Dipl.-Soziologin Birgit Kleber, Dipl.-Forstwirtin Andrea Maldonado, Daniel Scheuregger, M. A., Dipl.-Sozialwissenschaftlerin Katja Ziprik

Aufbau des Anschriften- und Gebäuderegisters für den Zensus 2011

Im Jahr 2011 findet in Deutschland nach über 20 Jahren erneut eine Zählung der Bevölkerung und der Wohnungen statt. Die aktuellen Bevölkerungs- und Wohnungszahlen basieren auf Fortschreibungen der jeweils letzten Volkszählung, die in der Bundesrepublik Deutschland 1987 und in der ehemaligen DDR 1981 stattfand. Bevölkerungszahlen bilden ein wesentliches Fundament des statistischen Gesamtsystems. Der Zensus 2011 wird Basisdaten zu Bevölkerung, Erwerbstätigkeit und Wohnsituation in Deutschland liefern, auf denen viele politische, wirtschaftliche und gesellschaftliche Planungsprozesse aufbauen. Die bei einem Zensus ermittelten amtlichen Einwohnerzahlen werden beispielsweise auch als Bemessungsgrundlage für den horizontalen und vertikalen Finanzausgleich der Gebietskörperschaften und für die Einteilung der Bundestagswahlkreise herangezogen. Mit dem Zensus 2011 findet ein grundlegender Methodenwechsel im Vergleich zu den bisher in Deutschland durchgeführten Volkszählungen statt. Die traditionelle Form einer Vollbefragung der Bevölkerung wird – vor allem aus Akzeptanz- und Kostengründen – durch einen registerbasierten Zensus ersetzt.

Um einen solchen Methodenwechsel durchführen zu können, sind intensive Vorarbeiten nötig. Am Beginn des Zensusprojekts steht dabei der Aufbau eines Anschriften- und Gebäuderegisters aus bestehenden Verwaltungsregistern. Das Anschriften- und Gebäuderegister bildet die Grundlage für die Erhebung, Koordination und Auswertung des Zensus.

Der Artikel beschreibt den Aufbau des Anschriften- und Gebäuderegisters. In der Einführung werden die gesetzlichen Rahmenbedingungen sowie die verwendeten Datenquellen dargestellt. Das Kapitel „Record-Linkage-Verfahren" gibt einen Überblick über gängige Methoden, wie Daten aufbereitet und zusammengeführt werden. Schließlich wird erläutert, wie diese Verfahren während des Aufbaus des Anschriften- und Gebäuderegisters umgesetzt wurden, und ausgewählte Ergebnisse werden vorgestellt. Abschließend erfolgt ein kurzer Ausblick auf künftige Aufgaben und Arbeiten.

1 Rahmenbedingungen

1.1 Die neue Zensusmethode im Überblick

Die nachlassende Bereitschaft der Bevölkerung, an statistischen Erhebungen teilzunehmen, und die großen Fortschritte in der Informationstechnologie beim Verarbeiten größerer Datenbestände führten zu einem Paradigmenwechsel bei der Datenerhebungsmethode des Zensus 2011. Die zensustypischen Merkmale sollen nicht mehr dadurch gewonnen werden, dass die gesamte Bevölkerung durch Interviewerinnen und Interviewer befragt wird, sondern durch eine Kombination aus Registerauswertungen, einer postalischen Gebäude- und Wohnungszählung sowie einer Haushaltebefragung bei etwa 10 % der Bevölkerung durch Interviewerinnen und Interviewer ermittelt werden. Die Informationen aus den verschiedenen Datenquellen werden auf Personen- und Anschriftenebene zu einem Datensatz zusammengeführt.

Der Zensus 2011 wird im Einzelnen folgende Datenquellen nutzen:

– Kernbestand sind die Daten der Melderegister (MR) aller Gemeinden. Diese Daten werden jeweils an drei Stichtagen geliefert und mit einer Mehrfachfallprüfung maschi-

nell und manuell um eventuelle Fehler bereinigt, wobei es keine Rückmeldung an die zuständigen Meldebehörden geben wird.

– Aufgrund der Fehleranfälligkeit der Melderegister-Daten in Sonderbereichen, wie Gemeinschaftsunterkünften, Justizvollzugsanstalten usw., wird in diesen Bereichen eine gesonderte Erhebung stattfinden.

– Bezüglich der verfügbaren erwerbsstatistischen Daten wird auf Daten der Bundesagentur für Arbeit (BA) und Daten des Bundes und der Länder über unmittelbar in einem Dienst- oder Dienstordnungsverhältnis stehende Personen zurückgegriffen. Das Register der Bundesagentur für Arbeit enthält Angaben zu sozialversicherungspflichtig Beschäftigten, arbeitslos gemeldeten und an arbeitsfördernden Maßnahmen teilnehmenden Personen.

– Ein weiteres Register – die Georeferenzierten Adressdaten Bund (GAB) – wird durch die Vermessungsbehörden bereitgestellt. Dieses Register enthält Geoinformationen für die jeweiligen Anschriften.

– In einer Haushaltestichprobe werden etwa 10 % der Bevölkerung durch Interviewerinnen und Interviewer befragt, um zusätzliche Informationen zu gewinnen, die nicht aus den Registern bezogen werden können.

– Weitere Angaben zu Gebäuden und Wohnungen werden durch eine Gebäude- und Wohnungszählung gewonnen. In dieser werden schätzungsweise 17,5 Mill. Hauseigentümer oder -verwalter schriftlich befragt.

Um dem Bedarf an aktuellen Bevölkerungszahlen gerecht zu werden und den entsprechenden Vorgaben der Europäischen Union (EU) nachkommen und die Verpflichtungen erfüllen zu können, die sich aus der EG-Verordnung über Volks- und Wohnungszählungen[1]) für das Jahr 2011 in allen Mitgliedstaaten ergeben (EU-Pflichtprogramm, siehe Schaubild 1), wurde auf nationaler Ebene 2007 das Zensusvorbereitungsgesetz 2011[2]) erlassen; auch das Zensusgesetz 2011[3]) ist am 16. Juli 2009 in Kraft getreten.

1.2 Die Aufgaben des Anschriften- und Gebäuderegisters

Die rechtliche Grundlage für die konkrete Vorbereitung des Zensus 2011 wurde in Deutschland mit dem Zensusvorbereitungsgesetz 2011 (ZensVorbG 2011) geschaffen, welches am 13. Dezember 2007 in Kraft trat. § 2 des ZensVorbG 2011 regelt hierbei insbesondere die Nutzung von Registerangaben zum Aufbau eines Anschriften- und Gebäuderegisters (AGR). Das Anschriften- und Gebäuderegister soll alle Anschriften von Gebäuden mit Wohnraum und bewohnten Unterkünften enthalten[4]) und dient im Zensus dazu,

1. den Ablauf der Gebäude- und Wohnungszählung sowie die Ablaufkontrolle aller primärstatistischen Erhebungen des Zensus zu steuern,

2. die beim Zensus vorgesehenen Stichprobenerhebungen vorzubereiten und aus ihm die Stichprobeneinheiten auszuwählen,

Schaubild 1

EU-Pflichtprogramm: Themen der Volks- und Wohnungszählungen

Demografische und geografische Merkmale	Erwerbs- und bildungsstatistische Merkmale	Haushalts- und familienstatistische Merkmale	Gebäude- und wohnungsstatistische Merkmale
• Geschlecht	• Aktueller Erwerbsstatus	• Haushaltstyp	• Art der Unterkunft
• Alter	• Ausgeübter Beruf	• Haushaltsgröße	• Gebäudetyp
• Familienstand	• Wirtschaftszweig	• Stellung im Haushalt	• Belegungsstatus
• Staatsangehörigkeit	• Stellung im Beruf	• Stellung in der Familie	• Baujahr des Gebäudes
• Geburtsland/-ort	• Arbeitsort	• Typ der Kernfamilie	• Nutzfläche und/oder Zahl der Räume
• Üblicher Aufenthaltsort	• Bildungsniveau	• Größe der Kernfamilie	• Zahl der Bewohner
• Vorheriger üblicher Aufenthaltsort und Datum der Ankunft oder üblicher Aufenthaltsort ein Jahr vor dem Zensus			• Besitzstatus
• Jemals im Ausland gelebt und Jahr der Ankunft in Deutschland			• Ausstattung der Wohnung
			• Heizung

1) Verordnung (EG) Nr. 763/2008 des Europäischen Parlaments und des Rates vom 9. Juli 2008 über Volks- und Wohnungszählungen (Amtsbl. der EU Nr. L 218, S. 14).
2) Gesetz zur Vorbereitung eines registergestützten Zensus einschließlich einer Gebäude- und Wohnungszählung 2011 (Zensusvorbereitungsgesetz 2011 – ZensVorbG 2011) vom 8. Dezember 2007 (BGBl. I S. 2808).
3) Gesetz zur Anordnung des Zensus 2011 sowie zur Änderung von Statistikgesetzen vom 8. Juli 2009 (BGBl. I S. 1781).
4) Die im Anschriften- und Gebäuderegister gespeicherten Merkmale sind im Einzelnen in § 2 ZensVorbG 2011 festgelegt.

3. die Erhebungen für den Zensus zu koordinieren, im Rahmen der Durchführung des Zensus die aus verschiedenen Quellen stammenden Daten zusammenzuführen und die in den Zensus einzubeziehenden Gebäude, Wohnungen und Personen auf Vollzähligkeit zu prüfen,

4. ein System von raumbezogenen Analysen und Darstellungen von statistischen Ergebnissen zu entwickeln und die Grundlage für eine kleinräumige Auswertung des Zensus zu schaffen.

Im April 2008 wurden die Daten der Meldebehörden, Vermessungsbehörden und der Bundesagentur für Arbeit an das Statistische Bundesamt übermittelt. Seither wurden die Daten aufbereitet und bereinigt. Aktuell wird die Zusammenführung der Datenquellen auf Anschriftenebene abgeschlossen, sodass ein erstes vorläufiges Anschriften- und Gebäuderegister entsteht, welches alle Anschriften speichert, die in mindestens einem der Register enthalten waren. Diejenigen Anschriften, die nur in einem Register enthalten waren, werden durch die Statistischen Ämter der Länder geprüft, gegebenenfalls korrigiert und zu einem späteren Zeitpunkt erneut in das vorläufige Anschriften- und Gebäuderegister integriert.

1.3 Struktur und Basisquellen des Anschriften- und Gebäuderegisters

Als Grundlage für den Aufbau des Anschriften- und Gebäuderegisters dienen nach § 4 bis 6 ZensVorbG 2011 die Daten der Melderegister, die Georeferenzierten Adressdaten Bund und die Daten der Bundesagentur für Arbeit. Die drei liefernden Stellen übermitteln ihre Daten elektronisch zu festgelegten Stichtagen an die Statistischen Ämter des Bundes und der Länder. Die mitgelieferten personenbezogenen Daten in den Registern der Meldebehörden und der Bundesagentur für Arbeit werden für den Aufbau des Anschriften- und Gebäuderegisters nicht gebraucht und deshalb von den benötigten Anschriftenmerkmalen abgetrennt, sodass die statistische Geheimhaltung gewahrt wird. Da eine Anschrift für mehrere Personen gültig sein kann, ist die Zahl der Anschriften in den MR- und BA-Daten immer kleiner als die Zahl der Personen. Zur Erstellung eines Anschriften- und Gebäuderegisters sind nicht die Personen von Bedeutung, sondern nur deren Anschriften.

Die drei genannten Quellen unterscheiden sich im Aufbau und in den darin enthaltenen Informationen. Die Registerauszüge der Vermessungsbehörden enthalten Anschriften aller Gebäude, die durch die Katasterbehörden der Länder vermessen wurden (Hauskoordinaten). Die amtlichen Hauskoordinaten definieren dabei die genaue Position eines Hauses. Diese Quelle enthält Anschriften von Gebäuden mit Wohnraum und Anschriften rein gewerblich genutzter Gebäude und damit den umfangreichsten Bestand an Registeranschriften.

Die Vermessungsdaten werden in vier jährlichen Lieferungen elektronisch an das Statistische Bundesamt übermittelt. Sie enthalten zu allen vermessenen Gebäuden in Deutschland die Angaben zu den relevanten AGR-Merkmalen, mit denen die ersten Zusammenführungen stattfinden. Die erste Lieferung erfolgte zum 1. April 2008. Sie enthält die Merkmale der Anschriften, die am 1. April 2007 gültig waren. Die zweite, dritte und vierte Lieferung erfolgt(e) jeweils bis zum 31. Juli der Jahre 2008, 2009 und 2010. Neben den Anschriftenmerkmalen enthalten die Vermessungsdaten Angaben zu den geografischen Hauskoordinaten. Diese Information ist von großer Bedeutung für die Identifikation von Anschriften, die sich in allen Angaben gleichen, aber dennoch physisch getrennt sind (sogenannte Pseudo-Identische Anschriften). Insbesondere in ländlichen Gebieten existieren aufgrund zum Beispiel von Eingemeindungen Straßen, die in allen oben angeführten Merkmalen identisch sind, aber dennoch physisch unterschiedliche Orte bezeichnen können. Die Verschiedenheit dieser Anschriften kann in den Daten nur aufgrund der geografischen Hauskoordinaten festgestellt werden.

Die zweite Quelle sind die Registerauszüge der Meldebehörden. Diese sind für den registerbasierten Zensus eine besonders wichtige Datenquelle, um zum einen die amtliche Bevölkerungszahl feststellen und zum anderen die demografische Grundstruktur der Bevölkerung abbilden zu können. Aufgrund des in Deutschland geltenden Melderechts sind hier alle Personen enthalten, die eine Wohnung bezogen haben. Ummeldungen sind innerhalb kurzer Fristen bei den jeweiligen Behörden zu melden. Aus diesen Gründen haben diese Daten eine hohe Aktualität und bilden daher u. a. eine wichtige Grundlage bei der Feststellung der amtlichen Bevölkerungszahl. Das heißt aber auch, dass dieser Datenbestand – im Unterschied zu dem der Vermessungsbehörden – nur Anschriften enthält, an denen Personen in Deutschland gemeldet sind. Die im Melderegister enthaltenen Anschriften bilden dementsprechend eine Untermenge der Anschriften der Vermessungsbehörden.

Von den Meldebehörden werden zwei Datenlieferungen zum Aufbau des Anschriften- und Gebäuderegisters herangezogen: zu den Stichtagen 1. April 2008 und 1. April 2010. Hierbei liefern die Gemeinden für jede gemeldete Person Angaben zur Anschrift und zu ausgewählten demografischen Merkmalen an das jeweils zuständige Statistische Landesamt. Dort werden die Daten auf Plausibilität geprüft, gegebenenfalls bereinigt und in die beim Statistischen Bundesamt vorgehaltene Datenbank geladen. Die Meldedaten werden dabei auf Personenbasis geliefert. Diese Personenmerkmale werden aus datenschutzrechtlichen Gründen nach der Übermittlung von den anderen Zensusmerkmalen abgetrennt. Zum Aufbau des Anschriften- und Gebäuderegisters werden zunächst die relevanten Merkmale zur gegenwärtigen Anschrift (Amtlicher Gemeindeschlüssel, Postleitzahl, Straße, Hausnummer, Hausnummernzusatz) genutzt. Dafür findet eine Aggregation der Melderegister-Datensätze auf die Merkmale der Anschrift statt, über die mit den anderen Quellen zusammengeführt wird.

Von der Bundesagentur für Arbeit wird zum Aufbau des Anschriften- und Gebäuderegisters nur eine Lieferung genutzt. Neben den Anschriftenmerkmalen enthält diese Lieferung den Erwerbsstatus aller am 30. September 2007 sozialversicherungspflichtig Beschäftigten und der am 13. März 2008 arbeitslos gemeldeten Personen. Die Angaben zum Erwerbsstatus werden für die Stichprobenmodellierung be-

nötigt. Die Anschriftendaten der Bundesagentur für Arbeit werden, wie beim Melderegister, auf Personenebene übermittelt und im ersten Schritt für die einzelnen Anschriften zusammengefasst. Im Zensus selbst wird die Zuordnung der Angaben zur Erwerbstätigkeit zu den Angaben aus den Melderegistern im ersten Schritt über die Anschrift und im zweiten über den Namen der erwerbstätigen Person erfolgen.

Wie bereits erläutert enthalten die verschiedenen Register unterschiedliche Datenmengen und auch in unterschiedlichem Maße für den Zensus relevante Angaben. Die GAB-Daten enthalten beispielsweise neben Wohngebäuden auch rein gewerblich genutzte Gebäude, die jedoch für die Erfassung der Wohnbevölkerung nicht relevant sind. Daher übersteigen die Georeferenzierten Adressdaten Bund auch die Anzahl der Anschriften in den Melderegistern und in den Registern der Bundesagentur für Arbeit. Die GAB-Daten enthalten hingegen keine Angaben zu Personen, weshalb die Anzahl der eingetragenen Datensätze hier geringer ausfällt als in den übrigen Registern.

Das zentrale Problem bei der Erstellung des Anschriften- und Gebäuderegisters ist die Zusammenführung der drei Registerbestände auf Anschriftenebene. Da in Deutschland für Anschriften kein eindeutiges und in allen Registern geführtes Verknüpfungsmerkmal (z. B. eine Identifikationsnummer) existiert, muss alternativ auf eine Kombination von Merkmalen zurückgegriffen werden, über welche eine deterministische Zusammenführung der Daten erfolgen kann. Als gemeinsame Merkmale, über die eine eindeutige Identifikation einer Anschrift möglich ist, enthalten die drei Quellen unter anderem die Merkmale

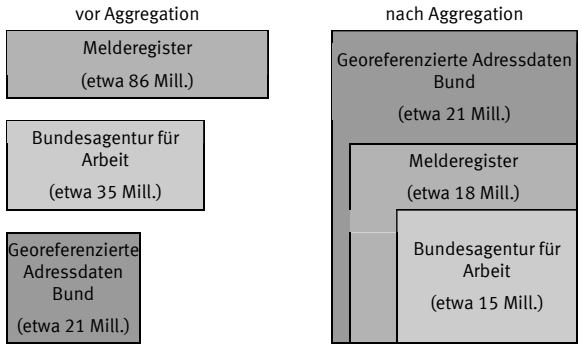
- Amtlicher Gemeindeschlüssel,
- Postleitzahl,
- Straße,
- Hausnummer und
- Hausnummernzusatz.[5])

Daher werden zunächst alle drei Register auf Anschriftenebene aggregiert. Durch dieses Verfahren verändern sich die Größenverhältnisse der Register zueinander (siehe Schaubild 2). Während die GAB-Daten, da hier nur Anschriftenangaben enthalten sind, bei der Lieferung den kleinsten Bestand darstellten, bilden sie nach der Aggregation die größte Menge.

Da in Deutschland sowohl für Anschriften als auch für Personen kein eindeutiges Identifikationsmerkmal existiert, ergeben sich für einen registergestützten Zensus bei der Zusammenführung der Einzelregister besondere Probleme. Im Folgenden werden methodische Aspekte der Datenintegration sowie deren Umsetzung beim Aufbau des Anschriften- und Gebäuderegisters dargestellt.

Schaubild 2

Mengenverteilung vor und nach der Aggregation der Registeranschriften[1])



1) Die Größenverhältnisse entsprechen nicht der vorhandenen Datensatzanzahl.

1.4 Record-Linkage-Verfahren

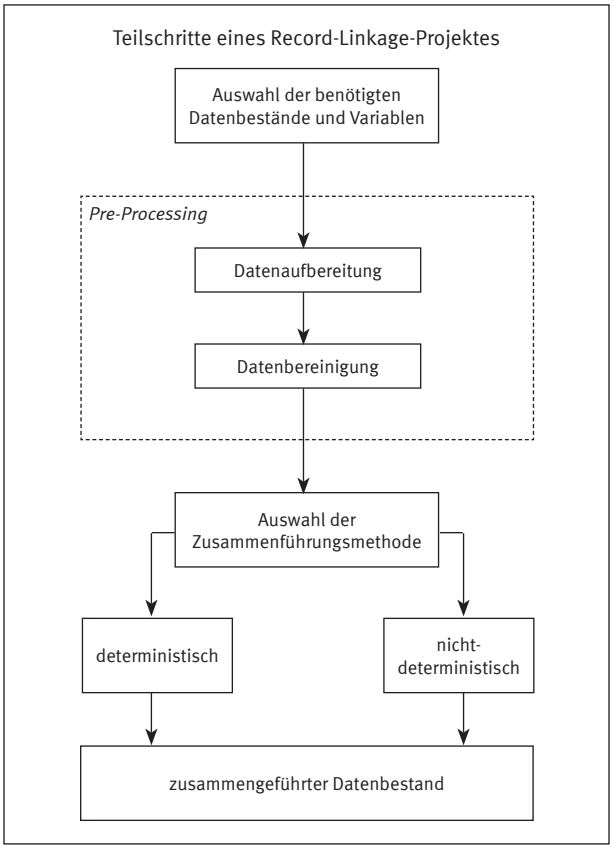
Unter Record-Linkage bzw. Datenzusammenführung wird eine Zusammenführung von Informationen aus unterschiedlichen Datenbeständen verstanden, deren Angaben zur gleichen Beobachtungseinheit gehören. Ziel dieser Verknüpfung ist es, die bereits in den Datenbeständen vorliegenden Informationen umfassender auszuwerten und mehr Informationen zu einer Einheit zu erhalten. Da durch Record-Linkage-Verfahren auch bereits für andere Zwecke erhobene Daten in neuen Kombinationen ausgewertet werden können, ist die Anwendung von Record-Linkage-Verfahren zudem ökonomisch effizient.[6]) Im Falle des Zensus 2011 können diese Vorteile durch die Zusammenführung administrativer Register, die originär für andere Zwecke erstellt wurden, genutzt werden. Grundsätzlich wird die Zusammengehörigkeit unterschiedlicher Datenbestände im Rahmen von Record-Linkage-Verfahren durch einen Paarvergleich von Merkmalen bestimmt.[7]) Für die eigentliche Zusammenführung sind jedoch vorbereitende Arbeiten an den jeweiligen Datenbeständen notwendig. Die Zusammenführung von Daten ist daher ein Arbeitsprozess, der sich in verschiedene Arbeitsschritte, die zusammenfassend in Schaubild 3 dargestellt sind, untergliedert.

Bei der Auswahl der Datenbestände sind zunächst die Ausgangsregister auf die darin enthaltenen Informationen zu prüfen. Es gilt abzuwägen, ob die enthaltenen Merkmale dem Erkenntnisziel entsprechen und die vorhandenen Daten eine Datenzusammenführung grundsätzlich ermöglichen. Für den Zensus 2011 wurde dies mit dem Zensustest 2001 geprüft. Dabei zeigte sich, dass die geprüften Register sowohl die für eine Zusammenführung notwendigen als auch die für den Zensus relevanten Informationen enthalten.[8])

Nach der Auswahl der Datenquellen wird das Pre-Processing durchgeführt, mit dem die Zusammenführungsvariablen für

5) Diese fünf Merkmale sind im Folgenden gemeint, wenn von Anschrift gesprochen wird.
6) Siehe Statistics New Zealand: "Data Integration Manual", 2006 (*www.stats.govt.nz*).
7) Siehe Winkler, E.: "Overview of Record Linkage and Current Research Directions" in U. S. Census Bureau (Hrsg.), Statistical Research Division, Research Report Series, Statistics #2006-2, Washington D. C., 2006, sowie Schürle, J.: „Record Linkage – Zusammenführung von Daten auf Basis des Modells von Fellegi und Sunter", Frankfurt a. M. 2004, S. 23.
8) Siehe Statistische Ämter des Bundes und der Länder: „Ergebnisse des Zensustests" in WiSta 8/2004, S. 813 ff.

Schaubild 3



die Verknüpfung der Daten in eine geeignete Form gebracht werden sollen. Das Pre-Processing kann, wie in Schaubild 3 dargestellt, in Datenaufbereitung und Datenbereinigung untergliedert werden. Die Datenaufbereitung umfasst dabei Arbeitsschritte, in denen die Registerbestände nach Regeln umkodiert oder in eine neue Anordnung gebracht werden. Hierzu zählen Arbeiten wie das Zerlegen von Zeichenketten (Parsing), die Standardisierung oder die Plausibilisierung, auf die weiter unten noch genauer eingegangen wird. Im zweiten Teilschritt – der Datenbereinigung – wird auf zusätzliche Informationen aus externen Datenquellen zurückgegriffen. Ziel dieser Arbeiten ist es, die Datenqualität zu verbessern, das heißt die Angaben in den Registern möglichst vollständig, korrekt und aktuell zu erhalten. Hierzu werden falsche Eintragungen identifiziert, korrigiert und die Schreibweisen aller Eintragungen in jedem Datenbestand in gleicher Weise standardisiert. Für das eigentliche Zusammenführen der Datenbestände ist nach Abschluss der Vorbereitungen ein geeignetes Verfahren auszuwählen. Hierbei kann zwischen deterministischen und nichtdeterministischen Verfahren der Zusammenführung unterschieden werden, die jeweils unterschiedliche Anforderungen an die Daten stellen.

1.4.1 Deterministische Zusammenführungen

Bei deterministischen Zusammenführungen werden Daten auf Basis der Identität von Merkmalen zusammengeführt. Hierzu wird ein eineindeutiger Identifikator gebildet, der den unterschiedlichen Informationen in beiden Datenbeständen zugewiesen wird. Liegt der Identifikator in beiden Datenbeständen vor, kann jeder Datensatz auf die Gleichheitsbedingung geprüft und zusammengeführt werden. Der Vorteil deterministischer Zusammenführungen liegt vor allem in einfachen Zusammenführungsregeln. Dieser Einfachheit des Verfahrens steht jedoch ein hoher Anspruch an die Datenqualität entgegen. Da Daten nur bei Identität zusammengeführt werden können, bedeutet jeder Informationsausfall (fehlende oder falsche Angaben), dass keine Integration des Datenbestandes möglich ist. Darüber hinaus ist dieses Verfahren auch intolerant gegenüber nichteindeutigen Informationen. Da Identität Eineindeutigkeit voraussetzt, können Datensätze nur zusammengeführt werden, wenn einem Element in einem Datensatz genau ein identisches Element im anderen Datenbestand entspricht.

Die Identität zweier Datensätze kann bei deterministischen Zusammenführungen über numerische und alphanumerische Zeichenketten oder Variablentupel[9]), die eine eindeutige Zuweisung ermöglichen, erfolgen. Da in Deutschland weder eine eineindeutige Kennung für Straßen noch für Anschriften noch für Personen existiert, wird die Eineindeutigkeit im Anschriften- und Gebäuderegister durch eine Kombination von Variablen erzielt. Im Falle der Anschriften sind dies etwa die genannten Anschriftenmerkmale.

1.4.2 Nichtdeterministische Zusammenführungen

Bei nichtdeterministischen Zusammenführungsverfahren werden die strengen Voraussetzungen deterministischer Verfahren abgeschwächt. Die Informationen für die Zusammenführung müssen nicht notwendigerweise eindeutig, gleich oder vollständig sein. Inhaltlich kann dies als eine Zusammenführung auf Basis von Plausibilität aufgefasst werden. In analoger Weise würde auch ein manuelles Zusammenführen erfolgen, also wenn die Paarigkeit zweier nicht übereinstimmender Datensätze durch einen Menschen beurteilt würde. Die an der Zusammenführung beteiligten Personen würden hierbei Vermutungen anstellen, um die Identität von ungleichen Datenbeständen zu bestimmen und eine Zusammenführung plausibel zu begründen. Dieser kognitive Prozess kann maschinell übersetzt werden und auf diese Weise können sonst mühsam durch Menschen vorgenommene Einzelfallprüfungen automatisiert werden.

Beim Aufbau des Anschriften- und Gebäuderegisters werden nichtdeterministische Zusammenführungen vor allem im Pre-Processing eingesetzt, da in großen Datenbeständen schon geringe Anteile problematischer Daten die manuellen Korrekturkapazitäten überfordern. Um dennoch nichtidentische, aber möglicherweise zusammengehörige Angaben identifizieren zu können, müssen Bedingungen formuliert werden, auf deren Basis die Zusammengehörigkeit der Daten vorausgesetzt werden kann. Da es sich bei den Angaben im Anschriften- und Gebäuderegister um alphanumerische Informationen handelt, werden hierzu Ähnlichkeiten von Zeichenketten durch Distanzmaße ermittelt. Mit Distanzmaßen wird versucht, die „Entfernung“ einer Zeichenkette

9) Kombinationen verschiedener Variablen.

zu einer Vergleichskette zu bestimmen und somit graduelle Abstufungen der Übereinstimmung numerisch auszuweisen. Bei der Aufbereitung der zensusrelevanten Registerbestände wurden hierbei vor allem die Levenshtein-Distanz und die Jaro-Winkler-Distanz verwendet.

Mit der Levenshtein-Distanz wird die Ähnlichkeit zweier Ketten durch das Minimum der notwendigen Operationen zur Überführung einer Zeichenkette in eine Vergleichskette ausgedrückt. Als Operation wird in diesem Zusammenhang sowohl eine Löschung als auch das Hinzufügen oder das Ersetzen eines Zeichens gewertet. Wird der Name „Meyer“ mit „Meier“ verglichen, kann „Meyer“ durch den Austausch des „y“ in „Meier“ überführt werden. Die Distanz beträgt somit 1. Um den Vergleich der Levensthein-Distanz mit anderen Ähnlichkeitsmaßen zu erhöhen, kann der Wert auch über die Stringlänge auf ein Intervall zwischen 0 und 100 normiert werden, wobei 100 vollständige Identität anzeigt:

$$(1) \qquad d_{s_1,s_2} = 1 - \frac{\sum a}{b}$$

mit a = Operation

b = Länge der Zeichenkette

Im Falle des obigen Beispiels ergäbe sich etwa:

$$(2) \qquad d_{s_1,s_2} = 1 - \frac{1}{5} = 0{,}8$$

Im Gegensatz zur Levenshtein-Distanz werden bei der Jaro-Winkler-Distanz nicht allein die notwendigen Operationen zur Transformation einer Zeichenkette gezählt, sondern darüber hinaus auch die Übereinstimmungen und Unterschiede zweier Zeichenketten als Anteile zueinander in Beziehung gesetzt. Ausgangsmaß für die Jaro-Winkler-Distanz ist zunächst die Jaro-Metrik:

$$(3) \qquad d_j = \frac{1}{3}\left(\frac{c}{s_1} + \frac{c}{s_2} + \frac{c-t}{c}\right)$$

mit c = Anzahl der übereinstimmenden Elemente

s_1 = Länge der ersten Zeichenkette

s_2 = Länge der zweiten Zeichenkette

t = Anzahl der notwendigen Änderungen.

Die Jaro-Metrik wird im Anschluss zur Jaro-Winkler-Distanz erweitert:

$$(4) \qquad d_w = d_j + \left(\ell p (1 - d_j)\right)$$

mit d_j = Jaro-Metrik

ℓ = Länge des übereinstimmenden Präfix bis maximal zur Länge 4

p = Konstanter Faktor (üblicherweise 0,1)

Am genannten Beispiel der Ähnlichkeit von „Meier“ und „Meyer“ ergibt sich danach:

$$(5) \qquad d_j = \frac{1}{3}\left(\frac{4}{5} + \frac{4}{5} + \frac{4-1}{4}\right) = 0{,}78$$

Und für die Jaro-Winkler-Distanz:

$$(6) \qquad d_w = 0{,}78 + \left(2 \cdot 0{,}1 \cdot (1 - 0{,}78)\right) = 0{,}83$$

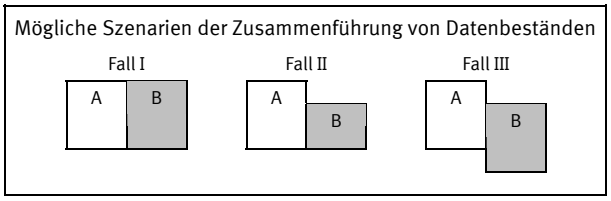
Mit der Anwendung von Distanzmaßen ist die Identität von Zeichenketten nur näherungsweise zu bestimmen, und es können in Abhängigkeit von den verwendeten Maßen gleichwertige Lösungen existieren. „Meyer“ etwa hätte zur alternativen Schreibweise „Mayer“ sowie zur Schreibweise „Meier“ eine Levenshtein-Distanz von 1. Der Fall ist damit maschinell allein auf Basis dieses Maßes nicht zu entscheiden. Die Jaro-Winkler-Distanz würde hier hingegen deutlicher unterscheiden, da hier die Präfixe „Mei“, „May“ und „Mey“ aus „Meier“, „Mayer“ bzw. „Meyer“ berücksichtigt würden. Aufgrund dessen wird die Jaro-Winkler-Distanz häufig vorgezogen oder ergänzend zur Levenshtein-Distanz verwendet.

Mit der Anwendung von Distanzmaßen kann die Identität von Zeichenketten nur näherungsweise bestimmt werden und es bleibt ein Entscheidungsspielraum für das Festsetzen eines Grenzwertes, ab dem eine Identität ungleicher Angaben unterstellt wird. Im Gegensatz zur an der formalen Identität orientierten deterministischen Zusammenführung enthalten nichtdeterministische Verfahren damit inhaltliche Annahmen, die Einfluss darauf haben, wie die Elemente in den unterschiedlichen Datenbeständen zugeordnet werden. Je nach Festlegung des Entscheidungskriteriums sind Fehler zu erwarten. Fordert man eine starke Ähnlichkeit, wird eine höhere Sicherheit für korrekte Zuordnungen erzielt. Gleichzeitig aber verringert sich so die Menge der zugeordneten Fälle und entsprechend bleibt auch eine große Korrekturmasse zurück. Setzt man die Kriterien dagegen etwas großzügiger an, erhöht sich zwar die Quote der zusammengeführten Datensätze, jedoch steigt damit auch das Risiko falscher Zuordnungen. Das konkrete Vorgehen bei nichtdeterministischen Verfahren umfasst daher stets einen Abwägungsprozess zwischen effizienter Verwertung von (unvollständigen) Informationen einerseits und dem Problem der Begrenzung von Fehlerraten andererseits.

Würden beim Einsatz von Distanzmaßen nicht die Vergleichsmengen begrenzt, müssten alle Zeichenketten mit allen anderen verglichen werden. In diesem Falle stiege die Anzahl der Paarvergleiche schnell an. Um die Zahl der Vergleichspaare bei der Berechnung von Ähnlichkeitsmaßen zu reduzieren, werden daher oft Blockingvariablen verwendet; mit diesen werden Untergruppen definiert, auf die der Paarvergleich begrenzt wird. Beim Aufbau des Anschriften- und Gebäuderegisters wird beispielsweise die Ähnlichkeit von Straßennamen nur für Straßen mit identischem amtlichen Gemeindeschlüssel bestimmt. Durch dieses Blocking über die AGS-Information erhöht sich zudem die Sicherheit einer korrekten Identifikation, da es wahrscheinlicher ist, dass ähnliche Straßen im gleichen Gemeindegebiet tatsächlich identisch sind, als es möglicherweise weit entfernte Straßen sind. Dieser Zuwachs an Sicherheit ist indes an die Qualität der Blockingvariablen gebunden. Fehlerhafte Angaben (etwa falsche Gemeindeschlüssel) in der Blockingvari-

ablen führen zu möglicherweise falschen Zuordnungen und können somit ungeprüft sogar negative Effekte hervorrufen. Unabhängig vom gewählten Zusammenführungsverfahren sind verschiedene Ergebnisvarianten möglich. Optimal wäre der in Schaubild 4 dargestellte Fall I. Die Zusammenführung ist korrekt und vollständig. Realistischer sind jedoch die Fälle II und III. Es verbleiben hierbei Restmengen, die einer weiteren Behandlung bedürfen, da sie nicht zusammengeführt werden konnten.

Schaubild 4



Um die Restmenge zu reduzieren, können im Anschluss weitere Schritte der Datenaufbereitung und/oder -bereinigung vorgenommen und Zuordnungsfehler nachträglich korrigiert werden. Beim Aufbau des Anschriften- und Gebäuderegisters werden die Restmengen der Anschriften gesondert analysiert und schließlich entweder maschinell weiterverarbeitet oder in Kooperation mit den Statistischen Ämtern der Länder geprüft.

Die Erstellung des Anschriften- und Gebäuderegisters kann damit nicht als statische Schrittfolge verstanden werden, da sie auch iterative Optimierungen umfasst, um die abschließende Quote bei der Zusammenführung zu erhöhen. In vielen Fällen ist auch eine manuelle Prüfung von Restmengen nicht vollständig zu vermeiden. Die sinnvolle Anwendung deterministischer und nichtdeterministischer Zusammenführungsmethoden ist daher nicht immer eine Entweder-oder-Entscheidung, sondern hängt von der Zielsetzung und der Arbeitsphase eines Projektes ab. Die Funktion des Anschriften- und Gebäuderegisters macht eine deterministische Zusammenführung notwendig. Bei der Datenaufbereitung bieten sich jedoch ergänzend nichtdeterministische Verfahren an, um Korrekturmassen zu bestimmen und zu bearbeiten, sodass beim Aufbau des Anschriften- und Gebäuderegisters verschiedene Ansätze kombiniert werden.

2 Pre-Processing

2.1 Datenaufbereitung

Da ein eindeutiger Identifikator für die Datenzusammenführung fehlt, werden die Register über die Einzelangaben der Anschriften zusammengeführt. Die Kombination dieser Angaben ist hinreichend, um eine eindeutige Identifikation zu garantieren und eine deterministische Zusammenführung zu ermöglichen. Jedoch müssen die Angaben zur Anschrift in allen Datensätzen über eine sehr hohe Datenqualität verfügen.

Während der Datenaufbereitung werden die Register auf Vollständigkeit geprüft und die Vergleichbarkeit durch das Parsing (Zerlegen von Zeichenketten) und die Standardisierung der Zusammenführungs-Variablen gewährleistet.

Da die Datenmenge, die für den Aufbau des Anschriften- und Gebäuderegisters genutzt wird, sehr umfangreich ist (siehe die Schaubilder 1 und 2), bedeuten schon geringe prozentuale Fehleranteile einen hohen manuellen Korrekturaufwand. Es ist daher das Ziel, die einzelnen Schritte der Datenaufbereitung weitestgehend maschinell durchzuführen.

2.1.1 Plausibilisierung

Bei der Plausibilitätskontrolle werden offensichtliche Unrichtigkeiten, wie etwa logisch oder sachlich widersprüchliche Angaben, eliminiert. Die Plausibilität der Registerangaben wurde von unterschiedlichen Stellen geprüft. Für die GAB-Datei wurde die Plausibilitätskontrolle vom Bundesamt für Kartographie und Geodäsie durchgeführt, sodass ein bereits geprüfter Datensatz an das Statistische Bundesamt geliefert wurde. Beim Statistischen Bundesamt wurde sodann auf Vollständigkeit der Datensätze und Aktualität des Amtlichen Gemeindeschlüssels geprüft.

Für den Datenbestand der Bundesagentur für Arbeit wurde eine Plausibilitätsprüfung durchgeführt. Die Vollzähligkeit der BA-Daten wurde auf Gemeindeebene kontrolliert. Es erfolgte des Weiteren ein Abgleich des Amtlichen Gemeindeschlüssels im Datenbestand der Bundesagentur für Arbeit mit dem aktuellen Amtlichen Gemeindeschlüssel aus dem Gemeindeverzeichnis-Informationssystem (GV-ISys)[10]) sowie ein Abgleich der Postleitzahlen mit aktuellen Angaben aus der PostDirekt-Datei[11]). Durch diesen Vorgang wurde die Korrektheit der vorhandenen Amtlichen Gemeindeschlüssel und Postleitzahlen bestätigt. Die beiden genannten Variablen durften per Definition zudem nur eine bestimmte Zeichenlänge besitzen: Postleitzahlen mussten fünf Zeichen enthalten, der amtliche Gemeindeschlüssel acht Zeichen.

Die Plausibilisierung der Daten der Melderegister erfolgte durch die Statistischen Ämter der Länder. Sie bestand aus einem Regelkatalog, der u. a. Vollständigkeitsprüfungen und Zeichenprüfungen enthielt. Zudem wurde kontrolliert, ob der achtstellige Amtliche Gemeindeschlüssel der Anschrift identisch ist mit dem achtstelligen Amtlichen Gemeindeschlüssel der liefernden Gemeinde. Stimmen beide Variablen nicht überein bedeutet das, dass sich die jeweilige Anschrift nicht in der Gemeinde, für die die Daten geliefert wurden, befindet. In diesen Fällen korrigierten die Statistischen Ämter der Länder die Daten für die entsprechende Gemeinde.

10) Das GV-ISys der Statistischen Ämter des Bundes und der Länder führt jede politisch selbstständige Gemeinde Deutschlands u. a. mit den Merkmalen Amtlicher Gemeindeschlüssel, Gemeindenamen, Postleitzahl des Verwaltungssitzes der Gemeinde bzw. Stadt.

11) Bei der PostDirekt-Datei handelt es sich um ein öffentlich zugängliches, kostenpflichtiges Verzeichnis aller gültigen Straßen der Deutschen Post Direkt GmbH. Dieses Verzeichnis wurde mit den amtlichen Straßenverzeichnissen von sieben Großstädten abgeglichen und ergänzt.

2.1.2 Parsing

Beim Parsing wurden die Angaben zu den Anschriften nach einheitlichem Muster aus einem Feld extrahiert und in getrennte Felder eingefügt. Besonders aufwendig war dies im Falle der BA-Daten, da Teile der Anschriften – das heißt Straßennamen, Hausnummer, Hausnummernzusätze usw. – in ein Feld geschrieben wurden. Zur maschinellen Auftrennung dieser Zeichenketten wurden reguläre Ausdrücke verwendet, mit denen Muster in Zeichenfolgen identifiziert und auf eine jeweils zu bestimmende Weise weiterverarbeitet werden können.

Zur Mustererkennung diente eine Kombination von Zeichen, welche als Indiz fungierten, um Grenzen von Zeichenketten zu bestimmen. So kann beispielsweise ein Bindestrich, der von alphabetischen Zeichen umgeben ist, einen Ortsteil (OT-Nord) oder einen Straßennamen (Hans-Meyer-Straße) darstellen. Die Semantik-Muster mussten so weit ausdifferenziert werden, bis die richtigen Inhalte eindeutig identifiziert werden konnten.

Durch weitere Spezifikation von Bedingungen und Musterdefinitionen konnten diese an die jeweiligen Inhalte der Datenbestände angepasst und in analoger Weise bei der Aufbereitung aller Merkmale angewandt werden.

Durch das Parsing konnte erreicht werden, dass in allen Dateien die gleichen Zusammenführungs-Variablen enthalten sind. Um diese Variablen in die gleiche Form zu bringen, müssen sie standardisiert werden. Durch die Standardisierung werden die Schreibweisen von Straßennamen, Hausnummern und Hausnummernzusätzen in den verschiedenen Datensätzen vereinheitlicht.

2.1.3 Standardisierung

Die Standardisierung erfolgt in einem mehrstufigen Verfahren, welches speziell auf die Datenquellen Georeferenzierte Adressdaten Bund, Melderegister und Bundesagentur für Arbeit abgestimmt wurde und jeweils auf den gleichen Standardisierungs-Regeln basiert. Zu diesen Regeln gehört etwa das Entfernen von Sonderzeichen und Zahlen, sofern diese keinen Teil des Straßennamens darstellen. Ein weiteres Beispiel ist die Vereinheitlichung der Schreibweise von Teilwörtern in einem String (STRAßE, Strasse, STRAß usw. zu STR).

Ein Indiz für die Qualität eines Registers ist die Reduktion der Daten nach der Standardisierung und anschließenden Gruppierung der Datenbestände über den Straßennamen in Verbindung mit dem Amtlichen Gemeindeschlüssel. Liegen die Daten in uneinheitlicher Schreibweise vor, greift die Standardisierung entsprechend stark und damit reduzieren sich die Datenbestände bei einer Gruppierung deutlich. Je stärker die Reduktion der Daten, desto geringer ist daher die Qualität der Ausgangsdaten. Die Register Georeferenzierte Adressdaten Bund und Melderegister wiesen bereits bei der Lieferung eine hohe Qualität auf. Die Standardisierung und Gruppierung (über Amtlichen Gemeindeschlüssel und Straßenname) führte nur zu einer geringfügigen Reduzierung der Datensätze um 1,1 % bei den MR- und ebenfalls 1,1 % bei den GAB-Daten. Bei den Daten der Bundesagentur für Arbeit hingegen lagen viele unterschiedliche Schreibweisen vor, sodass ein Großteil der Schreibweisen vereinheitlicht werden musste. Im Vergleich zum unstandardisierten Datenbestand auf der Ebene Amtlicher Gemeindeschlüssel/STR reduzierte sich die Anzahl der Datensätze dadurch um 54,5 %.

Bei Straßennamen, die unvollständig, veraltet oder falsch sind, kann auch eine Standardisierung nicht weiterhelfen. Beim Zusammenführen würden die veralteten oder nicht mehr existenten Angaben nicht zu Treffern in den anderen Datenbeständen führen. Daher muss im nächsten Schritt des Pre-Processings die Datenqualität verbessert werden.

2.2 Datenbereinigung

In diesem Schritt soll vor allem die Datenqualität – das heißt die Vollständigkeit, Korrektheit und Aktualität der Angaben – verbessert werden. Hierzu werden die Registerdaten mit externen Datenbeständen oder weiteren Registern zusammengeführt und abgeglichen, um die Zusammenführungs-Variablen, sofern notwendig, zu korrigieren. Die Datenaktualität des Amtlichen Gemeindeschlüssels wurde durch den Abgleich mit dem Gemeindeverzeichnis-Informationssystem erreicht. Die Aktualisierung von Straßennamen konnte teilweise anhand von Listen über Straßenumbenennungen maschinell durchgeführt werden. Dass die Zusammenführungs-Variablen korrekt sind, wurde durch den Abgleich mit den verschiedenen Registern und der PostDirekt-Datei gewährleistet. Hierunter fallen das Auflösen von Abkürzungen und die Bereinigung von Straßennamen.

2.2.1 Datenaktualität

Untersuchungen der drei Datenbestände ergaben, dass Merkmale, deren Ausprägungen über die Zeit veränderbar sind, wie zum Beispiel der Amtliche Gemeindeschlüssel oder die Straßennamen, in den verschiedenen Datenbeständen unterschiedlich aktuelle Stände aufweisen. Infolge von Gebietsstandsänderungen oder Straßenumbenennungen, die manche Bundesländer besonders stark betreffen, können veraltete Angaben dann zunächst nicht mit den entsprechenden Datensätzen aktuellerer Datenquellen zusammengeführt werden. Damit eine Zusammenführung auf der Basis von Anschriften ein Maximum an richtigen Treffern erzielt, ist es notwendig, dass den Datenbeständen gleiche bzw. ähnliche Bezugszeitpunkte zugrunde liegen. Die Register wurden zwar alle zum gleichen Stichtag an das Statistische Bundesamt geliefert, dennoch mussten die Bezugszeiträume der Merkmale Amtlicher Gemeindeschlüssel und Straßenname zwischen den Datenquellen angeglichen werden. Die Angaben zum Amtlichen Gemeindeschlüssel wurden aus dem Gemeindeverzeichnis-Informationssystem GV-ISys maschinell aktualisiert.

Für die Aktualisierung von Straßennamen, die in jedem Jahr im ganzen Bundesgebiet geändert werden, hat sich bisher jedoch noch keine vollständig IT-gestützte Lösung gefunden. Während diese Umbenennungen im Melderegister zeitnah erfasst werden, werden sie in die Vermessungsdaten zeitverzögert eingearbeitet. Im Register der Bundesagentur für Arbeit werden sie gar nicht erfasst. Erfolgte beispielsweise

die Meldung des Arbeitgebers vor dem Zeitpunkt der Straßenumbenennung, bleibt die Adresse unverändert. Da kein vollständiges Register deutscher Straßen mit aktuellen Straßennamen existiert, kann kein maschineller Abgleich wie bei der Aktualisierung des Amtlichen Gemeindeschlüssels vorgenommen werden. Ein Teil der Umbenennungen konnte jedoch in die Datenbestände maschinell eingearbeitet werden, wenn den Statistischen Ämtern der Länder Listen der Straßenumbenennungen vorlagen. Für die übrigen Fälle war eine manuelle Korrektur des Straßennamens durch die Statistischen Ämter der Länder vorgesehen.

2.2.2 Abkürzungen

Insbesondere im Register der Bundesagentur für Arbeit sind vielfältige Schreibweisen der Straßennamen vorhanden, die häufig auch Abkürzungen enthalten. Problematisch ist, dass die Vielfalt an Abkürzungen unpaarige Straßen erzeugt, da sich die Strings in den unterschiedlichen Registern voneinander unterscheiden. Da die Abkürzungen zudem nicht einheitlich gestaltet sind, können sie nicht automatisiert erkannt werden; somit kann keine generelle Regel zum Ausschreiben von bestimmten Zeichenkombinationen getroffen werden. Daraus folgend musste ein Kennzeichen für eine Abkürzung gefunden werden, durch das es zumindest möglich wurde, einen Teil der Abkürzungen zu identifizieren und mithilfe anderer Datenbestände zu bereinigen.

Der Punkt stellt in den meisten Strings des Merkmals Straßenname eine Abkürzung dar. Daher wurde er als eine Art Platzhalter ausgewählt, der die Stelle des Strings kennzeichnet, an der Zeichen fehlen. Diese wurden mithilfe einer anderen Datei, die bereits einen hohen Qualitätsstandard hat (und die Straßen daher als vorhanden angesehen werden), aufgefüllt. Um keine falschen Treffer zu erzeugen, durften Zeichen nur aufgefüllt werden, wenn genau ein Gegenstück bei der Zusammenführung gefunden wurde. In diesem Sinn kann die Art der Zusammenführung als deterministische Zusammenführung bezeichnet werden, denn es wird davon ausgegangen, dass die Identität aufgrund der Eineindeutigkeit gegeben ist.

2.2.3 Bereinigung der Straßennamen

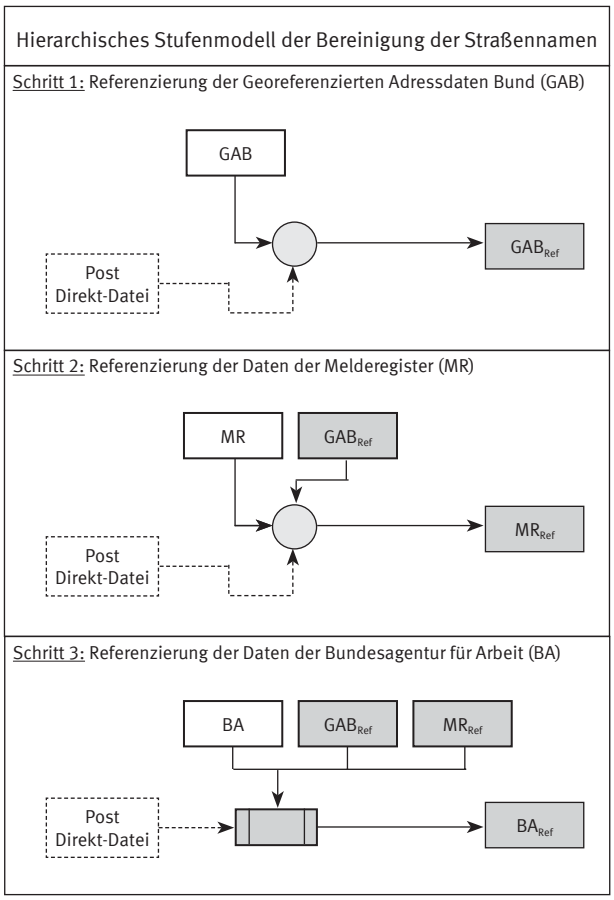
Um das Ergebnis der bevorstehenden Zusammenführung der Anschriften aus den drei Datenquellen zu optimieren, wurde ein weiterer Schritt zur Datenbereinigung des Merkmals Straßenname vorgenommen. Dabei wurde die Schreibweise des Straßennamens korrigiert sowie die Straße zur richtigen Gemeinde zugeordnet. Die Bereinigung der Straßennamen erfolgte wie in Schaubild 5 dargestellt in einem Stufenprozess, der aus drei Hauptschritten besteht. Die Ergebnisse der einzelnen Schritte waren die standardisierten und referenzierten Datenbestände der Georeferenzierten Adressdaten Bund (GAB), der Melderegister (MR) und der Bundesagentur für Arbeit (BA).

Schritt 1: Referenzierung der Straßennamen der GAB-Datei mit der PostDirekt-Datei sowie den Straßenverzeichnissen großer Städte

Im ersten Schritt wurden die GAB-Straßennamen mit den Straßennamen der PostDirekt-Datei referenziert. Dabei

Schaubild 5

Hierarchisches Stufenmodell der Bereinigung der Straßennamen



wurde zunächst zu jedem Straßennamen in der GAB-Datei (Grunddatenbestand) maschinell ein entsprechender Straßenname in der PostDirekt-Datei (Referenzdatenbestand) gesucht. Straßennamen, die in der GAB-Datei, aber nicht in der PostDirekt-Datei vorlagen, wurden in einem zweiten manuellen Schritt in den Statistischen Ämtern der Länder einzeln auf Richtigkeit geprüft. Bei der manuellen Überprüfung zogen die Statistischen Ämter der Länder weitere externe, unabhängige Datenquellen heran, wie beispielsweise Straßenverzeichnisse einzelner Gemeinden sowie öffentlich zugängliche Internet-Straßenverzeichnisse.

Da die GAB- und die PostDirekt-Datei in sich weitgehend konsistente Datenbestände darstellen, wurde ausschließlich die Methode der deterministischen Datenzusammenführung verwendet. Im bundesweiten Durchschnitt wurden lediglich 3 % der Straßennamen aus der GAB-Datei nicht in der PostDirekt-Datei gefunden und mussten somit manuell geprüft werden. Die Zahl der zu prüfenden Straßennamen betrug bundesweit etwa 36 000 (von 1,15 Mill. Straßennamen insgesamt). Tabelle 1 gibt einen nach Bundesländern aufgegliederten Überblick über die Zahl der zu prüfenden Fälle.

Schritt 2: Referenzierung der Straßennamen aus der MR-Datei mit der GAB-Datei

Die Referenzierung der Straßennamen aus der MR-Datei erfolgte anhand der im ersten Schritt bereinigten GAB-Datei.

Tabelle 1: Zu prüfende Straßennamen aus dem Datenbestand der Georeferenzierten Adressdaten Bund (GAB) im Rahmen der Referenzierung mit der PostDirekt-Datei

Bundesland	Straßen insgesamt	Zu prüfende Straßen	Anteil zu prüfender Straßen an Straßen insgesamt
	Anzahl		%
Baden-Württemberg	178 723	14 145	7,9
Bayern	226 931	4 509	2,0
Berlin	9 415	73	0,8
Brandenburg	36 331	1 236	3,4
Bremen	5 103	386	7,6
Hamburg	7 651	20	0,3
Hessen	89 110	2 622	2,9
Mecklenburg-Vorpommern	24 053	1 510	6,3
Niedersachsen	140 254	2 831	2,0
Nordrhein-Westfalen	182 960	2 029	1,1
Rheinland-Pfalz	74 614	1 592	2,1
Saarland	14 810	501	3,4
Sachsen	48 035	943	2,0
Sachsen-Anhalt	34 828	1 345	3,9
Schleswig-Holstein	43 704	1 491	3,4
Thüringen	34 957	996	2,8
Deutschland	1 151 479	36 229	3,1

Die Vorgehensweise entsprach der unter Schritt 1 beschriebenen Referenzierung der GAB-Datei, wobei hier die MR-Datei als Grunddatenbestand und die GAB-Datei als Referenzdatei diente.

Die Ergebnisse des Abgleichs der Straßennamen aus der MR-Datei mit den Straßennamen aus der referenzierten GAB-Datei sind in der Tabelle 2 nach Bundesländern differenziert dargestellt.

Tabelle 2: Zu prüfende Straßennamen aus dem Datenbestand der Melderegister (MR) im Rahmen der Referenzierung mit der GAB-Datei

Bundesland	Straßen insgesamt	Zu prüfende Straßen	Anteil zu prüfender Straßen an Straßen insgesamt
	Anzahl		%
Baden-Württemberg	163 289	3 474	2,1
Bayern	222 639	3 161	1,4
Berlin	10 117	165	1,6
Brandenburg	37 260	2 056	5,5
Bremen	4 572	38	0,8
Hamburg	7 607	135	1,8
Hessen	89 596	2 655	3,0
Mecklenburg-Vorpommern	23 774	1 724	7,3
Niedersachsen	135 257	2 502	1,8
Nordrhein-Westfalen	190 417	3 016	1,6
Rheinland-Pfalz	76 323	3 825	5,0
Saarland	14 366	101	0,7
Sachsen	48 280	934	1,9
Sachsen-Anhalt	35 013	1 804	5,2
Schleswig-Holstein	44 021	2 551	5,8
Thüringen	37 085	2 124	5,7
Deutschland	1 139 616	30 265	2,7

Die Menge der manuell korrigierten Straßennamen aus der MR-Datei betrug rund 30 000 (2,7 %) von 1,14 Mill. Straßennamen in der MR-Datei. Die Korrekturen wurden in die MR-Datei eingepflegt, bevor der dritte und letzte Schritt der Datenbereinigung durchgeführt wurde.

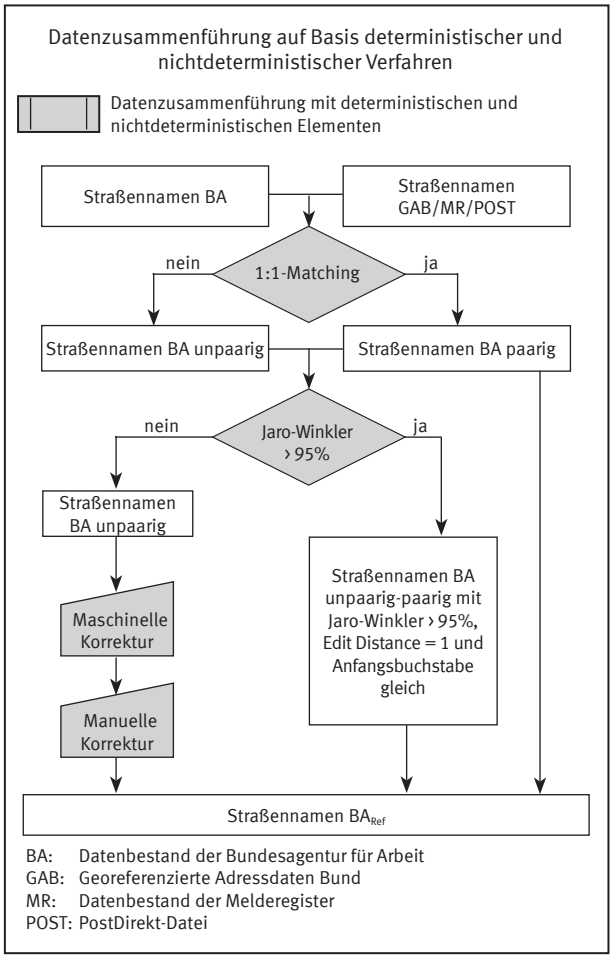
Schritt 3: Referenzierung der Straßennamen aus der BA-Datei mit sich selbst sowie mit der MR- und der GAB-Datei

Im Gegensatz zu den MR- und GAB-Straßennamen bestand in der BA-Datei eine breite Vielfalt an Schreibweisen ein und desselben Straßennamens. Untersuchungen zeigten jedoch, dass in der überwiegenden Zahl der Fälle neben vielen falsch erfassten Schreibweisen auch die richtige Schreibweise enthalten war. So entstand die Idee, innerhalb der BA-Datei mithilfe der richtigen Schreibweise eines Straßennamens alle zugehörigen falschen Schreibweisen zu korrigieren. Hierzu war die Abgrenzung der richtigen Schreibweise eines Straßennamens von allen falschen Schreibweisen notwendig. Die richtigen Straßennamen definierte man als diejenigen Straßen, die in mindestens einem weiteren Datenbestand (GAB, MR oder PostDirekt) zu finden waren. Danach erfolgte die Referenzierung der Gruppe nicht korrekter Straßennamen mit der Gruppe richtiger Straßennamen über Ähnlichkeitsfunktionen. Der gesamte Prozess ist in Schaubild 6 dargestellt. Die unpaarigen BA-Straßennamen wurden über die Ähnlichkeitsfunktionen Jaro-Winkler- und Levenshtein-Distanz mit den paarigen/korrekten Straßennamen der jeweils selben Gemeinde abgeglichen. Die automatisierte maschinelle Zuordnung und Korrektur eines falschen Straßennamens zu einem richtigen erfolgte dann, wenn alle folgenden Bedingungen erfüllt waren:

- Jaro-Winkler-Distanz größer 0,95
- Levenshtein-Distanz gleich 1
- Anfangsbuchstaben gleich
- Unpaarigem Straßennamen konnte genau ein richtiger Straßenname zugeordnet werden

Übrig blieb eine immer noch sehr hohe Zahl von unpaarigen Straßennamen, sodass diese Restmengen nicht vollständig zur manuellen Korrektur weitergegeben wurden. Diejenigen Straßennamen unter ihnen, zu denen ähnliche richtige Straßennamen mit der Bedingung Jaro-Winkler größer als 90 % gefunden wurden, wurden samt den ähnlichen Straßennamen an die Statistischen Ämter der Länder zur manuellen Korrektur weitergegeben. Bei diesen Korrekturarbeiten war das Heranziehen externer Referenzdatenbestände nicht notwendig. Hier galt es, Fall für Fall die Entscheidung zur treffen, ob die gegebene Ähnlichkeit mit einer richtigen Schreibweise die Gleichheit der beiden Straßen bedeutete. Die auf diese Weise korrigierbaren Straßennamen gingen in die Menge der referenzierten BA-Straßennamen ein. Die nicht korrigierbaren Fälle gingen in die Restmenge der unpaarigen Straßennamen ein. Die Menge der von den Statistischen Ämtern der Länder manuell korrigierten Straßennamen umfasste 25 % der Gesamtzahl der BA-Straßennamen. Aufgrund der eingeschränkten Datenqualität der BA-Datei war die verbleibende Restmenge noch so groß, dass von einer manuellen Korrektur durch die Statistischen Ämter der Länder abgesehen wurde und stattdessen die Restmenge der unpaarigen Straßennamen durch die Einbeziehung transitiver Beziehungen bei Paarvergleichen in das maschinelle Verfahren weiter reduziert wurde.

Schaubild 6



Die in diesem Kapitel dargestellten Verfahren zur Datenaufbereitung und -bereinigung haben die Qualität der Daten so weit verbessert, dass mit der Zusammenführung auf Anschriftenebene begonnen werden konnte.

3 Aufbau des vorläufigen Anschriften- und Gebäuderegisters

Ziel des Aufbaus des Anschriften- und Gebäuderegisters ist es, alle bewohnten Unterkünfte und Gebäude mit Wohnraum in Deutschland zu erfassen. Das vorläufige Anschriften- und Gebäuderegister wird auf Basis der Zusammenführung der GAB-, MR- und BA-Datenlieferungen vom April 2008 erstellt. Da die einzelnen Register selbst fehlerhaft sind, kann nicht vorausgesetzt werden, dass die enthaltenen Angaben auch tatsächlich existente Anschriften bezeichnen und vollständig zusammengeführt werden können. Bei der Integration entstehen daher Teilmengen, denen eine unterschiedliche Qualität unterstellt wird. Diese Teilmengen lassen sich in drei Kategorien einteilen. Die sich aus dieser Einteilung ergebenden Schnittmengen sind in Schaubild 7 grafisch zusammengefasst. Die erste, mit hellem Raster dargestellte Kategorie enthält Datensätze, die nur in einem Datenbestand vorkommen und mit keinem Datensatz aus einem anderen Register zusammengeführt werden konnten. Diese Kategorie besteht aus den Teilmengen die mit den Ziffern 5, 6 und 7 gekennzeichnet sind. Die zweite Kategorie umfasst Datenbestände, die in zwei Registern enthalten sind und zusammengeführt werden konnten (Ziffern 2, 3 und 4). Zu dieser Kategorie könnten zum Beispiel zwei Straßen gehören, die bei den Georeferenzierten Adressdaten Bund und im Datenbestand der Melderegister auftauchen, aber nicht in dem der Bundesagentur für Arbeit. Die dritte Kategorie setzt sich aus Datensätzen zusammen, die in allen drei Registern vorhanden sind (Ziffer 1).

Schaubild 7

Als „gültig“ werden Wohnanschriften gewertet, die in mindestens zwei Datenquellen vorkommen, das heißt die in Schaubild 7 mit den Nummern 1 bis 4 bezeichneten Mengen. Anschriften, die beim maschinellen Abgleich in nur einem Register gefunden wurden, werden einer manuellen Prüfung unterzogen.

Die Schnittmengen lassen eine qualitative Einordnung der Angaben in den Registern zu, für die jeweils unterschiedliche Arbeitsschritte notwendig werden.

Um die Qualität einer Anschrift darzustellen, wurden Qualitätskennzeichen eingeführt. Diese beschreiben, bis zu welchem Grad die jeweilige Anschrift mit einer Anschrift aus einer weiteren Datenquelle zusammengeführt werden konnte. Die Qualitätskennzeichen veranschaulichen jeweils die Zusammenführung zwischen den Datenbeständen Georeferenzierte Adressdaten Bund – Melderegister und Melderegister – Bundesagentur für Arbeit bzw. Georeferenzierte Adressdaten Bund – Bundesagentur für Arbeit. Sie bestehen aus einer fünfstelligen Zeichenkette, wobei jede Stelle der Zeichenkette ein Anschriftenmerkmal darstellt, wie das folgende Beispiel zeigt:

Beispiel des Qualitätskennzeichens

Amtlicher Gemeindeschlüssel	Postleitzahl	Straße	Hausnummer	Hausnummernzusatz
A	A	A	A	A

Ist eine Stelle der Zeichenkette mit einem „A“ belegt, so ist das entsprechende Merkmal bei beiden betroffenen Datenbeständen identisch. Steht dort ein „O“, so konnte die Ausprägung des Merkmals nur in einer Datenquelle gefunden werden. So beschreibt beispielsweise die Ausprägung „AAAAO“ des Qualitätskennzeichens Georeferenzierte Adressdaten Bund – Melderegister eine GAB-Anschrift, die über den Amtlichen Gemeindeschlüssel, die Postleitzahl, Straße und Hausnummer, aber nicht über den Hausnummernzusatz mit einer MR-Anschrift zusammengeführt werden konnte. Sind zwei Anschriften aus zwei Quellen über alle fünf Anschriftenmerkmale identisch, so wird die höchste Qualität „AAAAA“ vergeben. Dies gilt auch für identische Anschriften, die in beiden Datenbeständen keinen Hausnummernzusatz besitzen.

Das Kennzeichen dient ebenfalls dazu, anzuzeigen, bei welchen Merkmalen der Anschrift möglicherweise eine Korrektur vorgenommen werden muss. Ist das Qualitätskennzeichen Georeferenzierte Adressdaten Bund – Melderegister einer Anschrift mit „AAAOO“ belegt, sind also Amtlicher Gemeindeschlüssel, Postleitzahl und Straßennamen sowohl im Datenbestand Georeferenzierte Adressdaten Bund als auch im Datenbestand Melderegister vorhanden, so kann man mit hoher Wahrscheinlichkeit davon ausgehen, dass der Fehler bei der Hausnummer der Anschrift zu suchen ist.

Tabelle 3 stellt beispielhaft für die GAB-Datei neben der Gesamtzahl der Anschriften die Zahl der Prüffälle nach der Zusammenführung mit dem Datenbestand Melderegister dar. Die Fallzahlen der unpaarigen Anschriften sind insgesamt und nach Qualitätskennzeichen differenziert dargestellt.

Tabelle 3: Zu prüfende GAB- sowie MR-Anschriften im vorläufigen Anschriften- und Gebäuderegister

Prüffälle Qualitätskennzeichen	Anschriften im AGR.vorläufig	
	Anzahl	%
Insgesamt	21 363 164	*100*
Prüffälle aus den Georeferenzierten Adressdaten Bund		
AAAOO	2 176 198	*10,2*
AOAOO	9 952	*0,0*
OAAAO	70 337	*0,3*
OAAOO	5 015	*0,0*
OOOOO	207 285	*1,0*
Zusammen ...	2 468 787	*11,6*
Prüffälle aus dem Datenbestand Melderegister		
Zusammen ...	139 025	*0,7*

Rund 11 % der GAB-Anschriften konnten nicht in der MR-Datei gefunden werden. Der hohe Anteil ist darauf zurückzuführen, dass die GAB-Datei alle Anschriften – einschließlich der Anschriften gewerblich genutzter und unbewohnter Gebäude – enthält, während die MR-Datei nur die Anschriften bewohnter Gebäude abbildet. Beim Großteil dieser unpaarigen Anschriften handelte es sich um Anschriften der Kategorie „AAAOO“, das sind Anschriften, die bis zur Straßenebene auch in der MR-Datei zu finden waren, für die sich jedoch keine passende Hausnummer und kein passender Hausnummernzusatz in der MR-Datei finden ließen.

Die Zahl der MR-Anschriften, die in keiner weiteren Datenquelle zu finden waren, umfasste mit knapp 140 000 Anschriften rund 0,7 % der Anschriften bundesweit.

Die Statistischen Ämter der Länder überprüfen gemäß § 14 ZensG 2011 bei Anschriften, die in das Anschriften- und Gebäuderegister ausschließlich aufgrund von Angaben einer der drei Quellen aufgenommen wurden, ob es sich dabei um Anschriften von Gebäuden mit Wohnraum handelt. Das Ergebnis der Datenzusammenführung der Register ist das vorläufige Anschriften- und Gebäuderegister. Die bei der manuellen Überprüfung der offenen Anschriften festgestellten Wohnanschriften werden bis zum 30. Juli 2010 in das vorläufige Anschriften- und Gebäuderegister eingestellt.

4 Ausblick

Der Großteil der drei für den Zensus genutzten Registerbestände konnte erfolgreich aufbereitet und auf Anschriftenebene zusammengeführt werden. Es gibt Anschriften aus dem Datenbestand der Bundesagentur für Arbeit, die bislang nicht endgültig ins Anschriften- und Gebäuderegister aufgenommen werden konnten und weiterverarbeitet werden. Darüber hinaus läuft ebenfalls die manuelle Prüfung von maschinell nicht zu bearbeitenden Anschriften in Kooperation mit den Statistischen Ämtern der Länder. Zudem werden besondere Korrekturmassen, wie Pseudo-Identische-Anschriften, bearbeitet. Hierunter fallen bundesweit etwa 40 000 Anschriften, die u. a. aufgrund von Eingemeindungen entstanden sind. Für diese Anschriften wird derzeit ein Bearbeitungsverfahren auf Basis von Geokoordinaten entwickelt. Im weiteren Verlauf wird das Anschriften- und Gebäuderegister im Wesentlichen entsprechend den oben dargestellten Vorgehensweisen vervollständigt und mit Update-Lieferungen der Einzelregister sowie Informationen aus dem Sonderanschriftenregister und der Gebäude- und Wohnungszählung bis zum Zensusstichtag aktuell gehalten. Damit wird die Aufbauphase des Anschriften- und Gebäuderegisters in der Zensusvorbereitung abgeschlossen.

Um die eingangs (Abschnitt 1.2) genannten Aufgaben im Rahmen der Durchführung des Zensus erfüllen zu können, wird vor allem die Entwicklung der Schnittstellen zu den jeweiligen Teilerhebungen bei künftigen Arbeiten im Vordergrund stehen.

Auszug aus Wirtschaft und Statistik

Herausgeber: Statistisches Bundesamt, Wiesbaden

Schriftleitung: Roderich Egeler
Präsident des Statistischen Bundesamtes
Verantwortlich für den Inhalt:
Brigitte Reimann,
65180 Wiesbaden
- Telefon: +49 (0) 6 11/75 20 86
- E-Mail: wirtschaft-und-statistik@destatis.de

Vertriebspartner: SFG Servicecenter Fachverlage
Part of the Elsevier Group
Postfach 43 43
72774 Reutlingen
Telefon: +49 (0) 70 71/93 53 50
Telefax: +49 (0) 70 71/93 53 35
E-Mail: destatis@s-f-g.com

Erscheinungsfolge: monatlich

Allgemeine Informationen über das Statistische Bundesamt und sein Datenangebot erhalten Sie:
- im Internet: www.destatis.de

oder bei unserem Informationsservice
65180 Wiesbaden

- Telefon: +49 (0) 6 11/75 24 05
- Telefax: +49 (0) 6 11/75 33 30
- www.destatis.de/kontakt