Dipl.-Sozialwissenschaftlerin Eva-Maria Diehl

Methoden der Mehrfachfallprüfung im Zensus 2011

Die kommunalen Melderegister sind die zentrale Datenquelle des Zensus 2011. Für die Ermittlung der Einwohnerzahlen mit dem Zensus 2011 ist es notwendig, diejenigen Daten, die für den Zensus von den Melderegistern an die Statistischen Ämter der Länder übermittelt wurden, daraufhin zu untersuchen, ob es mehrfache Anmeldungen ein und derselben Person gibt und ob diese melderechtlich korrekt vorgenommen wurden. Ziel der Einwohnerermittlung ist es, jede Person in Deutschland nur einmal zu zählen, und zwar an ihrer Hauptwohnung beziehungsweise ihrer alleinigen Wohnung. Die mit dem Zensus 2011 ermittelten neuen Einwohnerzahlen haben weitreichende Konsequenzen. Sie wirken sich auf die Verteilung der Steuereinnahmen auf Bund, Länder und Gemeinden, die Einteilung der Wahlkreise, die Sitzverteilung im Bundesrat und vieles mehr aus. Deshalb ist es äußerst wichtig zu ermitteln, wer am Zensusstichtag, dem 9. Mai 2011, unter welcher Anschrift mit Hauptwohnung oder alleiniger Wohnung gemeldet war.

In diesem Beitrag wird dargestellt, wie beim Zensus 2011 die potenziellen Melderegisterdubletten ermittelt wurden und wie die statistische Bereinigung der erkannten Dubletten erfolgte. Mit diesen Schritten konnte der Wohnungsstatus für jede Person an einer Anschrift festgelegt und damit die Hauptwohnung beziehungsweise alleinige Wohnung bestimmt werden.

1 Sinn und Zweck der Mehrfachfallprüfung

Die ermittelten Einwohnerzahlen haben direkten Einfluss auf den Länderfinanzausgleich, den kommunalen Finanzausgleich, die Einteilung der Wahlkreise sowie das Stimmgewicht der Bundesländer im Bundesrat. Auch vonseiten der Europäischen Union sind die Qualitätsanforderungen für Bevölkerungsdaten sehr hoch. So wird in Artikel 11 Absatz 5 der Geschäftsordnung des Rates der Europäischen Union das Verfahren bei der Beschlussfassung mit qualifizierter Mehrheit im Europäischen Rat festgelegt.[1] Ein Kriterium ist, dass die Mitgliedstaaten, die eine qualifizierte Mehrheit bilden, mindestens 62 % der Gesamtbevölkerung der Europäischen Union repräsentieren.

Ziel des registergestützten Zensus 2011 ist es, die Einwohnerzahlen zum Zensusstichtag 9. Mai 2011 zu ermitteln. Die wichtigste Datenquelle hierfür sind die Melderegister, welche in Deutschland gemeindeweise geführt werden. Bei solch einer dezentralen Führung ist nicht auszuschließen, dass Personen in mehreren Gemeinden gleichzeitig mit mehr als einer alleinigen Wohnung oder Hauptwohnung oder ausschließlich mit (einer oder mehreren) Nebenwohnung(en) gemeldet sind. Da die Anmeldungen in den aufnehmenden Meldebehörden nicht zeitgleich mit der Verarbeitung der elektronischen Abmeldungen in den abgebenden Meldebehörden stattfinden, kommt es zu zeitlichen Verzögerungen (Time-Lags). Davon zu unterscheiden sind mehrfache Eintragungen zum Wohnsitz bei einer Haupt- und einer oder mehreren Nebenwohnungen, die laut Melderecht ebenso zulässig sind wie eine einzige Eintragung zum Wohnsitz bei einer alleinigen Wohnung. Dubletten (zwei Personendatensätze, die für dieselbe Person stehen) müssen also nicht grundsätzlich problematisch sein, erschweren aber die Verarbeitung insofern,

1 Amtsblatt der EU Nr. L 319 vom 20. Oktober 2004, Seite 16.

als unzulässige von zulässigen Mehrfachanmeldungen zu unterscheiden sind.

Um unzulässige Mehrfachanmeldungen aufdecken zu können, mussten die von den Melderegister führenden Stellen an die statistischen Ämter übermittelten Daten der verschiedenen Meldebehörden zunächst in eine Gesamtdatei und formal auf einen Standard gebracht werden. Dabei war davon auszugehen, dass die Angaben aus den Melderegistern nicht immer „korrekt" sind, also Tippfehler, falsche oder ungenügende Angaben und so weiter vorliegen können. Über verschiedene Standardisierungen der Namensfelder wurde die Grundlage für die Mehrfachfallprüfung geschaffen (siehe Abschnitt 2.1). Nach Abschluss dieser Schritte galt es, diejenigen Datensätze zu bestimmen, welche potenziell identisch sein könnten. Hierbei wurden Verfahren der Ähnlichkeitssuche angewendet (siehe Abschnitt 2.2). Es ergaben sich potenzielle Dublettenpaare, welche zunächst maschinell und eine kleine Teilmenge davon auch manuell auf Paarigkeit geprüft wurden (siehe Abschnitt 2.3). Die als zusammengehörig identifizierten Paare wurden primärstatistisch und/oder maschinell aufgelöst, indem die Annahme „Dublette" entweder verworfen oder bestätigt wurde. Letztendlich waren nur Personen an der Hauptwohnung beziehungsweise alleinigen Wohnung relevant für die Ermittlung der Einwohnerzahlen. Ziel der Mehrfachfallprüfung im Zensus 2011 war es also festzustellen, wer zum Zensusstichtag (9. Mai 2011) unter welcher Anschrift mit Hauptwohnung beziehungsweise alleiniger Wohnung gemeldet war.

Da die Melderegister führenden Stellen prozessorientiert arbeiten, der Zensus 2011 aber stichtagsbezogen durchgeführt wurde, mussten Vorkehrungen getroffen werden, um die Meldeverhältnisse zum Zensusstichtag abbilden zu können. Der Zensustest im Jahr 2001 hat in diesem Zusammenhang deutlich gezeigt, dass es sinnvoll ist, mit zwei zeitlich versetzten Meldedatenabzügen zu arbeiten.[2] Beim Zensus 2011 erfolgte der erste Meldedatenabzug zum 9. Mai 2011 (MR.1), also dem Zensusstichtag, der zweite zum 9. August 2011 (MR.2). Aus letzterem hatten nur solche Personendatensätze für die zu ermittelnde Bevölkerungszahl eine Relevanz, deren Ereignishorizont bis einschließlich Stichtag reichte. Geburten, Um- und Zuzüge, die nach dem 9. Mai 2011 erfolgt sind, waren nicht stichtagsrelevant.

2 Vorgehen bei der Mehrfachfallprüfung

Zentrales Ziel der Mehrfachfallprüfung war es, solche Personen zu ermitteln, die mehrmals in den Registern erfasst waren, sowie diejenigen Personen, die ausschließlich mit Nebenwohnsitz gemeldet waren. In einem ersten Schritt wurden doppelt erfasste Personen gesucht (Dublettensuche). Bei der Dublettensuche handelte es sich datentechnisch um einen Vergleich des Datenbestandes mit sich selbst, bei dem das Ziel verfolgt wurde, doppelte Datensätze zu kennzeichnen, die zur gleichen Person gehörten. Es war dabei unerheblich, welchen Wohnungsstatus die Personen in den einzelnen Datensätzen aufwiesen. Dabei konnte es vorkommen, dass Elemente der gefundenen Paare auch mit anderen Datensätzen eine Dublette bildeten. Die dadurch entstandenen Dublettenpaare wurden dann jedoch weiterhin auf Pärchenebene verarbeitet, auf Transitivität geprüft (wenn Person A und Person B als identisch identifiziert werden, Person B und Person C ebenfalls, so müssen auch Person A und Person C als identisch betrachtet werden) und entschieden. Nach Feststellung der doppelt/mehrfach geführten Personen wurden die unterschiedlichen Ausprägungen des Wohnungsstatus in die Prüfung einbezogen, um entscheiden zu können, ob die Datensätze für Personen standen, die

- ausschließlich mit mehreren Nebenwohnungen,
- mehrfach mit Hauptwohnung,
- mehrfach mit alleiniger Wohnung oder
- mit Hauptwohnung und alleiniger Wohnung gemeldet waren.

Darüber hinaus wurden solche Personen betrachtet, die nur mit Nebenwohnsitz ohne korrespondierenden Hauptwohnsitz im Datenbestand enthalten waren.

Schaubild 1 stellt den schematischen Ablauf der Mehrfachfallprüfung in ihren Teilschritten dar.

Im Vorfeld der Mehrfachfallprüfung wurden die in den Meldedaten geführten Namensfelder mit unterschiedlichen Verfahren standardisiert. Hierauf wird im nächsten Abschnitt eingegangen.

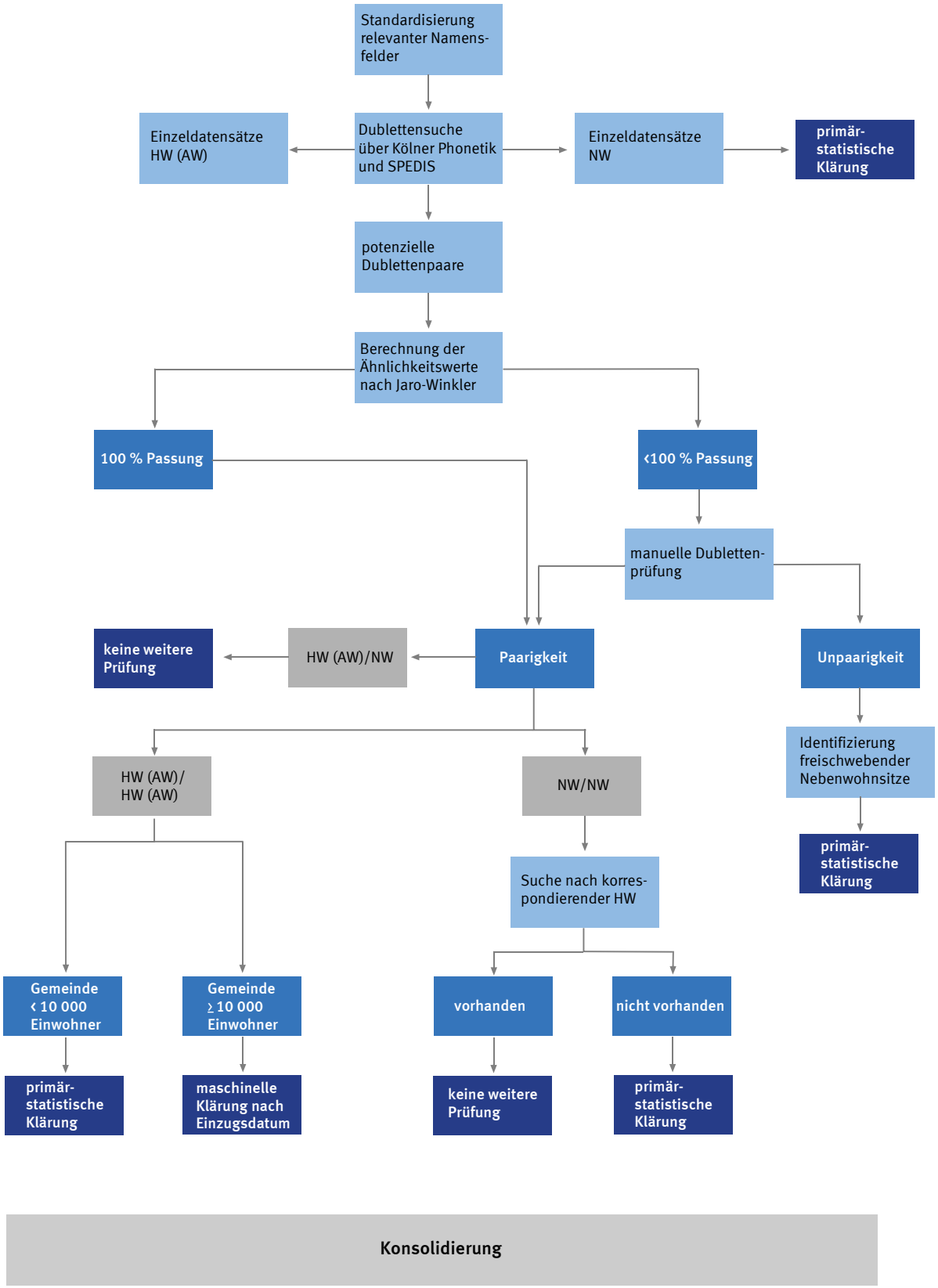
2.1 Standardisierung der Namensfelder

In den meisten Gesellschaften werden Personen Namen gegeben. In Verbindung mit dem Geburtsdatum und dem Geschlecht ist der Name oftmals ein Alleinstellungsmerkmal. Namen können im Laufe eines Lebens geändert werden. In Deutschland bleibt der bei der Geburt zugewiesene Name in der Regel unverändert und ist in den Meldedaten enthalten. Die Meldebehörden sind zudem dazu angehalten, die formalen Vorgaben zu Geburtsnamen, Familiennamen, Vornamen und so weiter für alle sich anmeldenden Personen anzuwenden, sodass auch Personen aus Kulturkreisen mit anderen Namensgebungstraditionen in den Melderegistern entsprechend abgebildet werden. Unbefüllte Geburtsnamensfelder (in der Regel ledige oder männliche Personen) wurden mit dem Inhalt des Familiennamensfeldes befüllt, da anzunehmen war, dass bei diesem Personenkreis der Geburts- und der Familienname identisch sind.

Bei der Mehrfachfallprüfung wurde somit zwar von der Annahme ausgegangen, dass es bestimmte Grundmerkmale gibt, die sich bei einem Menschen nach der Erstaufnahme in das Melderegister in der Regel nicht ändern. Dabei handelt es sich um den Geburtsnamen, das Geburtsdatum, den Vornamen, das Geschlecht und den Geburtsort. Gleichwohl war davon auszugehen, dass bei der Aufnahme der unveränderlichen Merkmale unterschiedlich vorgegangen wird

2 Siehe dazu Michel, B.: „Ergebnisse der Mehrfachfallprüfung im Rahmen des Zensustests" in WiSta 9/2004, Seite 979 ff., sowie Statistische Ämter des Bundes und der Länder: „Ergebnisse des Zensustests" in WiSta 8/2004, Seite 813 ff., hier: Seite 817 f.

Schaubild 1 Ablauf der Mehrfachfallprüfung im Zensus 2011



HW = Hauptwohnung; AW = alleinige Wohnung; NW = Nebenwohnung.

2012 - 01 - 0389

und die Namensangaben voneinander abweichen können. Dies kann aufgrund von unterschiedlichen Transkriptionen von einem ausländischen Sprachraum in den deutschen, Schreibfehlern, fehlerhaften oder unvollständigen Registereintragungen und Ähnlichem entstehen. Um zu vermeiden, dass Personen mit mehreren Wohnungen durch solche Vorkommnisse nicht als zusammengehörig identifizierbar waren, war es also notwendig, Standardisierungen durchzuführen. Entsprechendes galt für die Geburtsorte.[3]

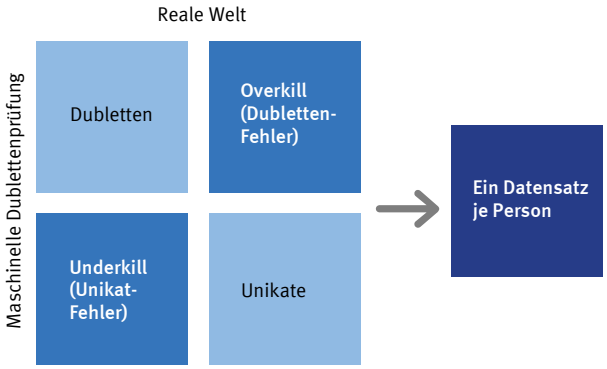
Um die Vornamens- und Geburtsnamensfelder zu standardisieren, war es in einem ersten Arbeitsschritt notwendig, Sonderzeichen zu entfernen und damit auch Hochkommata zu eliminieren, sowie Diakritika[4] auf ihren Stammbuchstaben zurückzuführen, da diese in früheren Jahren im Meldewesen nicht zulässig waren und einer zeitlich konsistenten Schreibweise im Wege standen. Des Weiteren wurden „Ä“, „Ö“ und „Ü“ sowie „ß“ durch „AE“, „OE“, „UE“ und „SS“ ersetzt. Diese Standardisierungen wurden in einem „gedoppelten“ Eingabefeld durchgeführt, sodass die originäre Schreibweise in den Namensfeldern erhalten blieb. In einem nächsten Schritt wurden alle für sich stehenden Worte mit einem Großbuchstaben beginnend, gefolgt von Kleinbuchstaben dargestellt, um die unterschiedliche Handhabung von Groß- und Kleinschreibungen in den Melderegistern auszugleichen. Ebenso wurden aus Gründen der Vereinheitlichung Bindestriche aus den oben genannten Namensfeldern entfernt. Zusätzlich wurden bestimmte im Namen mitgeführte Bestandteile wie „genannt“ oder „verheiratete“ vereinheitlicht. Einträge wie „Vatersname“ oder auch „Eigenname“ wurden herausgenommen. Abkürzungen wie „v.“, „u.“ wurden zu „von“ beziehungsweise „und“ umgesetzt. Die Entfernung mehrfacher und führender Leerzeichen schloss sich an. Eine weitere Form der Standardisierung lag in der Entfernung von Adelstiteln aus dem Namensfeld und ihrer Übertragung in das dafür vorgesehene Feld der Namensbestandteile. Ähnliches galt für akademische Titel. Namenszusätze wie „Junior“ und „Senior“ und ihre Abkürzungsformen wurden entfernt, da sie die erste Stufe des Abgleichs behinderten. Im Unterschied hierzu wurden Namenszusätze, beispielsweise von asiatischstämmigen Personen, beibehalten.

Mit der Durchführung dieser Schritte wurden neue Datenfelder angelegt, in denen alle nun standardisierten Vor- und Geburtsnamen getrennt voneinander aufsteigend alphabetisch sortiert wurden, aber auch ein Feld, welches zusätzlich eine Kombination aus standardisierten Vor- und Geburtsnamen enthielt. Diese neuen Datenfelder bildeten die Grundlage für die Dublettensuche und die Berechnung der verschiedenen Ähnlichkeitswerte, auf die später eingegangen wird.

2.2 Dublettensuche

Bei einer Dublettenprüfung können sich – wie in Schaubild 2 dargestellt – zwei Arten von Fehlern zeigen. Zum einen kann Personen die Dubletteneigenschaft fälschlicherweise zugeschrieben werden, das heißt es werden zu viele vermeintliche Paare gefunden. Solch ein Verfahren würde einen Dubletten-Fehler (Overkill) erzeugen. Zum anderen kann ein Unikat-Fehler (Underkill) entstehen, bei dem Personen, die mehr als einmal im Datenbestand vorhanden sind, nicht als Dubletten gefunden werden. Somit liegt die Aufgabe einer „guten“ Dublettensuche darin, diese beiden Fehler gleichzeitig zu minimieren. Hierfür wurden die an die Statistischen Ämter der Länder übermittelten Daten aus den Melderegistern zu einem zentralen Datenbestand im Statistischen Bundesamt zusammengefasst.[5]

Schaubild 2 Fehlerquellen bei der Dublettensuche



2012 - 01 - 0390

Um sicherzugehen, dass alle potenziellen Dubletten gefunden werden konnten, wurden, wie in den Abschnitten 2.2.1 und 2.2.2 beschrieben, zwei voneinander unabhängige Verfahren zu deren Identifizierung eingesetzt (phonetische Algorithmen und SPEDIS-Verfahren). Damit nicht zu viele, sehr unwahrscheinliche Dubletten geprüft werden mussten, wurde anschließend ein Verfahren angewendet, anhand dessen sehr unähnliche Paare ausgesteuert werden konnten (Jaro-Winkler-Funktion) – hierauf wird in Abschnitt 2.2.3 eingegangen.

2.2.1 Phonetische Algorithmen

Mit der „Kölner Phonetik“, die auf die relevanten Zusammenführungsvariablen angewandt wurde, können Text-Daten auf Ähnlichkeit geprüft werden. Diese Methode stellt einen phonetischen Algorithmus dar, der Worten anhand ihres Sprachklanges einen phonetischen Code (Zeichenfolge) zuweist. Das Ziel besteht darin, identisch klingenden Worten den gleichen Code zuzuordnen, um darüber eine Ähnlichkeitssuche zu ermöglichen. Der Vorteil der „Kölner Phonetik“ liegt im Unterschied zu anderen Verfahren, wie beispielsweise SOUNDEX, darin, dass sie speziell auf die deutsche Sprache abgestimmt ist.[6]

Die „Kölner Phonetik“ weist jedem Buchstaben eines Wortes eine Zahl zwischen „0“ und „8“ zu. Für die Wahl der ent-

3 In einer späteren Ausgabe dieser Zeitschrift (voraussichtlich im August 2012) wird ein Beitrag zum Thema „Auswertung der Geburtsorte im Zensus 2011“ veröffentlicht werden.

4 Diakritika sind an Buchstaben angebrachte kleine Zeichen wie Punkte, Striche, Häkchen oder Kreise.

5 Siehe dazu Michel, B. (Fußnote 2), hier: Seite 980 f.

6 Siehe hierzu auch Schürle, J.: „Record Linkage – Zusammenführung von Daten auf Basis des Modells von Fellegi und Sunter“ , Frankfurt 2004, derselbe: „Automatisierte Zusammenführung von Daten – Das Modell von Fellegi und Sunter“ in WiSta 4/2005, Seite 406 ff., sowie Postel, H. J.: „Die Kölner Phonetik. Ein Verfahren zur Identifizierung von Personennamen auf Grundlage der Gestaltungsanalyse.“, IBM-Nachrichten/19. Jahrgang/1969, Seite 925 ff.

sprechenden Ziffer wird maximal ein nebenstehender Buchstabe als Bezug verwendet. Einen Sonderfall stellen Anlaute (Wortanfänge) dar. Für diese gelten gesonderte Regeln. Zusätzlich werden Doppelbuchstaben auf einen Buchstaben reduziert. Um möglichst alle Konstellationen in der Datenzusammenführung abgreifen zu können, werden die Namen in allen Kombinationen gespeichert. Abschließend werden führende Nullen und mehrere aufeinanderfolgende gleiche Codes entfernt.

Als Ergebnis entstehen Zahlencodes für jedes berechnete Eingabefeld. Diese berechneten Eingabefelder werden in einem weiteren Schritt gegeneinander abgeglichen, wobei übereinstimmende Paarbildungen einen Hinweis auf eine Grundähnlichkeit darstellen. So kann durch einen kombinatorischen Vergleich die Suchmasse der potenziellen Dubletten stark eingeschränkt werden.[7]

Beispiel: Ausgehend von den beiden Namen Lars Wolter und Lars Walther ergibt sich nach Standardisierung und Sortierung der für die „Kölner Phonetik“ verwendeten Felder ebenfalls Lars Wolter und Lars Walther. Es lassen sich für die beiden Namen folgende Codes ermitteln:

Lars Wolter – 5078305207; nach Streichung der Nullen bleibt 5783527 übrig.

Lars Walther – 50783052-07; auch hier ergibt sich im Endeffekt 5783527.

Das Beispiel zeigt, dass durch die Kölner Phonetik auch Namen (Personendatensätze) paarig erscheinen, die offensichtlich nur mit einer geringen Wahrscheinlichkeit die gleiche Person repräsentieren. Aus diesem Grund ist die Identifikation von paarigen Datensätzen nur über die „Kölner Phonetik“ nicht ausreichend, um die Eigenschaft einer Dublette zu belegen. Sie eignet sich aber durchaus dazu, aus einem großen Datenbestand potenziell paarige Datensätze herauszufinden, die sich dann anhand weiterer Wahrscheinlichkeitsverfahren, wie nachfolgend dargestellt, qualitätssichern lassen.

Ferner wurde das SOUNDEX-Verfahren für den angloamerikanisch geprägten Personenkreis genutzt, womit auch asiatische Namen wie Yu, die durch die „Kölner Phonetik“ nicht ausreichend bewertet werden, angemessen ausgewertet wurden.

7 Um auszuschließen, dass durch die Einschränkung der Suchmasse etwaige Dubletten nicht erkannt wurden, wurde die nachfolgend beschriebene SPEDIS-Variante auch auf den Gesamtbestand angewendet.

2.2.2 SPEDIS-Verfahren

Der mithilfe der „Kölner Phonetik“ ermittelte maximale (paarige) Datenbestand wurde in einem weiteren Schritt auf einen „potenziellen Datenbestand“ reduziert. Die Dublettensuche durch SPEDIS unterscheidet sich von der nach der „Kölner Phonetik“ dadurch, dass hier kein Code für die einzelnen Buchstaben berechnet wird, sondern „Kosten“ ermittelt werden, um eine Zeichenkette (String) in eine andere zu überführen (auch Austauschkosten genannt). So wird berechnet, wie „teuer“ es ist, einen String (zum Beispiel Name) in einen anderen zu überführen. Es ist also potenziell möglich, jedes Wort in ein anderes zu überführen. Aus diesem Grund müssen hier sehr strenge Regeln gesetzt werden, die es im Anschluss an die Berechnung erlauben, von mit hoher Wahrscheinlichkeit bestehenden Dubletten zu sprechen (Näheres zu den Ähnlichkeitswerten siehe Kapitel 5).

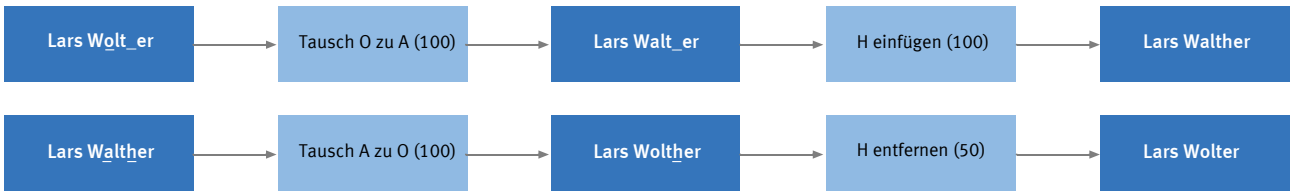
Um den endgültigen SPEDIS-Wert zu erhalten, werden die „Kosten“ für die benötigten Veränderungen des Abfragestrings in den Zielstring aufsummiert und durch die Länge des Zielstrings dividiert. Hierdurch werden die anfallenden Kosten im Verhältnis zur Stringlänge gewichtet. Erforderlich ist es auch, Abfrage- und Zielstring zu tauschen und hierfür erneut die „Kosten“ zu ermitteln.

Am Beispiel aus Abschnitt 2.2.1 ergeben sich bei wechselseitiger Betrachtung der beiden Datensätze unterschiedliche „Kosten“ (siehe hierzu auch Schaubild 3). Wählt man die Abfolge von Lars Wolter im Vergleich zu Lars Walther, ergibt sich ein SPEDIS-Wert von 16. Da Lars übereinstimmt, müssen keine Transformationen vorgenommen werden. „Wolter“ in „Walther“ zu überführen bedarf jedoch eines Austausches des „o“ durch „a“ und das Einfügen des „h“, was „Gesamtkosten“ von 200 ergibt. Teilt man diese durch die Länge des Zielstrings 12, erhält man 16. Wird das Ganze umgekehrt, erhält man den Wert 13, da das „a“ wieder durch das „o“ ersetzt werden muss und ein „h“ entfernt wird. Somit ergibt sich ein Wert von 150 für die „Kosten“, welche in diesem Beispiel durch 11 dividiert werden.[8]

Um besser beurteilen zu können, ob die durch SPEDIS gefundenen Paare tatsächlich potenzielle Dubletten waren, wurden die SPEDIS-Werte auf einen Wert im Intervall zwischen 0 und 1 normiert. 0 bedeutete in diesem Fall, dass die Strings komplett unterschiedlich waren, 1 hingegen, dass es sich mit sehr hoher Wahrscheinlichkeit um paarige Strings handelte. Da das SPEDIS-Verfahren nicht symmetrisch arbei-

8 Siehe hierzu auch die SAS OnlineDoc 9.2 unter dem Suchbegriff SPEDIS function.

Schaubild 3 Beispiel einer SPEDIS-Wert-Berechnung



2012 - 01 - 0391

tet und die Reihenfolge der in die Berechnung gegebenen Abfragestrings das Ergebnis beeinflusst, wurden für einen Abgleich immer zwei SPEDIS-Werte ermittelt.

2.2.3 Jaro-Winkler-Funktion

Um zu bestätigen, ob die Dubletteneigenschaft mit hoher Wahrscheinlichkeit vorliegt, gaben die normierten SPEDIS-Werte einen ersten gesicherten Anhaltspunkt. Da dies den im Zensus 2011 gesetzten Qualitätsansprüchen noch nicht genügte, wurde zusätzlich die Wahrscheinlichkeit der Paarigkeit auf Grundlage der Jaro-Winkler-Funktion ermittelt. Dieses Verfahren berücksichtigt speziell die Transpositionen einzelner Buchstaben und ist vor allem für kurze Wörter geeignet; außerdem gewichtet es Übereinstimmungen am Wortanfang höher als die am Wortende. Die Jaro-Winkler-Funktion vergibt normierte Werte zwischen 0 und 1, wobei 0 einen kompletten Unterschied angibt (Müller verglichen mit Schmitt) und 1 eine perfekte Übereinstimmung darstellt. Jaro-Winkler-Werte, die unter einem Schwellenwert von 0,75 lagen (was einer Ähnlichkeit von weniger als 75 % entspricht), erwiesen sich als überwiegend unpaarig.

Die Berechnung der Jaro-Winkler-Werte funktioniert wie folgt:

Die Jaro-Winkler-Distanz[9] ist eine Weiterentwicklung des Jaro-Distanzmaßes[10] mit der Formel

$$\Phi = W1 \times c/d + W2 \times c/r + Wt \times (c-\tau)/c,$$

wobei

W1 = frei wählbare Gewichtung der Zeichen im ersten der beiden Datensätze (standardmäßig 1/3)

W2 = frei wählbare Gewichtung der Zeichen im zweiten der beiden Datensätze (standardmäßig 1/3)

Wt = frei wählbare Gewichtung der Transposition (standardmäßig 1/3)

d = Länge der Zeichenkette im ersten Datensatz

r = Länge der Zeichenkette im zweiten Datensatz

τ = Anzahl der Transpositionen des Zeichens

c = Anzahl der gemeinsamen Zeichen im Zeichenketten-Paar

Hierbei ist zu beachten, dass der Ergebniswert = 0 ist, wenn c = 0 ist.

Das Ergebnis der Berechnung laut oben genannter Formel ist sodann Bestandteil der eigentlichen Jaro-Winkler-Berechnung, wonach

$$\text{Jaro-Winkler} = \Phi + [lp(1-\Phi)],$$

wobei

l = Länge des übereinstimmenden Präfix bis maximal zur Länge 4

p = Konstante (üblicherweise 0,1)

Somit bedeutet das Beispiel von Lars Wolter gegenüber Lars Walther für den Jaro-Wert Folgendes:

$$\Phi = 1/3 \times 10/11 + 1/3 \times 10/12 + 1/3 \times (10-1)/10$$

$$\Phi = 0{,}603$$

Wird dies in die Formel zur Jaro-Winkler-Berechnung eingefügt, so resultiert das nachstehende Endergebnis:

$$\text{Jaro-Winkler} = 0{,}603 + [4 \times 0{,}1 \times (1 - 0{,}603)]$$

$$\text{Jaro-Winkler} = 0{,}7618$$

Dieses Vorgehen wurde auf den zusammengefassten String „Vor- und Geburtsname standardisiert und aufsteigend alphabetisch sortiert“ und drei Variablen mit jeweils einer Namensinformation angewandt. Bei den einzelnen Strings handelte es sich um den standardisierten Vornamen, den standardisierten und aufsteigend alphabetisch sortierten Vornamen sowie den standardisierten Geburtsnamen. Hierfür wurden jeweils beide Dublettenelemente einbezogen.

2.3 Dublettenprüfung

Nachdem auf Grundlage der Ergebnisse der phonetischen Algorithmen, SPEDIS und Jaro-Winkler potenzielle Dublettenpaare ermittelt worden waren, erfolgte die Qualitätssicherung der Dubletten als wesentlicher Teil der Mehrfachfallprüfung. Diese bestand in zwei aufeinander aufbauenden Arbeitsschritten, einem maschinellen Abgleich und einer manuellen Sichtprüfung.

2.3.1 Maschinelle Qualitätssicherung

Wie oben gezeigt, konnten aufgrund sehr geringer Ähnlichkeiten (identifiziert durch SPEDIS und Jaro-Winkler) potenzielle Mehrfachfälle aufgetrennt und somit von vornherein als nicht identisch klassifiziert werden. Im Gegensatz dazu war es jedoch auch möglich, Paarungen von Anfang an als identisch zu klassifizieren. Dies war immer dann der Fall, wenn die Grundmerkmale Geschlecht, Geburtsdatum, Geburtsort, Vorname und Geburtsname bereits in den Rohdaten beziehungsweise nach der Standardisierung vollständig übereinstimmten, was einem Jaro-Winkler-Wert von 1 in allen standardisierten Feldern entsprach.[11] Hierzu wurden die potenziellen Dublettenpaare horizontal in einem Daten-

9 Siehe hierzu Winkler, W. E.: “String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage”, Proceedings of the Section on Survey Research Methods, American Statistical Association/1990, Seite 354 ff., sowie Winkler, W. E.: “Overview of Record Linkage and Current Research Directions”, Research Report Series, 2006 (*www.census.gov/srd/papers/pdf/rrs2006-02.pdf*; abgerufen am 13. Juni 2012).

10 Siehe Jaro, M. A.: “Advances in Record Linkage Methodology as Applied to the 1985 Census of Tampa, Florida”, Journal of the American Statistical Society 84 (406)/1989, Seite 414 ff.

11 Eine kleine Ausnahme bildeten dabei die Geburtsorte. Diese wurden auch als identisch betrachtet, wenn sich in einem Feld, welches sich aus mehreren Ortsbezeichnungen zusammensetzte, komplette Strings (also vollständige Worte), die eine eindeutige Identifizierung erlaubten, wiederfanden. Dies wurde so beispielsweise für Frankfurt gegenüber Frankfurt am Main umgesetzt, wobei immer beachtet werden muss, dass für die Personendatensätze ohnehin Geburtsdatum und Geschlecht als Blockingvariablen fungierten und die restlichen Grundmerkmale mindestens mit hoher Wahrscheinlichkeit zu ein und derselben Person gehörten.

satz abgebildet, sodass die maschinellen Prüfungen effizient und übersichtlich erfolgen konnten.

2.3.2 Manuelle Qualitätssicherung

Zur Überprüfung der nicht zu 100 % übereinstimmenden Dublettenpaare wurden diese Prüffälle einschließlich ihrer Ähnlichkeitswerte in eine eigens hierfür erstellte Anwendung eingelesen. Diese Anwendung ermöglichte es, Parameterwerte vorzugeben, durch die Dublettenpaare als identisch zu klassifizieren waren. Paarungen, die dieser Grundlage nicht voll entsprachen, wurden in die manuelle Inaugenscheinnahme geschickt. Hier stellte die Anwendung den Bearbeiterinnen und Bearbeitern nicht nur die Grundmerkmale der Mehrfachfallprüfung zur Verfügung, sondern zusätzlich alle weiteren relevanten Merkmale, auf deren Grundlage sich eine eindeutige Entscheidung treffen ließ, ob es sich um eine Dublette handelte oder nicht. Hierzu zählten unter anderem der Familienstand, das Eheschließungs-/Scheidungsdatum, Namensbestandteile, Informationen zur derzeitigen und Zuzugsanschrift, zum Ehe-/Lebenspartner und zu den Kindern. Lagen hier eindeutige Informationen vor, dass es sich um dieselbe Person handelt, wurden die beiden Personendatensätze als einem Individuum zugehörig (paarig) bewertet.

Eindeutige Hinweise auf Paarigkeit zweier Personendatensätze lagen vor, wenn zusätzlich zu den beiden Blockingvariablen (Variablen, bei denen zwischen den beiden betrachteten Dublettenelementen keine Unterschiede vorliegen dürfen) Geburtsdatum und Geschlecht mehrere der folgenden Merkmale übereinstimmten:

- Familienstand „Verheiratet“ & Eheschließungsdatum & Merkmale des Ehegatten
- Familienstand „Verheiratet“ & Eheschließungsdatum unbefüllt & Merkmale des Ehegatten
- Familienstand „Geschieden“ oder „Verwitwet“ & Datum der Beendigung der letzen Ehe
- Namen und Geburtsdaten der Kinder
- Identischer Geburtsnummerneintrag
- Vom Geburtsnamen abweichender identischer Familienname bei erwachsenen Personen (bei Kindern könnte es sich beispielsweise um eine Adoption von Zwillingen handeln)
- Vorname und Geburtsname sind identisch, wenn man sie dreht (beispielsweise Vorname Vanh und Geburtsname Thi bei Person 1 und umgekehrt bei Person 2). Dies ist jedoch nur bei Namen aus dem asiatischen Raum zuverlässig anwendbar.
- Die Zuzugsanschrift der einen Person entspricht der „aktuellen“ Anschrift der anderen Person.

Diese Informationen stellten jedoch nur Indizien für eine Übereinstimmung von zwei Personendatensätzen dar, entscheidend war das Gesamtbild. So konnte es beispielsweise sein, dass nur bei einem der beiden Personendatensätze Kinder eingetragen waren, weil diese nach der Anmeldung bei der zuständigen Meldebehörde geboren und erfasst wurden, bei der anderen Meldebehörde die Geburt jedoch nicht eingetragen wurde, da die Weitergabe der Information nicht erfolgte (dieses Phänomen findet man fast ausschließlich an Nebenwohnungen, aber auch bei älteren, augenscheinlich inaktiven Meldedatensätzen).

3 Sonderfälle bei der Mehrfachfallprüfung

Wie in vielen Bereichen der Statistik stellte sich auch bei der Mehrfachfallprüfung im Zensus 2011 das Problem, dass mit herkömmlichen Methoden zwar ein Großteil der interessierenden Fälle ermittelt werden konnte, dass es jedoch auch einige Spezialfälle von Personendatensätzen gab, welche nur mit besonderen Mitteln identifiziert und behandelt werden konnten. Nachfolgend werden diese dargestellt.

3.1 Uneindeutiges Geburtsdatum

Im Normalfall ist das Geburtsdatum, welches im Zusammenhang mit der Geburt attestiert wurde, das amtliche Geburtsdatum. In seltenen Ausnahmen ist es jedoch nicht auszuschließen, dass Personen ihren Geburtstag ändern lassen. Dies ist beispielsweise möglich, wenn die entsprechenden Personen aus der Türkei stammen.[12] Sie haben die Möglichkeit, sich in ihrem Heimatland einmal im Leben ein neues Geburtsdatum zuweisen zu lassen. Die Begründung dafür ist, dass in vielen ländlichen Gebieten prinzipiell der Erste eines Monats als Tag der Geburt verzeichnet wird oder Kinder erst lange Zeit nach der Geburt gemeldet werden und die Eltern sich an das genaue Datum bis dahin nicht mehr erinnern können. Dies kann später korrigiert werden.[13]

Etwas anders stellt sich der Fall für Personen aus Griechenland dar. Dort wurden im Bürgerkrieg zwischen 1946 und 1949 in erheblichem Umfang Urkunden zerstört, sodass viele vor und in dieser Zeit geborene Personen keinen amtlichen Nachweis mehr über ihr Geburtsdatum hatten und es nachträglich eintragen lassen mussten.

Bei Migrantinnen und Migranten kann es vorkommen, dass sie ihr Geburtsdatum gar nicht kennen und bei ihrer ersten Anmeldung bei einer deutschen Meldebehörde ihr Alter schätzen beziehungsweise eine solche Schätzung von einer Mitarbeiterin oder einem Mitarbeiter des Amtes vorgenommen wird, wenn die Angaben der Antragstellerin oder des Antragstellers offensichtlich unplausibel sind. Die melderechtlichen Regelungen sehen in solch einem Fall vor, dass

12 Türkisches Zivilgesetzbuch IV. Berichtigung, 1. Allgemein, Artikel 39; Türkisches Personalstandsgesetz Artikel 25 (1) über das Meldeverfahren; Türkisches Verwaltungsgesetz Nummer 15926, Artikel 174 vom 3. Juli 1995.

13 In den nächsten Jahrzehnten sollte sich diese Problematik jedoch zumindest für Personen, die in der Türkei geboren wurden, auflösen. Seit dem 29. April 2006 ist es dort gesetzlich vorgeschrieben, ein noch nicht festgelegtes Geburtsdatum von Amts wegen erfassen zu lassen. Ist nur das Geburtsjahr bekannt, ist es zwischenzeitlich so, dass der 1. Juli als entsprechendes Geburtsdatum festgelegt wird. Sind nur der Monat und das Jahr bekannt, so wird der 1. des Monats als Geburtsdatum eingetragen. Ist nichts bekannt, ist die Geschäftsleitung der entsprechenden Meldebehörde dazu berechtigt, ein Geburtsdatum festzulegen. Siehe hierzu das türkische Gesetz über die Aufgaben der Personenstandsbehörden, Artikel 39 (1) und (2).

der 1. Januar einzutragen ist und das geschätzte Geburtsjahr, welches das Alter des beziehungsweise der Zuziehenden treffen könnte.

Weiterhin ist es auch bei in Deutschland geborenen Personen möglich, dass sich schlichtweg Erfassungsfehler ergeben und diese zu einer späteren Zeit korrigiert werden müssen. Dabei handelt es sich meist um einen Unterschied von wenigen Tagen oder um Zahlendreher zwischen den beiden Datumsangaben.

Für die Mehrfachfallprüfung im Zensus 2011, welche auf dem Grundgedanken basierte, dass das Geburtsdatum ein unveränderliches Personenmerkmal darstellt, ergab sich das Problem, dass Personendatensätze der eben erwähnten Personengruppen gegebenenfalls nicht als Paare gefunden worden wären. Dies hätte eine direkte Auswirkung auf die Einwohnerzahlen gehabt. So war es notwendig, eine spezielle Vorgehensweise zu entwickeln, mit deren Hilfe auch potenzielle Dubletten, die dieses abweichende Merkmal aufwiesen, aufgedeckt werden konnten. Um dem gerecht zu werden, wurde diese Personengruppe mit einer Bandbreite von +/– einem Geburtsjahr untersucht, wobei besondere Anforderungen an die Paarigkeit der anderen persönlichen Merkmale gestellt wurden. Die Ergebnisse der Untersuchung wurden einer Sichtprüfung unterzogen, die ihrerseits wiederum qualitätsgesichert wurde.

3.2 Überprüfung des konsolidierten Datenbestandes

Der konsolidierte Datenbestand des Zensus 2011 enthält alle stichtagsrelevanten Datensätze (zum 9. Mai 2011), die durch die Melderegister führenden Stellen in den Datenlieferungen MR.1 und MR.2 an die statistischen Ämter übermittelt wurden. In § 3 Absatz 2 Zensusgesetz 2011[14] wurde neben den Datenlieferungen der Melderegister zu den Stichtagen 9. Mai 2011 (Berichtszeitpunkt) und 9. August 2011 zusätzlich eine Datenlieferung zum 1. November 2010 (MR.05) angeordnet, welche im konsolidierten Datenbestand ebenfalls mit den stichtagsrelevanten Datensätzen enthalten ist.

Es gilt zu beachten, dass letztlich jede in Deutschland gemeldete Person unter einer Anschrift nur mit einem Datensatz im konsolidierten Referenzdatenbestand enthalten sein dürfte. Der Referenzdatenbestand besteht primär aus der Verknüpfung des Anschriften- und Gebäuderegisters[15] mit denjenigen Daten, die von den Melde- und erwerbsstatistischen Registern[16] an die statistischen Ämter übermittelt wurden. Hierzu werden die drei Datenlieferungen (MR.1, MR.2 und MR.05) der Melderegister im Referenzdatenbestand abgelegt. Dieser Datenbestand wird im Laufe der Zensusdurchführung um die Ergebnisse der weiteren Erhebungsteile ergänzt. Der Referenzdatenbestand nimmt also dynamisch die Erhebungsergebnisse auf.

Aus unterschiedlichen Gründen (Beantragen eines neuen Personalausweises, wobei ein Fehler auffällt, Nachreichen von Unterlagen, Ändern von Leitbändern und so weiter) können Veränderungen in Personendatensätzen entstehen. Daher war es notwendig, den konsolidierten Datenbestand daraufhin zu untersuchen, inwieweit es hierdurch zu Doppelerfassungen im Referenzdatenbestand unter einer Anschrift gekommen sein könnte. Personen, die im Verlauf des Zensus 2011 konstant unter derselben Anschrift gemeldet waren, aber aus verschiedenen Gründen bei den Aktualisierungslieferungen neu in den Referenzdatenbestand aufgenommen worden waren, mussten über die oben genannten Dublettensuchverfahren identifiziert und anhand der Anschrift extrahiert werden, um sie gesondert zu betrachten. Häufig gab es Änderungen in der Schreibweise von Vor- oder Geburtsnamen (zum Beispiel Susanne gegenüber Suzanne) und/oder dem Geburtsort (Frankfurt wurde zu Frankfurt am Main), was zu einer Veränderung der personenbezogenen Ordnungsnummer (PON) führte. Diese Ordnungsnummer macht eine Person an einer Anschrift eindeutig identifizierbar. Sie leitet sich aus den – in der Regel – unveränderlichen Grundmerkmalen einer Person, also dem Geburtsnamen, Vornamen, Geburtsdatum, Geschlecht und Geburtsort, ab. Wenn sich eines dieser zur PON-Bildung herangezogenen Merkmale änderte, ergab dies eine neue personenbezogene Ordnungsnummer und das Datenbanksystem behandelte diesen Datensatz wie einen neuen, da die Integration der einzelnen Melderegisterdatenlieferungen in den Referenzdatenbestand über Anschrift und personenbezogene Ordnungsnummer erfolgte. Ziel war es, dass die personenbezogene Ordnungsnummer unter einer Anschrift über den gesamten Zensusverlauf konstant blieb. Dies war deshalb geboten, da personenbezogene Daten aus anderen Erhebungsteilen des Zensus 2011, zum Beispiel aus der Haushaltebefragung, über die personenbezogene Ordnungsnummer, welche in der ursprünglichen Datenlieferung MR.05 zum 1. November 2010 berechnet wurde, an den Referenzdatenbestand angebunden werden. Die betreffenden Datensätze mussten also in einem gesonderten Arbeitsschritt ermittelt werden.

Auch hierfür konnte die Anwendung zur Prüfung von Melderegisterdubletten eingesetzt werden, wobei sich bereits im Vorfeld maschinell viele Datensätze zuverlässig klären ließen, da – entgegen den ursprünglichen Annahmen – alle Datensätze der Melderegisterdatenlieferungen ein über die Zeit konstantes Ordnungsmerkmal der Person[17] (OMP) im jeweiligen Melderegister enthielten. Zusätzlich wurden alle an einer Anschrift zusammengeführten Datensätze daraufhin untersucht, ob es sich bei den zu prüfenden Paarungen tatsächlich um die gleiche Person handelte. Die für paarig befundenen Pärchen wurden anschließend zusammengeführt, sodass ein aktualisierter Datensatz entstand, welcher

14 Gesetz über den registergestützten Zensus im Jahre 2011 (Zensusgesetz 2011 – ZensG 2011) vom 8. Juli 2009 (BGBl. I Seite 1782).

15 Siehe hierzu Kleber, B./Maldonado, A./Scheuregger, D./Ziprik, K.: „Aufbau des Anschriften- und Gebäuderegisters für den Zensus 2011“ in WiSta 7/2009, Seite 629 ff.

16 Siehe hierzu §§ 4 und 5 Zensusgesetz 2011. Zum Jahresende 2012 soll in dieser Zeitschrift ein Artikel zur Erhebung, Aufbereitung und Zusammenführung der Erwerbsregister im Zensus 2011 erscheinen.

17 Die Melderegister enthalten aus technischen und organisatorischen Gründen der Datenhaltung Ordnungsnummern, die die Datensätze jeder Person kennzeichnen. Diese Ordnungsnummern werden auch verwendet, um Bezüge zwischen den Datensätzen von Personen (zum Beispiel bei verheirateten Personen oder bei Eltern-/Kind-Beziehungen) datentechnisch eindeutig darstellen zu können. Die Nutzung dieser Ordnungsnummern im Zensus 2011 vereinfacht es, Haushalts- und Familienzusammenhänge abzubilden.

aber die ursprüngliche personenbezogene Ordnungsnummer enthielt.

Besonderes Augenmerk musste in diesem Zusammenhang auf mögliche Paarungen gelegt werden, bei denen es sich um Zwillinge handeln konnte. Dabei wurde unter anderem von der Grundannahme ausgegangen, dass Zwillinge bei der gemeinsamen Anmeldung in der Meldebehörde nahe beieinanderliegende Ordnungsmerkmale der Personen (OMPs) zugewiesen bekommen. War dies der Fall, wurden der Geburtsname, das Geburtsdatum und die Anschriften-ID als Blockingvariablen verwendet. Bei numerisch ausgerichteten Ordnungsmerkmalen der Personen (OMPs) sollte zusätzlich eine Abweichung von +/– 1 vorliegen, jedoch durfte der Wert nicht +/– 10 überschreiten, um einen ersten Hinweis auf die Eigenschaft „Zwillingspaar" zu liefern. Diese potenziellen Zwillingsdubletten wurden sodann über weitere Merkmale (wie Geburtsnummerneinträge, Verzeigerungen[18] zu den Eltern) auf die Eigenschaft „Zwilling" hin untersucht. Hierbei kamen sowohl maschinelle Vergleichsverfahren (zum Beispiel Verzeigerungen zu den Eltern mussten identisch sein) als auch eine Inaugenscheinnahme zum Einsatz.

3.3 Mehrfachfallprüfung für Personen an Sonderanschriften

Die Erhebung an Sonderanschriften unterscheidet zwischen sensiblen und nicht sensiblen Sonderbereichen. Als sensibel gelten Sonderbereiche, bei denen die Information über die Zugehörigkeit für die Betroffenen zu einer sozialen Benachteiligung führen könnte. Hierzu zählen beispielsweise Justizvollzugsanstalten oder Behindertenwohnheime. Um den Personenkreis, der in sensiblen Sonderbereichen lebt, besonders zu schützen, wurde mit dem Bundesbeauftragten für den Datenschutz und die Informationsfreiheit eine spezielle Vorgehensweise erarbeitet. Ziel war es, die Hilfsmerkmale (siehe hierzu § 8 Absatz 1 Nr. 2 Zensusgesetz 2011) aus der Erhebung in sensiblen Sonderbereichen so früh wie möglich zu löschen und die für spätere Auswertungen vorzuhaltenden statistischen Informationen ausschließlich in anonymisierter Weise abzulegen.

Die Aufgabe der Mehrfachfallprüfung bestand darin, dem für die Erhebung in sensiblen Sonderbereichen zuständigen Projektteil zum frühestmöglichen Zeitpunkt diejenigen potenziellen Dublettenpaare zu übergeben, bei denen sich mindestens eine Person an einer sensiblen Sonderanschrift befand. Aus diesem Grund wurde beschlossen, die Mehrfachfallprüfung für die Wohnungsstatusfeststellung auf dem zuerst zur Verfügung stehenden Melderegisterbestand MR.05 durchzuführen. Die potenziellen Dubletten konnten damit bereits im Mai 2011 aufbereitet zur Verfügung gestellt werden.

Um die betreffenden Personen zu ermitteln, wurden Dublettenpaare aus dem Bestand MR.05 (Stichtag 1. November 2010) extrahiert, bei denen sich ein oder beide Dublettenelemente an einer sensiblen Sonderanschrift befanden. Zusätzlich wurden alle Einzeldatensätze von Personen, die ausschließlich an einer sensiblen Sonderanschrift gemeldet waren, dem zuständigen Projektteil übergeben. Im Anschluss wurde durch das Teilprojekt „Sonderbereiche" der endgültige Wohnungsstatus innerhalb und außerhalb der Sonderbereiche für angetroffene Personen festgelegt. Im Verlauf der Erhebung ergaben sich potenzielle Fehlbestände (eine Person befand sich laut Melderegister nicht an einer Sonderanschrift, wurde aber an einer solchen erhoben) und potenzielle Karteileichen (eine Person befand sich laut Melderegister in einem Sonderbereich, wurde dort jedoch nicht erhoben). Die vorläufigen Fehlbestände an sensiblen Sonderanschriften, auf Basis des Datenmaterials MR.05, wurden abschließend durch den Projektteil „Mehrfachfallprüfung" am aktuellen Referenzdatenbestand vorbeigeführt. Damit wurde geprüft, ob und gegebenenfalls wie diese Personen zum Zensusstichtag an und eventuell noch außerhalb der Sonderanschrift gemeldet waren. Diese Informationen flossen zeitnah an den Projektteil „Sonderbereiche" zurück, um dort abschließend verarbeitet zu werden.

Von einer Befragung im Rahmen der primärstatistischen Klärung, wie bei Personen an Nicht-Sonderanschriften, waren Bewohner von sensiblen Sonderbereichen wegen ihrer besonderen Datenschutzinteressen ausgeschlossen.

Im Falle von nicht sensiblen Sonderbereichen besteht im Unterschied zu den sensiblen Sonderbereichen kein besonderes Schutzbedürfnis für diejenigen Personen, die an solchen Anschriften gemeldet sind. Deshalb war eine vorzeitige Löschung der Hilfsmerkmale nicht nötig.

Potenzielle Dubletten, bei denen mindestens eine Person an einem nicht sensiblen Sonderbereich gemeldet war, wurden im Rahmen einer Mehrfachfallprüfung auf dem Referenzdatenbestand ermittelt. Dabei wurden sowohl Dubletten als auch nur an nicht sensiblen Sonderanschriften gemeldete Personen an den Projektteil „Sonderbereiche" übergeben. Für die sich durch die Erhebung ergebenden vorläufigen Fehlbestände erfolgte eine erneute Mehrfachfallprüfung im gesamten aktuellen Referenzdatenbestand. Ziel war es, zu identifizieren, ob und gegebenenfalls wie diese Personen zum Zensusstichtag an und eventuell noch außerhalb der Sonderanschrift gemeldet waren. Diese Informationen flossen zeitnah an den Projektteil „Sonderbereiche" zurück, um dort abschließend verarbeitet zu werden. In dessen Fachverfahren erfolgte die endgültige Feststellung des Personenkreises, der in Sonderbereichen lebt.

4 Primärstatistische Klärung

Nach Abschluss der maschinellen und manuellen Dublettenprüfung wurden die ermittelten Dublettenpaare (nur Hauptwohnungsdubletten) sowie Datensätze von Nebenwohnungen, die mit keiner Hauptwohnung verknüpft werden konnten, zur weiteren Klärung an die Statistischen Landesämter übergeben. Die Statistischen Ämter der Länder stellten für alle Personen, die nur mit Nebenwohnungen gemeldet waren, sowie für Personen mit mehr als einer alleinigen Wohnung oder Hauptwohnung, die in Gemeinden mit weniger als 10 000 Einwohnern gemeldet waren,

18 Verzeigerungen verweisen von einem Personendatensatz auf den Datensatz einer anderen Person, wie zum Beispiel des gesetzlichen Vertreters oder der Kinder. Sie werden im Zensus 2011 auch für die Haushaltegenerierung genutzt.

den Wohnungsstatus fest (§ 15 Absätze 3 und 4 Zensusgesetz 2011). Hierbei wurde sichergestellt, dass keine an einer sensiblen Sonderanschrift gemeldeten Personen in die Befragung gelangten. Nach Herausfilterung dieser Personen wurde eine Liste erstellt, mit deren Hilfe die Statistischen Ämter der Länder Fragebogen zur Ermittlung des Wohnungsstatus versenden konnten. Generell konnten die zuständigen Statistischen Ämter der Länder Unstimmigkeiten beim Wohnungsstatus auch durch Begehungen und somit Rückfragen an der zu klärenden Anschrift oder durch eine telefonische Befragung klären. Da die beiden letzten Alternativen jedoch einen hohen zeitlichen und personellen Aufwand erforderten, erwies sich die erste Methode in der Regel als die Methode der Wahl. Im Fragebogen waren der zu klärende Melderegistereintrag sowie Informationen zur rechtlichen Grundlage und zur Auskunftspflicht enthalten. Der beziehungsweise die Auskunftspflichtige konnte den Wohnungsstatus bestätigen, diesen ändern oder auch die richtige Hauptwohnung, die zum 9. Mai 2011 bestand, beziehungsweise weitere zum Stichtag bestehende Wohnungen angeben.

Nach Erhalt und Auswertung der zurückgesandten Befragungsunterlagen wurden die Befragungsergebnisse von den Statistischen Landesämtern über eine speziell dafür entwickelte IT-Anwendung an das Statistische Bundesamt übermittelt. Hier wurde anschließend nach letzten Plausibilisierungen die aktive Hauptwohnungsanschrift ermittelt (siehe hierzu Kapitel 5).

Hauptwohnungsdubletten, deren Anschriften sich ausschließlich in Gemeinden mit mindestens 10 000 Einwohnern befanden, wurden maschinell entschieden (§ 15 Absatz 2 Zensusgesetz 2011).

5 Konsolidierung im Rahmen der Mehrfachfallprüfung

Als abschließender Schritt der Mehrfachfallprüfung war es notwendig, die Befragungsergebnisse aus der primärstatistischen Erhebung mit dem Referenzdatenbestand zusammenzuführen und auf Stimmigkeit zu prüfen.

Bei diesem Arbeitsschritt wurde sichergestellt, dass nach Integration der Befragungsergebnisse keine unplausiblen Daten mehr im Referenzdatenbestand enthalten waren. Dabei war besonders zu beachten, dass stets den primärstatistischen Ergebnissen aus der Mehrfachfallprüfung – in Bezug auf Wohnungsstatus und Existenz der Person – Vorrang vor den im Melderegister vorliegenden Angaben eingeräumt wurde. Sie sind also per Definition richtig, selbst wenn sich dadurch Widersprüche zu den Meldedaten ergeben. So entstanden bei der Konsolidierung drei Datenmassen. Zum einen gab es Personendatensätze, welche primärstatistisch geklärt wurden. Weiter lagen solche vor, die maschinell geklärt wurden. Dabei handelte es sich um temporäre Dubletten sowie um Hauptwohnungsdubletten, bei denen sich beide Elemente in Gemeinden mit mehr als 10 000 Einwohnern befanden. Die dritte Datenmasse stellten korrekte Personendatensätze dar, die keiner weiteren Bereinigung bedurften.

5.1 Auswertung der Ergebnisse der primärstatistischen Klärung

Bei der primärstatistischen Klärung wurden nur Einzelpersonen an Anschriften befragt und nicht – wie in der Haushaltebefragung[19] – alle unter einer Anschrift lebenden Personen. Diese Befragungsergebnisse mussten bei der Auswertung der Rückläufe wieder zu Dubletten und Dublettenketten zusammengesetzt werden. Die Übersicht 1 „Fallkonstellationen des Befragungsergebnisses der primärstatistischen Klärung und ihre maschinellen Konsequenzen“ zeigt die theoretischen Rücklaufmöglichkeiten im Rahmen der primärstatistischen Klärung und deren Auflösung auf. Diese lassen sich entsprechend auf Dublettenketten übertragen, welche mehr als zwei Elemente enthalten.

Bei der primärstatistischen Klärung wurden Personen zu ihren Personendatensätzen befragt. Die Personendatensätze wurden im Zuge der Konsolidierung entsprechend dem erzielten Befragungsergebnis plausibilisiert und maschinell sowie manuell konform zu den vorliegenden Gesamtinformationen bei Unplausibilitäten korrigiert. Anschließend wurde die korrekte Ausprägung des Wohnungsstatus geprüft und gegebenenfalls korrigiert. In Fällen, in denen keine Antwort erfolgte, wurde der Wohnungsstatus aus dem Referenzdatenbestand übernommen und zur Weiterverarbeitung genutzt.

Abschließend wurde der endgültig zu vergebende Wohnungsstatus für jeden Personendatensatz festgelegt. So wurde eine Wohnung, die mit keiner weiteren Wohnung verbunden war, aber im Datenbestand als Hauptwohnung klassifiziert war, als alleinige Wohnung festgelegt. Im umgekehrten Fall musste eine als alleinige Wohnung gekennzeichnete Wohnung, die mit weiteren Nebenwohnungen verbunden werden konnte, als Hauptwohnung charakterisiert werden. Diese Informationen wurden an den entsprechenden Personendatensätzen abgelegt.

5.2 Auflösung von Dubletten in Gemeinden mit mehr als 10 000 Einwohnern und temporärer Dubletten

Parallel zur Verarbeitung der Rücklaufergebnisse aus der primärstatistischen Klärung wurden Personendatensätze aus Hauptwohnungsdubletten, welche sich ausschließlich in Gemeinden mit mehr als 10 000 Einwohnern befanden, maschinell geklärt. Hier wurde anhand der Einzugs- beziehungsweise Anmeldedaten, welche zu den entsprechenden Datensätzen hinterlegt waren, die aktuelle Anschrift ermittelt. Abhängig davon, ob diesen Dublettenelementen eine Nebenwohnung zugeordnet werden konnte, war auch hier abschließend die korrekte Ausprägung des Wohnungsstatus zu vergeben.

Gleichermaßen wurde mit temporären Dubletten verfahren, wobei hier die ältere Wohnung (mit dem älteren Einzugsdatum) als für den Zensus 2011 nicht relevant angesehen wurde. Als temporäre Dubletten wurden solche Dubletten

19 Siehe hierzu § 7 Zensusgesetz 2011.

Übersicht 1

Fallkonstellationen des Befragungsergebnisses der primärstatistischen Klärung und ihre maschinellen Konsequenzen[1]

Fallkonstellationen	Rücklauf primärstatistische Klärung	Maschinelle Konsequenz
HW(1) – HW(2)	Beide unzustellbar	Beide HW erhalten Löschkennzeichen.
	Nur HW(1) zustellbar, Antwort HW	HW(2) Löschkennzeichen, HW(1) bleibt HW.
	HW(1) ohne Antwort, HW(2) unzustellbar	
	Nur HW(1) zustellbar, Antwort NW	Variante A: HW(2) Löschkennzeichen, HW(1) gibt NW an, zugehörige HW wird gefunden. Ergebnis NW. Variante B: HW(2) Löschkennzeichen, HW(1) gibt NW an, zugehörige HW wird nicht gefunden. Ergebnis Löschkennzeichen.
	HW(1) keine Antwort, HW(2) antwortet NW	Variante A: HW(1) bleibt HW, HW(2) gibt NW an, zugehörige HW wird gefunden. Ergebnis NW. Variante B: HW(1) bleibt HW, HW(2) gibt NW an, zugehörige HW wird nicht gefunden. Ergebnis Löschkennzeichen.
	HW(1) antwortet HW, HW(2) antwortet NW mit anderer HW	
	Beide zustellbar, beide keine Antwort	Einzugsdatum unterschiedlich: Die HW(1) mit dem aktuelleren Einzugsdatum wird HW, die ältere HW(2) Löschkennzeichen. Einzugsdatum identisch, identische Anschrift: MR-Datensatz mit aktuellerem Existenzkennzeichen [HW(1)] bleibt bestehen. Einzugsdatum identisch, unterschiedliche Anschrift: Beibehaltung beider Datensätze.
	Zustellbar, eine Antwort (HW)	
	Beide antworten HW	
	Beide HW zustellbar, Antwort: HW und zugehörige NW	Keine maschinelle Konsequenz, da das Ergebnis korrekt für die Einwohnerzahlberechnung vorliegt.
	Beide HW zustellbar, beide antworten mit NW	Variante A: Für beide neuen NW wird HW gefunden. Beide werden NW. Variante B: HW(1) gibt NW an, zugehörige HW wird gefunden und bleibt NW. HW(2) gibt NW an, zugehörige HW wird nicht gefunden, Löschkennzeichen. Variante C: Für beide neuen NW wird keine HW gefunden. Beide Löschkennzeichen.
NW(1) – NW(2)	Beide kommen unzustellbar zurück	Beide erhalten Löschkennzeichen.
	NW(1) unzustellbar, NW(2) Antwort HW	NW(1) erhält Löschkennzeichen. NW(2) wird zu HW.
	NW(1) unzustellbar, NW(2) mit Antwort NW	Variante A: NW(1) erhält Löschkennzeichen, NW(2) bleibt NW, zugehörige HW wird gefunden. Variante B: NW(1) erhält Löschkennzeichen, NW(2) gibt NW an, zugehörige HW wird nicht gefunden, Löschkennzeichen.
	NW(1) unzustellbar, NW(2) ohne Antwort	
	Zustellbar, aber keine Antwort	Variante A: Für NW wird HW gefunden. Beide bleiben NW. Variante B: NW(1), zugehörige HW wird gefunden, bleibt NW. NW(2) gibt NW an, zugehörige HW wird nicht gefunden, Löschkennzeichen. Variante C: Für beide NW wird keine HW gefunden. Beide Löschkennzeichen.
	Beide zustellbar, NW(1) keine Antwort, NW(2) antwortet NW	
	Beide zustellbar, beide antworten NW	
	NW(1) keine Antwort, NW(2) antwortet HW	Keine maschinelle Konsequenz, da das Ergebnis korrekt für die Einwohnerzahlberechnung vorliegt.
	NW(1) antwortet HW, NW(2) NW	
	Beide zustellbar, beide antworten mit HW	Einzugsdatum unterschiedlich: Die angegebene HW(1) mit dem aktuelleren Einzugsdatum wird HW, die ältere angegebene HW(2) Löschkennzeichen. Einzugsdatum identisch, identische Anschrift: Angegebene HW mit aktuellerem Existenzkennzeichen bleibt bestehen. Einzugsdatum identisch, unterschiedliche Anschrift: Beibehaltung beider Datensätze.
NW	Unzustellbar	NW erhält Löschkennzeichen.
	Zustellbar, aber keine Antwort	Variante A: Bleibt NW, zugehörige HW wird gefunden. Variante B: HW wird nicht gefunden, Löschkennzeichen.
	Zustellbar, Antwort HW	NW wird zu HW.
	Zustellbar, Antwort NW	Variante A: Bleibt NW, zugehörige HW wird gefunden. Variante B: HW wird nicht gefunden, Löschkennzeichen.

HW = Hauptwohnung, NW = Nebenwohnung

1 Dubletten, bei denen im weiteren Verlauf der Klärung für einen Datensatz die Feststellung zu löschen erfolgte, werden im Referenzdatenbestand nicht gezählt. Anstatt diese Datensätze in der Datenbank physisch zu löschen, wurden sie mit einem Löschkennzeichen versehen, das es ermöglicht, sie bei Auswertungen auszusteuern.

Übersicht 2

Überblick über die bei der Mehrfachfallprüfung durchgeführten Bereinigungen

Zu prüfende Dublettenart	Bereinigungsart	Vorgehen	Konsequenz
Freischwebende Nebenwohnungen und Hauptwohnungsdubletten in Gemeinden mit weniger als 10 000 Einwohnern	Primärstatistische Klärung	siehe Übersicht 1	siehe Übersicht 1
Hauptwohnungsdubletten in Gemeinden mit mindestens 10 000 Einwohnern	Maschinell	Der aktuellere Datensatz wurde beibehalten.	Die neuere Hauptwohnung blieb bestehen, die ältere erhielt ein Löschkennzeichen.
Temporäre Dubletten	Maschinell	Der Datensatz mit dem älteren Einzugsdatum wurde auf zensusirrelevant gesetzt.	Die neuere Wohnung blieb bestehen, die ältere wurde gelöscht.
Veränderte personenbezogene Ordnungsnummer (PON) unter einer Anschrift	Primär maschinell, Qualitätssicherung manuell	Beibehaltung der alten PON, Übernahme der aktuellen Merkmale	Alte PON wurde für aktuellen Datensatz übernommen, maschinelle Zusammenführung zu einem aktuellen Datensatz

angesehen, die sich in der Datenlieferung MR.1 befanden und zu denen in der Datenlieferung MR.2 eine stichtagsrelevante neue Haupt- oder alleinige Wohnung bekannt wurde. Um diese Paare aufzulösen, wurde das Dublettenelement mit dem jüngeren Einzugsdatum aus der Datenlieferung MR.2 als aktuell gültige Wohnung erfasst und das Gegenstück als zensusirrelevant verworfen.

5.3 Korrekt im Referenzdatenbestand abgelegte Personendatensätze

Unter Personendatensätzen, die korrekt im Referenzdatenbestand abgelegt sind, sind Datensätze zu verstehen, die den melderechtlichen Vorschriften entsprechen: Jede meldepflichtige Person darf nur mit einer alleinigen oder Hauptwohnung in Deutschland am Zensusstichtag gemeldet sein und diese gemeldete Person darf eine oder mehrere Nebenwohnungen besitzen.

Prinzipiell war im gesamten Prozess der Konsolidierung darauf zu achten, dass die Bereinigungen der verschiedenen Dublettenarten und primärstatistischen Klärungen (Ausnahme: Ergebnisse der Haushaltebefragung) aufeinander abgestimmt wurden, sodass keine unzulässigen Dubletten mehr im Referenzdatenbestand vorhanden waren.

6 Fazit

Es ist anzunehmen, dass sich die Qualität und die Aktualität der Einträge in den Melderegistern im Vergleich zum Zeitpunkt des Zensustests 2001 verbessert haben. Dies könnte unter anderem an der Einführung der Steuer-Identifikationsnummer zum 1. Juli 2007 liegen. Im Zuge ihrer Einführung haben die zuständigen Finanzämter jedem Bürger und jeder Bürgerin an seiner/ihrer Hauptwohnung seine/ihre Steuernummer zugeteilt. Hierfür lieferten alle Meldeämter einen Datenauszug aus ihren Melderegistern an das Bundeszentralamt für Steuern. Durch die Verarbeitung der unzustellbaren Postrückläufer und die Rückmeldungen der Bürgerinnen und Bürger ist davon auszugehen, dass sich die Aktualität der Melderegister gesteigert hat. Eine Qualitätsverbesserung könnte auch durch eine besser abgestimmte elektronische Registerführung entstanden sein, deren Kern ein automatisiertes Rückmeldeverfahren zwischen den einzelnen Melderegister führenden Stellen auf der Basis eines einheitlichen Standards OSCI-XMeld[20] ist. Ob diese Annahme zutreffend ist, wird sich im Zuge der Auswertung der Ergebnisse der Mehrfachfallprüfung zeigen.

Wie schon beim Zensustest zeigte sich, dass bei den Prüfarbeiten im Zuge der Mehrfachfallprüfung flexibel und zeitnah auf unvorhersehbare Datenkonstellationen (wie beispielsweise mehrfach unter einer Anschrift gemeldete Personen) reagiert werden musste.

Zusammenfassend kann festgestellt werden, dass es notwendig ist, bei der Durchführung eines Zensus eine zentrale Prüfung der Melderegisterdaten auf Mehrfachanmeldungen durchzuführen, um einerseits Hauptwohnungsinhaber mit ihren Nebenwohnungen zu verbinden und andererseits temporär vorhandene Mehrfachanmeldungen mit Hauptwohnungen zu erkennen. Übersicht 2 gibt einen Überblick über die vorgenommenen Bereinigungen. Die hier dargestellten Verfahren haben sich für die Durchführung des Zensus 2011 als belastbar und zuverlässig herausgestellt und tragen somit zu einem validen Zensusergebnis bei. Zu den zahlenmäßigen Ergebnissen der Mehrfachfallprüfung im Zensus 2011 wird erneut in dieser Zeitschrift berichtet werden, wenn die Einwohnerzahlen veröffentlicht werden.

20 Für weitere Informationen siehe OSCI XMeld Expertengremium: Spezifikation OSCI-XMeld 1.8 (Fassung vom 31. Januar 2012), Koordinierungsstelle für IT-Standards (KoSIT), Bremen.