegangen, dass bei der Planung der Verfahren für die Datenaufbereitung einige zeitliche Restriktionen beachtet werden mussten. Hierzu gehörten Schritte zur Vervollständigung und Qualitätssicherung der Daten, in die Auskunftspflichtige beziehungsweise Erhebungsstellen eingebunden waren. Da mit einem vollständigen Abschluss der Erhebungsphase erst zehn bis zwölf Monate nach dem Stichtag 9. Mai 2011 gerechnet wurde, Rückfragen beziehungsweise Begehungen aber möglichst stichtagsnah erfolgen sollten, konnte mit der Imputation der Datensätze nicht gewartet werden, bis wirklich alle erwarteten Angaben eingegangen waren. Aus diesem Grund wurden insgesamt zwei CANCEIS-Läufe durchgeführt. Der erste CANCEIS-Lauf startete, nachdem ein Großteil der Datensätze eingegangen war. Der zweite Lauf wurde nach dem vollständigen Abschluss der Erhebungsphase durchgeführt.[27]

4 Bewertung der eingesetzten Imputationsverfahren

Abschließend gilt es die Frage zu beantworten, wie sich die eingesetzten Imputationsverfahren bewährt haben, welche (unerwarteten) Probleme aufgetreten sind und welche Schlussfolgerungen sich daraus für die Entwicklung und den Einsatz von Imputationsmethoden in künftigen Gebäude- und Wohnungszählungen ziehen lassen.

4.1 Grenzen und Probleme beim Einsatz der Imputationsverfahren

Als größte Herausforderung erwies sich die Tatsache, dass bei der Konzeption der Plausibilitätsprüfungen nicht alle Fehler bedacht werden konnten, die während der Erhebung tatsächlich aufgetreten sind. Insbesondere die kontinuierliche Aufbereitung der Daten und die eingeschränkten Möglichkeiten für Tests mithilfe von Echtdaten (siehe Abschnitt 3.2) waren von Nachteil. Einige Fehler wurden erst zu einem Zeitpunkt erkannt, als die Plausibilisierung und Korrektur der Daten bereits weit fortgeschritten und die Imputation der Datensätze mit CANCEIS bereits abgeschlossen war. Dies betraf vor allem die eingangs beschriebenen Beleglesefehler zur Anzahl der Wohnungen im Gebäude, aber auch kleinere Fehler – die nur in geringem Umfang aufgetreten sind – wie Wohnheime, die mindestens eine bestimmte Anzahl an Wohnungen aufweisen mussten. Aus diesem Grund war es nötig, insgesamt zwölf nachträgliche Korrekturen zu entwickeln. Nach der Veröffentlichung erster Ergebnisse im Mai 2013 wurden noch einmal vier weitere Korrekturen umgesetzt.

Diese späte Identifikation zusätzlicher Fehler blieb nicht folgenlos. So mussten zunächst Prozeduren entwickelt werden, mit deren Hilfe sämtliche noch fehlerhaften Datensätze identifiziert und korrigiert werden konnten. Aufgrund der großen Komplexität der Beleglesefehlerproblematik konnte dabei nicht vermieden werden, dass – anders als ursprünglich vorgesehen – umfangreiche manuelle Arbeiten durch die Statistischen Landesämter geleistet werden mussten (siehe Abschnitte 2.1.2 und 3.3.2). Daneben verursachte dies Schwierigkeiten in Bezug auf die Datenqualität, da die Imputationsverfahren ja bereits abgeschlossen waren. Durch die neuen Probleme wurden jetzt teilweise Datensätze als fehlerhaft identifiziert, die ursprünglich als fehlerfrei angesehen worden waren und deshalb unter Umständen bereits als Spender in CANCEIS fungiert hatten. In solchen Fällen wurden auch die entsprechenden Empfängerdatensätze korrigiert. Trotzdem bleibt die Problematik bestehen, dass es nicht unwahrscheinlich ist, dass CANCEIS – bei einer rechtzeitigen Identifikation aller Fehler – aufgrund anderer Distanzen gegebenenfalls auch andere Spenderdatensätze als „Nächste Nachbarn“ identifiziert und damit vielleicht auch andere Werte für Merkmale eingesetzt hätte. In welchem Umfang das der Fall gewesen wäre, kann im Nachhinein nicht gesagt werden. Allerdings ist der Anteil der Gebäude, die von nachträglichen Korrekturen betroffen waren, mit 9 % nicht sehr hoch. Ein Fünftel dieser Gebäude waren ursprünglich Spender.

Aufgrund verschiedener Gegebenheiten waren die Statistischen Landesämter in unterschiedlichem Umfang von den nachträglich identifizierten Fehlern betroffen. Jedoch traten einige Probleme nicht in allen Ländern auf. Aus diesem Grund entschied jedes Bundesland separat, welche Korrekturen in seinem Datenbestand umgesetzt werden sollten und welche nicht.

4.2 Menge der Korrekturen

Das Ziel von Imputationsverfahren ist es, fehlende und unplausible Werte zu ersetzen und dabei eine Verzerrung der Daten zu verhindern, die durch Fehler und Antwortausfälle entstehen kann. Imputationsverfahren können und sollen aber nicht anschriftenscharf die Wirklichkeit reproduzieren. Dies bedeutet, dass es in Einzelfällen immer Abweichungen zur Realität geben wird. Dennoch tragen Imputationsverfahren dazu bei, die Datenqualität zu verbessern. Im Folgenden sollen einige Zahlen die Datenqualität der Gebäude- und Wohnungszählung 2011 näher beleuchten.

27 In diesem zweiten CANCEIS-Lauf wurden nur die unplausiblen Datensätze imputiert, die erst nach dem ersten Lauf eingegangen waren. Datensätze, die im ersten CANCEIS-Lauf Empfänger waren, wurden aus dem zweiten CANCEIS-Lauf ausgeschlossen, da sie sonst von CANCEIS als potenzielle Spender identifiziert und gegebenenfalls zur Imputation verwendet worden wären.

Tabelle 1 Imputationsraten je Gebäudemerkmal
Prozent

	Art des Gebäudes	Anzahl der Wohnungen	Baujahr	Gebäudetyp	Eigentums-verhältnisse des Gebäudes	Heizungsart
Anteil der Gebäude						
ohne Korrekturen[1]	95,9	92,2	97,1	98,4	81,5	93,5
mit deterministischer Imputation	1,6	7,4	0,0	0,0	10,4	0,0
mit CANCEIS-Imputation	2,5	0,4	2,9	1,6	8,1	6,5

1 Enthält vollständig imputierte und durch Erhebungsstellen begangene Gebäude (Unit-Nonresponse-Fälle).

Die Tabellen 1 und 2 zeigen die Imputationsraten je Merkmal.[28] Der Anteil der Gebäude, bei denen Korrekturen notwendig waren, schwankt je nach Merkmal zwischen 1,6 % und 18,5 %.

Die Imputationsraten der Wohnungsmerkmale fallen deutlich höher aus. Hierbei muss jedoch berücksichtigt werden, dass der Anteil der Wohnungen, in denen alle Merkmale mit CANCEIS imputiert wurden, bei 10 % liegt. Diese Wohnungen sind in die Auswertung mit eingeschlossen. Wie in Abschnitt 2.1.2 beschrieben, wurden ganze Wohnungen imputiert, wenn das Merkmal „Anzahl der Wohnungen" Rückschlüsse auf mehr Wohnungen im Gebäude zuließ als tatsächlich übermittelt worden waren. Das Merkmal „Eigentumsverhältnisse der Wohnung" weist mit einer Imputationsrate von rund 45 % einen sehr hohen Wert auf. Wie bereits in Abschnitt 2.1.1 erläutert, handelt es sich um ein Merkmal, das nur bei Eigentumswohnungen ausgefüllt werden musste. Eine Reihe von Auskunftspflichtigen ging aber davon aus, dass auch bei anderen Wohnungen, etwa bei Eigenheimen, eine Angabe notwendig war.

Insgesamt wurden in der Gebäude- und Wohnungszählung 2011 einzelne Merkmale bei rund 51 % der Gebäude deterministisch und bei rund 30 % der Gebäude mithilfe von CANCEIS imputiert. Darunter sind allerdings auch Gebäude, die mithilfe beider Imputationsverfahren korrigiert wurden. Schließt man die Gebäude, bei denen nur das Merkmal „Eigentumsverhältnisse der Wohnung" unplausibel war, aus der Auswertung aus, so sinkt der Anteil der Gebäude mit deterministischen Imputationen auf rund 25 % und der Anteil der Gebäude mit durch CANCEIS imputierten Merkmalen auf etwa 26 %.

28 Sämtliche Auswertungen wurden auf den Daten, die im Mai 2013 veröffentlicht wurden, vorgenommen. Die nachträglichen Anpassungen, die durch die Bildung von Haushalten notwendig waren, sind daher nicht berücksichtigt.

Manuelle Korrekturen lassen sich nur schwer beziffern. Da bei der Konzeption der Verfahren manuelle Arbeiten nur in Ausnahmefällen vorgesehen waren, wurde eine Dokumentation mit Qualitätskennzeichen nicht in ausreichendem Maß integriert. Insgesamt kann davon ausgegangen werden, dass etwa 1 % der Gebäude ausschließlich manuelle Korrekturen aufweisen, also in den genannten Auszählungen nicht enthalten sind.

Um die Vollständigkeit und Qualität der Daten, wie sie von den Auskunftspflichtigen übermittelt wurde, einschätzen zu können, ist es sinnvoll, nicht nur Imputationsraten der einzelnen Merkmale zu betrachten. Vielmehr sollte man auch auswerten, wie viel Prozent der Datensätze von Anfang an plausibel waren, also ohne Korrekturen und Imputationen ausgekommen sind. Hier liefert die Gebäude- und Wohnungszählung 2011 als schriftliche Befragung folgendes Bild: Insgesamt waren Daten über 31 % der Gebäude vollständig fehlerfrei.[29] In diesem Ergebnis ist nicht berücksichtigt, welche Merkmale jeweils Fehler aufwiesen. So wiegt beispielsweise eine fehlende Angabe zum Vorhandensein eines WCs in der Wohnung weniger schwer als etwa eine unplausible Wohnfläche. Schließt man wieder die Gebäude aus, bei denen nur das Merkmal „Eigentumsverhältnisse der Wohnung" unplausibel war, so waren die Daten über rund 57 % der Gebäude von Anfang an plausibel.

Darüber hinaus lässt sich ein Zusammenhang feststellen zwischen notwendigen Korrekturen und dem Erhebungsweg, auf dem die Daten eingegangen sind. Insgesamt füllten

29 Neben den Fällen von Item Nonresponse gab es auch Gebäude, zu denen ursprünglich keine Daten eingegangen sind, entweder weil keine Auskunftspflichtige/kein Auskunftspflichtiger recherchiert werden konnte oder weil keine Angaben übermittelt wurden (Unit Nonresponse). Auch diese Fälle wurden imputiert (etwa 2,3 % der Gebäude). Wo dies nicht möglich war, wurden die Gebäudeangaben mithilfe von Interviewern/Interviewerinnen der kommunalen Erhebungsstellen erhoben (rund 3 % der Gebäude). Diese Ergebnisse sind bei der Auswertung der vollständig fehlerfreien Gebäude mit berücksichtigt.

Tabelle 2 Imputationsraten je Wohnungsmerkmal
Prozent

	Art der Wohnungs-nutzung	Wohnfläche	Raumzahl	Badewanne/ Dusche vorhanden	WC vorhanden	Eigentums-verhältnisse der Wohnung	Wohnungstyp (Ferien-/ Freizeit- oder Diplomaten-wohnung)
Anteil der Wohnungen							
ohne Korrekturen[1]	87,1	86,2	86,2	86,9	86,6	54,7	84,1
mit deterministischer Imputation	0,2	0,4	0,5	2,9	3,1	28,5	5,6
mit CANCEIS-Imputation	12,7	13,4	13,3	10,2	10,3	16,8	10,3

1 Enthält vollständig imputierte und durch Erhebungsstellen begangene Gebäude (Unit-Nonresponse-Fälle).

Tabelle 3 Anteil der Gebäude mit und ohne Korrekturen nach Form des Dateneingangs[1]
Prozent

	Gebäude	
	ohne Korrekturen	mit Korrekturen
Ausschließlich online	51	49
Ausschließlich Papier	26	74
Ausschließlich Datenübermittlung für Wohnungsunternehmen (CORE)	40	60
Unterschiedlicher Dateneingang	7	93

1 Es sind nur Gebäude eingeschlossen, zu denen Angaben von Auskunftspflichtigen übermittelt wurden (kein Unit Nonresponse).

etwa 30 % der Auskunftspflichtigen die Fragebogen online aus. Tabelle 3 zeigt, dass etwa 51 % dieser ausschließlich online übermittelten Gebäudedaten vollständig fehlerfrei waren. Dies trifft aber nur auf ein Viertel der Gebäude zu, bei denen alle Angaben über Papierfragebogen erhoben wurden. Dies zeigt, dass der Online-Fragebogen dazu beigetragen hat, die Fehlermenge spürbar zu verringern. Vermutlich hängt dies unter anderem mit den in den Fragebogen integrierten Meldungen zusammen, die angezeigt wurden, wenn Fragen weggelassen oder inkonsistent beantwortet wurden. Es besteht in diesem Zusammenhang weiterer Untersuchungsbedarf hinsichtlich der Frage, ob einzelne Gruppen von Auskunftspflichtigen bestimmte Übermittlungswege bevorzugen.

4.3 Schlussfolgerungen für die Entwicklung von Imputationsverfahren

Aus den Erfahrungen lassen sich einige Schlussfolgerungen für die Konzeption der Datenaufbereitung für eine künftige Gebäude- und Wohnungszählung ziehen, zumindest wenn sie wieder als schriftliche Befragung in Vollerhebung stattfinden sollte.

Umfangreicherer Pretest des Fragebogens

Die bisherigen Ausführungen haben deutlich gemacht, dass Maßnahmen zur Minimierung von Item Nonresponse ein erster wichtiger Schritt sind, um die Datenqualität zu verbessern. Dabei ist insbesondere die Konzeption des Fragebogens – ob in Papierform oder online – von entscheidender Bedeutung. Da der GWZ-Fragebogen von vielen Millionen Befragten beantwortet werden musste, erscheint ein Pretest mit nur 18 Probandinnen und Probanden als zu gering. Dieser konnte nur einen kleinen Anteil der möglichen Konstellationen erfassen, die zu Problemen bei der Beantwortung der Fragen geführt haben.

Bei einer künftigen Gebäude- und Wohnungszählung sollte der Fragebogen in einem umfassenden Feldtest mit großen Teilnehmerzahlen getestet werden.

Kombination maschineller und manueller Verfahren?

Trotz des Anspruchs, aufgrund der großen Datenmenge überwiegend maschinelle Verfahren zur Plausibilisierung und Korrektur der Daten einzusetzen, hat sich gezeigt, dass in der Gebäude- und Wohnungszählung 2011 manuelle Arbeitsschritte dringend erforderlich waren. Dennoch sollte der Umfang der manuellen Arbeiten auf den Prüfstand gestellt werden, um beim nächsten Zensus den Zeitraum bis zur Veröffentlichung der Ergebnisse zu verkürzen. Hierfür muss insbesondere analysiert werden, ob und inwieweit manuelle Prüfungen und Korrekturen wirklich die Qualität der Daten in größerem Umfang und im Vergleich zu maschinellen Verfahren verbessern können. Dabei müssen Aufwand und Nutzen abgewogen werden. Einerseits gibt es Zusammenhänge, die sich nur schwer maschinell operationalisieren lassen und für die deshalb Erfahrungen mit bestimmten örtlichen Gegebenheiten von Vorteil sein können. Dies können maschinelle Verfahren nur schwer leisten. Andererseits kann es aber auch passieren, dass Menschen Ergebnisse entsprechend ihres eigenen Erfahrungshorizonts abgleichen und bestimmte Konstellationen verzerrend korrigieren („creative editing“).

Unabhängig von der Frage, welchen Stellenwert die manuellen Korrekturen in einer künftigen Gebäude- und Wohnungszählung haben werden, wäre auch der Einsatz neuer Verfahren zur Unterstützung der Plausibilisierung der Daten hilfreich – etwa die Prüfung der räumlichen Verteilung bestimmter Merkmalsausprägungen oder Kennzahlen mithilfe von Karten (GIS-Systeme). Auf diese Weise ließe sich zum Beispiel relativ schnell erkennen, ob in einzelnen Regionen mehr Ferien- und Freizeitwohnungen vorhanden sind als angenommen oder ob Leerstandsquoten in bestimmten Gebieten deutlich unter den Erwartungen beziehungsweise über bestimmten Vergleichsdaten liegen.

Auch eine stärkere Einbeziehung der Melderegister oder der Ergebnisse anderer Erhebungsteile (zum Beispiel der Haushaltsstichprobe) in die maschinellen Prüfungen könnte dazu beitragen, insgesamt die Konsistenz und damit die Qualität der Ergebnisse weiter zu verbessern.

Das Konzept einer kontinuierlichen Plausibilisierung muss überdacht werden

Wie erläutert, wurde in der Gebäude- und Wohnungszählung 2011 mit der Plausibilisierung nicht gewartet, bis alle Daten eingegangen waren, sondern die Fehlererkennung startete, sobald alle Angaben zu einem Gebäude vorhanden waren.

Die Imputation der Datensätze mit CANCEIS erfolgte zwar später, aber aufgrund der in Abschnitt 3.2 beschriebenen zeitlichen Rahmenbedingungen auch in zwei Läufen. Neben vielen Vorteilen hatte diese Vorgehensweise den entscheidenden Nachteil, dass ein Großteil der Imputationen mit CANCEIS bereits zu einem Zeitpunkt abgeschlossen war, als noch Fehler in den Daten vermutet und daher noch Prüfarbeiten notwendig waren. Die nach den CANCEIS-Imputationen durchgeführten Korrekturen könnten nachträglich das Imputationsergebnis beeinflusst haben, waren jedoch nur in relativ geringem Umfang notwendig.

In einer künftigen Gebäude- und Wohnungszählung sollte sichergestellt sein, dass die Imputationsverfahren erst zu einem Zeitpunkt starten, zu dem an den Datensätzen keine Veränderungen mehr vorgenommen werden (müssen).

Umfassendere Tests im Vorfeld

Bei der Gebäude- und Wohnungszählung 2011 war es mangels entsprechender vergleichbarer Daten aus vorangegangenen Erhebungen nicht möglich, die Verfahren im Vorfeld mit einer ausreichenden Menge an echten Einzeldaten zu testen und anzupassen. Dies wird bei einer künftigen Gebäude- und Wohnungszählung anders sein. Eine weitere mögliche Lösung könnte aber auch darin bestehen, zu Beginn der Datenaufbereitung die gesamten Daten testweise zu plausibilisieren oder die Fehlererkennungs- und Korrekturverfahren anhand einer Stichprobe der bis zu einem bestimmten Zeitpunkt eingegangenen Daten zu überprüfen. Eventuell könnten so bestimmte Probleme – wie bei der Gebäude- und Wohnungszählung 2011 die Beleglesefehler – früher erkannt und bereinigt werden. Allerdings ist die Entwicklung eines solchen Testverfahrens komplex und es müsste mit einem entsprechenden zeitlichen Aufwand bei der Umsetzung gerechnet werden.

Insgesamt lässt sich ein positives Fazit zum Umgang mit Unplausibilitäten in der Gebäude- und Wohnungszählung 2011 ziehen. Die Entscheidung für die eingesetzten Imputationsverfahren war richtig. Trotz einiger Schwierigkeiten ließ sich CANCEIS gut in den Datenaufbereitungsprozess der Gebäude- und Wohnungszählung 2011 integrieren und zur Imputation der Daten verwenden.

Auszug aus Wirtschaft und Statistik

Herausgeber
Statistisches Bundesamt, Wiesbaden

www.destatis.de

Schriftleitung
Dieter Sarreither,
Vizepräsident des Statistischen Bundesamtes

Redaktion: Ellen Römer
Telefon: +49 (0) 6 11 / 75 23 41

Ihr Kontakt zu uns
www.destatis.de/kontakt

Statistischer Informationsservice
Telefon: +49 (0) 6 11 / 75 24 05

Abkürzungen

WiSta	=	Wirtschaft und Statistik
MD	=	Monatsdurchschnitt
VjD	=	Vierteljahresdurchschnitt
HjD	=	Halbjahresdurchschnitt
JD	=	Jahresdurchschnitt
D	=	Durchschnitt (bei nicht addierfähigen Größen)
Vj	=	Vierteljahr
Hj	=	Halbjahr
a. n. g.	=	anderweitig nicht genannt
o. a. S.	=	ohne ausgeprägten Schwerpunkt
St	=	Stück
Mill.	=	Million
Mrd.	=	Milliarde

Zeichenerklärung

p	=	vorläufige Zahl
r	=	berichtigte Zahl
s	=	geschätzte Zahl
–	=	nichts vorhanden
0	=	weniger als die Hälfte von 1 in der letzten besetzten Stelle, jedoch mehr als nichts
.	=	Zahlenwert unbekannt oder geheim zu halten
. . .	=	Angabe fällt später an
X	=	Tabellenfach gesperrt, weil Aussage nicht sinnvoll
I oder —	=	grundsätzliche Änderung innerhalb einer Reihe, die den zeitlichen Vergleich beeinträchtigt
/	=	keine Angaben, da Zahlenwert nicht sicher genug
( )	=	Aussagewert eingeschränkt, da der Zahlenwert statistisch relativ unsicher ist

Abweichungen in den Summen ergeben sich durch Runden der Zahlen.