Dipl.-Mathematikerin Sarah Gießing (Statistisches Bundesamt), Dr. Felix Heinzl (Bayerisches Landesamt für Statistik und Datenverarbeitung), Dipl.-Soziologin Birgit Kleber (Statistisches Bundesamt), Dipl.-Wirtschaftsinformatiker (FH) Achim Wilke (Landesamt für Digitalisierung, Breitband und Vermessung, Bayern)

Geheimhaltung beim Zensus 2011

Dieser Artikel beschreibt die im Zensus 2011 angewandten statistischen Geheimhaltungsverfahren. Dabei liegt der Fokus auf der Geheimhaltung für Auszählungen aus dem Zensus-Einzeldatenbestand. Hierbei kam erstmals für eine solch umfangreiche Datenmenge das datenverändernde Geheimhaltungsverfahren SAFE zum Einsatz. Zunächst werden die Vorüberlegungen zur statistischen Geheimhaltung der Zensusdaten und die verschiedenen im Zensus 2011 eingesetzten Geheimhaltungsverfahren skizziert, bevor Kapitel 3 auf die Besonderheiten und Herausforderungen bei der Anwendung von SAFE im Zensus 2011 eingeht.

Dieser Aufsatz ist in Zusammenarbeit zwischen dem Statistischen Bundesamt, dem Bayerischen Landesamt für Statistik und Datenverarbeitung sowie dem Landesamt für Digitalisierung, Breitband und Vermessung (Bayern) entstanden und wird auch in der Ausgabe November 2014 der Zeitschrift „Bayern in Zahlen“ erscheinen.

1 Voraussetzungen und Überlegungen zur Geheimhaltung im Zensus 2011

In der Vorbereitungsphase des Zensus 2011 zeigte sich, dass traditionelle Zellsperrverfahren für den Großteil der Zensusdaten als Geheimhaltungsmethode aus methodischer Sicht nicht geeignet sind, da eine vollständige und konsistente Geheimhaltung aller geplanten Veröffentlichungen durch Sperrverfahren nicht realisierbar sein würde. Die Gründe hierfür sind vielfältig: Das Auswertungs- und Tabellenprogramm ist komplex, es besteht nicht nur aus einem statischen, sondern auch aus einem dynamischen Teil, es sind Datenproduktionen zu mehreren Zeitpunkten verteilt zu liefern, Datennutzer sind nicht nur die statistischen Ämter selbst, sondern auch Wissenschaftlerinnen und Wissenschaftler über die Forschungsdatenzentren oder auch Statistikerinnen und Statistiker in abgeschotteten Statistikstellen der Kommunen. Zudem gibt es verschiedene Erhebungsteile (zum Beispiel Registerauswertungen, Gebäude- und Wohnungszählung) mit unterschiedlichen Erhebungseinheiten (Personen, Haushalte, Familien, Gebäude, Wohnungen), über die hinweg die Geheimhaltung gewährleistet werden muss.

Eine Alternative zu Sperrverfahren stellen datenverändernde Verfahren dar. Datenverändernde Verfahren unterscheiden sich von informationsreduzierenden Verfahren dadurch, dass sie geheim zu haltende Informationen nicht unterdrücken, sondern durch veränderte Ergebnisse ersetzen. Den Nutzerinnen und Nutzern werden der Realität sehr ähnliche Ergebnisse präsentiert. Diese sollen die Eigenschaften der Gesamtheit sehr gut widerspiegeln, gleichzeitig aber auch verhindern, dass sicher Rückschlüsse über Einzelangaben gezogen werden können. Für den Zensus 2011 wurden zunächst drei datenverändernde Geheimhaltungsverfahren empirisch anhand eines ausgewählten exemplarischen Datenquaders der Daten der Volkszählung 1987 getestet:

- das Mikroaggregationsverfahren SAFE („Sichere Anonymisierung für Einzeldaten“),
- das Zufallsüberlagerungsverfahren des Australischen Statistischen Amts (ABS) und
- die invariante post-tabulare Methode (Shlomo, Young, 2008).

In den Entscheidungsprozess, welches Geheimhaltungsverfahren letztlich eingesetzt werden sollte, flossen unterschiedliche Kriterien ein: Aufdeckungsrisiko, Genauigkeit (Informationsverlust, verstanden als maximale Abweichung zwischen Original-Tabellenwert und dem durch das Geheimhaltungsverfahren veränderten Tabellenwert), Additivität[1], tabellenübergreifende Konsistenz, aber auch weitere Aspekte wie zum Beispiel der Aufwand, die Entwicklungszeit und die Entwicklungskosten für entsprechende Software.

Nach den Tests entschlossen sich die Statistischen Ämter des Bundes und der Länder aus einer Reihe von Gründen für den Einsatz des datenverändernden Geheimhaltungsverfahrens SAFE. Zum einen handelt es sich bei SAFE um ein in der amtlichen Statistik bereits eingesetztes und erprobtes Verfahren, welches mit relativ wenigen Anpassungen in die Zensusdatenbank integriert werden konnte. Zum anderen sind die Qualitätsanforderungen zur Additivität und Konsistenz[2] bei der Tabellierung mit SAFE automatisch gegeben und die Genauigkeit lässt sich zumindest für eine Auswahl an Merkmalen und Tabellen kontrollieren. Außerdem sind auch nachträglich gewünschte Ad-hoc-Auswertungen, die über den definierten (mit SAFE geheim gehaltenen) Merkmalskranz hinausgehen, möglich. Die Daten werden mit SAFE ausreichend geschützt, sodass die Aufdeckung von Einzelfällen gemäß der Forderung in § 16 Bundesstatistikgesetz verhindert wird.

2 Eingesetzte Geheimhaltungsverfahren

Die Erhebungsteile des Zensus 2011 lassen sich unter Geheimhaltungsaspekten in zwei wesentliche Gruppen einteilen: auf der einen Seite die Zensusmodule, die einen voll umfassenden Einzeldatenbestand abbilden (Registerauszählung, Erhebungen an Anschriften mit Sonderbereichen, Gebäude- und Wohnungszählung, Haushaltegenerierung), auf der anderen Seite die Haushaltebefragung auf Stichprobenbasis. Für die beiden Gruppen sind unterschiedliche Verfahren geeignet, um den Datenschutzansprüchen zu genügen. Da im Zensus 2011 für bestimmte Auswertungen auch noch eine Mischform zur Anwendung kommt, lassen sich die Geheimhaltungsverfahren in drei Kategorien einteilen: Geheimhaltung bei Auszählungen aus dem Zensus-Einzeldatenbestand (Abschnitt 2.1), Geheimhaltung bei Auswertungen aus der Haushaltsstichprobe (Abschnitt 2.2) und Geheimhaltung bei Auswertungen aus kombinierten Daten des Zensus-Einzeldatenbestands und der Haushaltsstichprobe (Abschnitt 2.3).

2.1 Geheimhaltung bei Auszählungen aus dem Zensus-Einzeldatenbestand

Tabellen, die durch reine Auszählung der Zensus-Einzeldaten erstellt werden, werden durch das Mikroaggregationsverfahren SAFE geheim gehalten. SAFE fand Anwendung auf:

- Daten zu Personen aus den korrigierten Registern (einschließlich der demografischen Daten zu Personen aus der Vollerhebung an Anschriften mit Sonderbereichen) sowie
- Daten zu Gebäuden, Wohnungen, Haushalten und Familien auf Basis der Gebäude- und Wohnungszählung und der Haushaltegenerierung.

Der Datenbestand wurde durch SAFE geringstmöglich verändert, sodass jede in den Originaldaten existierende Merkmalskombination im geheim gehaltenen Datenbestand (der sogenannten „anonymen Lösung“) mindestens dreimal oder gar nicht mehr auftrat. Es gilt: Ist eine Ausprägungskombination im Datenbestand ein Unikat, so ist sie mit einer Wahrscheinlichkeit von mindestens zwei Dritteln nach der Anonymisierung nicht mehr im Datenbestand vorhanden und wird somit in einer Tabelle zu einer 0. Das Kriterium für die Bestimmung der anonymen Lösung war, dass Abweichungen in zentralen Auswertungstabellen [dazu zählen Tabellen, die über die Zensusdatenbank[3] online verfügbar sind, sowie das Tabellenprogramm, das an das Statistische Amt der Europäischen Union (Eurostat) zu liefern war] minimiert und alle wichtigen statistischen Ergebnisse verlässlich abgebildet werden (Minimierung der Maximalabweichung).

Um dieses Ziel zu erreichen, wurden vorher definierte Auswertungstabellen, sogenannte Kontrolltabellen (siehe Abschnitt 3.1), zur Vorgabe des Optimierungsziels innerhalb des SAFE-Ablaufs in den Prozess eingebunden. Für diese Tabellen wurden während der Anonymisierung die Abweichungen zu den Originaltabellen kontrolliert, das heißt möglichst klein gehalten. Damit wurde erreicht, dass die Auswertungen der anonymen Daten die Struktur der Originaldaten gut widerspiegeln, auch wenn sie nicht völlig mit den Originalergebnissen identisch sind.

2.2 Geheimhaltung bei Auswertungen aus der Haushaltsstichprobe

Neben den Merkmalen der Erhebungsteile mit Vollerhebungscharakter wurde im Zensus 2011 eine Reihe von Merkmalen ausschließlich im Rahmen der Haushaltsstichprobe erhoben (zum Beispiel Angaben zur Bildung oder zum Beruf). Aus den Ergebnissen dieser Stichprobe werden Aussagen für die Gesamtheit der Bevölkerung über Hochrechnungen abgeleitet. Da in einer Stichprobe nur ein Teil der Bevölkerung befragt wird, ist ein Rückschluss auf eine einzelne Person kaum möglich, da es in der Grundgesamtheit noch weitere, nicht durch die Stichprobe erfasste Personen mit dieser Merkmalskombination geben kann. Entsprechende Rückschlüsse können also normalerweise nicht gezogen werden. Unter bestimmten Umständen sind jedoch auch aus Stichprobenergebnissen Aussagen über einzelne Personen ableitbar. Da bei den Befragungen im Zensus 2011 immer alle Personen an einer Anschrift befragt wurden, herrscht für bestimmte Personen Teilnahmekenntnis. Jede befragte Person kann zum Beispiel davon ausgehen, dass auch die Nachbarn an derselben Anschrift in der Stichprobe enthalten sind.

1 Additivität der Tabellen bedeutet, die Innenfelder jeder Tabelle addieren sich auf die Randfelder.

2 Die Konsistenz ist gegeben, sofern es sich bei den Merkmalen innerhalb einer SAFE-Anwendung um Merkmale der gleichen statistischen Einheit handelt. Ist dies nicht gegeben, kann es im geheim gehaltenen Datenbestand zu unplausiblen Datensätzen kommen.

3 *https://ergebnisse.zensus2011.de/*

Die Stichprobenergebnisse sind aufgrund der Zufallsauswahl der Stichprobenpersonen mit einem sogenannten Stichprobenzufallsfehler (Unschärfe im Ergebnis) behaftet; sie werden zudem durch Rundung der ermittelten Schätzwerte auf ein Vielfaches von zehn dargestellt.[4] Überdies werden einzelne Ergebnisse unterdrückt, wenn sie als statistisch unzuverlässig bewertet werden. Konkret wird ein hochgerechnetes Ergebnis für eine bestimmte Region dann gesperrt, wenn die zugrunde liegende Fallzahl folgende Schranke unterschreitet:

$$Fallzahl < \frac{1 - Auswahlsatz}{0{,}15^2}$$

Der Auswahlsatz hängt dabei von der jeweils betrachteten Region ab.

Insgesamt stellt bei hochgerechneten Ergebnissen diese Kombination aus Zufallsfehler, Rundung und der Kennzeichnung von statistisch nicht belastbaren Werten bereits eine hinreichende Geheimhaltung sicher.

2.3 Geheimhaltung bei Auswertungen aus kombinierten Daten des Zensus-Einzeldatenbestands und der Haushaltsstichprobe

Ein Teil der Personenergebnisse des Zensus 2011 wird durch die Kombination von Auszählung des Zensus-Einzeldatenbestands und hochgerechneten Stichprobendaten gebildet. So wurden die Merkmale Wirtschaftszweig, Erwerbsstatus und Stellung im Beruf zum Teil gleichzeitig in der Stichprobe und durch Registerangaben der Bundesagentur für Arbeit zu allen sozialversicherungspflichtig Beschäftigten (ausgenommen ausschließlich geringfügig Beschäftigte) sowie Registerangaben der öffentlichen Arbeitgeber zu Beamtinnen und Beamten, Richterinnen und Richtern, Soldatinnen und Soldaten sowie den Dienstordnungsangestellten[5] erfasst. Somit ist für einen bestimmten Personenkreis eine Registerauswertung möglich und nur für nicht im Register erfasste Personen muss die entsprechende Merkmalsausprägung aus der Stichprobe zugeschätzt werden.

Die aus dem Register ausgezählten Ergebnisse unterliegen der Geheimhaltung mit SAFE, wohingegen die Teile, die aus der Stichprobe zugeschätzt werden, durch die unter Abschnitt 2.2 beschriebenen Verfahren geschützt sind. Die Ergebnisse werden zudem auf ein Vielfaches von zehn gerundet dargestellt.

3 Herausforderungen beim Einsatz von SAFE im Zensus 2011

Um Daten mit SAFE innerhalb einer vertretbaren Laufzeit verarbeiten zu können, sind schnelle Zugriffe bei SAFE auf die Daten und alle Kontrolltabellen die grundlegende Voraussetzung. Aufgrund der Datenvolumina und des umfangreichen Kontrolltabellenprogramms stellte dies beim Zensus 2011 eine besondere Herausforderung dar. Das Ergebnis und die Laufzeit werden durch vier Faktoren bestimmt:

- Dateigröße respektive Anzahl der Datensätze,
- Anzahl der SAFE-relevanten (= geheim zu haltenden) Merkmale und deren Aggregationen,
- Anzahl der Kontrolltabellen,
- Parameter zur Begrenzung der Abweichungen.

Bei den Arbeiten für die konkrete Umsetzung und Implementierung in der Zensusdatenbank zeigte sich, dass aufgrund von Kapazitätsproblemen das Verfahren nicht auf den kompletten Datenbestand (als Ganzes)[6] angewandt werden konnte.

Im Zensus 2011 wurde diesem Umstand begegnet, indem sogenannte „Merkmalsscheiben“ (siehe hierzu Abschnitt 3.2) gebildet und separat mit SAFE geheim gehalten wurden.

3.1 Kontrolltabellen

Wie bereits in Abschnitt 2.1 erläutert, ermöglicht SAFE die kontrollierte Geheimhaltung vorab definierter Auswertungstabellen. In einer Auswertungstabelle werden die gemeinsamen Häufigkeiten von mehreren ausgewählten Merkmalen ausgewiesen. Ist eine Auswertungstabelle als Kontrolltabelle definiert, werden die Abweichungen, die durch die Veränderungen durch SAFE vorgenommen werden (also die Nach-SAFE-Häufigkeiten zu den Vor-SAFE-Häufigkeiten) „kontrolliert“, das heißt sie werden durch das SAFE-Verfahren so klein wie möglich gehalten. Es werden also sinnvollerweise solche Tabellen als Kontrolltabellen definiert, die man als „zentrale“ Ergebnistabellen ansieht, bei denen es besonders wichtig ist, dass die geheimhaltungsbedingte Datenveränderung möglichst klein bleibt. Fachliche und räumliche Aggregationen (zum Beispiel Gruppierungen des Alters oder der regionalen Ebene) wurden ebenfalls als eigene Kontrolltabellen abgebildet, damit entstehende Abweichungen sich in grob gegliederten Auswertungen nicht „aufschaukeln“.

Zu beachten ist, dass die Kontrolle der durch SAFE bewirkten Veränderungen nur innerhalb der jeweiligen Scheibe (zum Scheibenzuschnitt siehe Abschnitt 3.2) und nur für jene Merkmale, die in den Kontrolltabellen enthalten sind, erfolgen kann. Aufgrund der extrem hohen Anzahl an Kombinationsmöglichkeiten aus den Zensusmerkmalen war eine kontrollierte Geheimhaltung aller möglichen Merkmalskombinationen aus Performanzgründen nicht möglich. Trotzdem wurden für die Zensusdatenbank und für die Datenlieferung an Eurostat insgesamt 3 850 Auswertungstabellen kontrolliert.

Die räumliche Ebene, auf der die kontrollierte Geheimhaltung Anwendung fand, bildete die Gemeinde, beziehungs-

4 Diese Rundung wird nicht primär zu Geheimhaltungszwecken durchgeführt, sondern um dem Zufallsfehler Rechnung zu tragen und somit zu verdeutlichen, dass es sich bei den hochgerechneten Ergebnissen lediglich um Schätzwerte und nicht um exakt ermittelte Häufigkeiten handelt.

5 Zusammengefasst werden diese Registerdaten als „Erwerbsregister“ bezeichnet.

6 87 Millionen Personendatensätze, 43 Millionen Gebäude- und Wohnungsdatensätze, 40 Millionen Haushaltsdatensätze.

weise im Falle der Stadtstaaten Berlin und Hamburg die Ebene der Stadtbezirke. Untergemeindliche Zensusergebnisse genügen zwar auch den Anforderungen an die statistische Geheimhaltung; sie können jedoch – wie die scheibenübergreifenden Auswertungen – höhere (nicht-kontrollierte) Abweichungen zu Auswertungen aus den unveränderten Daten – sprich den Daten vor Geheimhaltung – haben.

Die Kontrolle über die regionale Zuordnung wurde dadurch erreicht, dass sämtliche Merkmalskombinationen nicht nur auf Gemeindeebene, sondern auch auf Verbandsgemeinde-, Kreis-, Regierungsbezirks-, Länder- und Bundesebene zusammen kontrolliert wurden. Dies setzt allerdings voraus, dass der für die Regionszugehörigkeit maßgebliche Amtliche Gemeindeschlüssel als Merkmal in SAFE berücksichtigt wird. Wie jedes andere Merkmal kann der Amtliche Gemeindeschlüssel damit auch Änderungen hinsichtlich seiner Häufigkeiten durch den Geheimhaltungslauf erfahren. Dies bewirkt, dass der Amtliche Gemeindeschlüssel bei manchen Datensätzen nach Geheimhaltung ein anderer sein kann als vorher. Eine Übersicht über die Personendatensätze, deren Amtlicher Gemeindeschlüssel durch SAFE verändert wurde, enthält Tabelle 1.

Tabelle 1 Übersicht über die durch SAFE gewechselten Personen nach regionaler Ebene

	Gewechselte Personen
Gemeinde	130 624
Verbandsgemeinde	76 286
Kreis	14 784
Regierungsbezirk	2 507
Land	1 104

3.2 Merkmalsscheiben

Die Zensusmerkmale und die damit verbundenen Kontrolltabellen wurden auf sogenannte Merkmalsscheiben für separate SAFE-Läufe verteilt, um akzeptable Rechenlaufzeiten und geringe scheibeninterne Vor-Nach-SAFE-Abweichungen[7] zu erreichen. Zunächst hatte man angenommen, die Zensusmerkmale nach rein fachlichen Gesichtspunkten auf die Merkmalsscheiben aufteilen zu können. Danach hätte man insgesamt drei bis vier getrennt geheim gehaltene Scheiben vorliegen gehabt: 1. Scheibe mit Personenmerkmalen, 2. Scheibe mit Gebäudemerkmalen, 3. Scheibe mit Haushalts- und Wohnungsmerkmalen sowie gegebenenfalls 4. Scheibe mit Familienmerkmalen[8]. Es zeigte sich jedoch in den Testläufen, dass sogenannte scheibenübergreifende Auswertungen häufig mit großen Ungenauigkeiten (hohe Abweichung des veränderten Zahlenwerts vom Original-Zahlenwert) behaftet waren. Es galt also, einen Kompromiss zu finden zwischen Performanz (möglichst viele kleine Merkmalsscheiben) auf der einen Seite und Genauigkeit (möglichst nur eine einzige Merkmalsscheibe) auf der anderen Seite. Die Aufteilung der nationalen Auswertungsmerkmale des Zensus 2011 in zwei unterschiedliche Merkmalsscheiben hat sich als guter Kompromiss erwiesen: 1. alle Auswertungsmerkmale zu Personen aus den Registern und 2. Auswertungsmerkmale zu Haushalten, Familien, Wohnungen und Gebäuden. Diese beiden Merkmalsscheiben wurden getrennt mit SAFE geheim gehalten. Die vorgenommenen Änderungen sowie deren Kontrolle fanden dementsprechend jeweils innerhalb der jeweiligen Merkmalsscheibe statt.

Da auch die an Eurostat zu liefernden Daten geheim zu halten waren, wurden diese wie folgt in das Scheibenkonzept integriert: Die EU-Personenmerkmale wurden auf der nationalen Personenscheibe mit geheim gehalten. Für die EU-Merkmale zu Wohnungen, Haushalten, Familien sowie für ein Merkmal zur Art der Unterkunft wurden jeweils eigene Scheiben gebildet, da deren Definitionen sich von den deutschen Auswertungsmerkmalen unterscheiden.

3.3 Matching-Verfahren

Grundsätzlich enthält der anonyme, das heißt mit SAFE geheim gehaltene, Datenbestand keine Identifikatoren aus Registern oder Erhebungen wie zum Beispiel den Haushaltsidentifikator, keine Zuordnungen zu kleinräumigen (innergemeindlichen) Gliederungssystemen und keine nicht für Standardauswertungen vorgesehenen Originalmerkmale aus den Erhebungsteilen. Neben einer systeminternen Datensatzkennung enthält der geheim gehaltene Datenbestand die für die Kontrolltabellen definierten Merkmale[9], die in den SAFE-Lauf eingegangen sind, sowie ein Zählmerkmal. Dieses Zählmerkmal gibt an, mit welcher Häufigkeit (im Folgenden mit n bezeichnet) eine identische Merkmalskombination im Datenbestand nach Geheimhaltung vorkommt. Für scheibeninterne Auswertungen sind nur diese Häufigkeiten relevant.

Für scheibenübergreifende Auswertungen benötigt man dagegen die Kenntnis der Identifikatoren (zum Beispiel den Haushaltsidentifikator), über die die Merkmale unterschiedlicher Scheiben verknüpft werden können. In einem ersten Schritt müssen deshalb den geheim gehaltenen Merkmalskombinationen – entsprechend ihrer jeweils von SAFE ermittelten Häufigkeiten (n) – wieder Originaldatenzeilen zugeordnet werden.

Dies erfolgt durch eine Ähnlichkeitssuche auf Basis eines Matching-Algorithmus. Mithilfe einer Priorisierungsliste von Merkmalen wird dabei gesteuert, welche Merkmale möglichst identisch zum Originaldatenbestand sein sollen und bei welchen Merkmalen Abweichungen bevorzugt in Kauf genommen werden. Typischerweise werden die regionalen Merkmale wie der Amtliche Gemeindeschlüssel sehr hoch priorisiert, um geheimhaltungsbedingte Regionswechsel beispielsweise von Personen und Wohnungen klein zu halten. Damit soll gewährleistet werden, dass auch auf tiefer regionaler Ebene – im Falle des Zensus 2011 auf Gemeindeebene beziehungsweise Stadtbezirksebene für Hamburg und Berlin – Ergebnisse weitestgehend unverändert blei-

7 Dies sind Abweichungen innerhalb eines Tabellenfeldes, wenn dieselbe Auswertung einmal mit Originaldaten und einmal mit geheim gehaltenen Daten vorgenommen wird.

8 Im Falle der Familienmerkmale musste zunächst untersucht werden, ob sie als eigene Scheibe geheim gehalten werden sollten oder zu welcher anderen Scheibe die entsprechenden Merkmale hinzugefügt werden könnten, ohne dadurch zu hohe Unplausibilitäten hervorzurufen.

9 Diese sind im Zensus 2011 aus den originalen Merkmalen abgeleitete Merkmale, sogenannte Auswertungsmerkmale.

ben. Gleichzeitig kann eine sinnvoll gewählte Priorisierungsliste die Zahl der SAFE-bedingten Änderungen insgesamt minimieren.

Das Matching stellt also eine Verknüpfung zwischen einem SAFE-Datensatz der Häufigkeit (*n*) und *n* Originaldatensätzen (vor Anwendung von SAFE) her. In den geheim gehaltenen Datenbestand können über diese Verknüpfung im zweiten Schritt sowohl die Identifikatoren der *n* gematchten Originaldatensätze übernommen werden als auch deren Zugehörigkeit zu einem räumlichen Gliederungssystem (zum Beispiel Blockseite). Auch nicht in SAFE einbezogene Originalmerkmale können in den geheim gehaltenen Datenbestand integriert werden. Um bei deren Auswertung ebenfalls den Anforderungen an die statistische Geheimhaltung zu genügen, werden jedoch nicht die Werte der *n* Originaldatensätze angespielt, sondern für jeden geheim gehaltenen Datensatz der Häufigkeit (*n*) die Ausprägung jeweils nur eines der *n* Originaldatensätze. Damit haben auch die herangespielten Originalmerkmale mindestens die Häufigkeit drei und sind damit implizit ebenfalls geheim gehalten.

3.4 Scheibenübergreifende Auswertungen

Wenn Merkmale ausgewertet werden sollen, die auf unterschiedlichen Scheiben geheim gehalten worden sind, dann spricht man von einer scheibenübergreifenden Auswertung. Wie bereits in Abschnitt 3.2 beschrieben, teilte man aus Gründen der Performanz die Datenmenge beim Zensus 2011 für die nationalen Merkmale in eine Scheibe mit Personenmerkmalen und in eine Scheibe mit Familien-, Haushalts-, Wohnungs- und Gebäudemerkmalen auf. Somit liegt beim Zensus 2011 eine scheibenübergreifende Auswertung zum Beispiel dann vor, wenn ein Personenmerkmal wie „Geschlecht“ mit einem Haushaltsmerkmal wie „Typ des privaten Haushalts (nach Familien)“ gemeinsam ausgewertet wird.

Nach der in Abschnitt 3.3 beschriebenen Matching-Prozedur können über die zugeordneten Identifikatoren die Merkmalsscheiben miteinander verknüpft werden, allerdings mit dem Nachteil, dass es bei den scheibenübergreifenden und damit nicht kontrollierten Auswertungen[10] zu höheren Abweichungen als bei „scheibeninternen“ (= kontrollierten) Auswertungen kommen kann (zur erreichten Genauigkeit siehe Abschnitt 3.5).

Bei der kombinierten Auswertung von Merkmalen aus unterschiedlichen Scheiben wird so vorgegangen, dass stets die „detaillierte“ statistische Einheit betrachtet wird. In obigem Beispiel werden also die weiblichen und männlichen Personen ausgewiesen, die in bestimmten Haushaltstypen leben. Diesem Grundsatz folgend werden im genannten Beispiel Personen statt Haushalte ausgezählt. Daher kann bei den Haushalten auch der nicht geheim gehaltene Datenbestand verwendet werden. Dies hat den Vorteil, dass dadurch die Vor-Nach-SAFE-Abweichungen reduziert werden.

10 Dazu zählen alle scheibenübergreifenden Auswertungen, Auswertungen von nicht für die Standardauswertung vorgesehenen Original-Erhebungsmerkmalen oder Auswertungen unterhalb der Gemeindeebene; siehe hierzu die Ausführungen in Abschnitt 3.1.

Die Aufteilung eines Datenbestandes auf mehrere Merkmalsscheiben ist im Allgemeinen dann problematisch, wenn inhaltlich-logische Verknüpfungen zwischen den statistischen Einheiten unterschiedlicher Merkmalsscheiben bestehen. Daher kann es, obwohl scheibenintern durch SAFE keine neuen Merkmalskombinationen entstehen können, bei scheibenübergreifenden Auswertungen sehr wohl zu solchen kommen. Daraus können Unplausibilitäten wie zum Beispiel die folgende resultieren: Der Originaldatenbestand enthält die Information über eine allein in einem Haushalt lebende 70 Jahre alte Person, die folglich in einem Seniorenhaushalt lebt. Während es bei dem Haushaltsmerkmal „Seniorenstatus des privaten Haushalts“ bei dieser Person auf der Scheibe mit den Familien-, Haushalts-, Wohnungs- und Gebäudemerkmalen keine Änderung gab, wurde auf der Scheibe mit den Personenmerkmalen das Merkmal Alter durch SAFE von 70 Jahre auf 15 Jahre geändert. Dies führt bei der scheibenübergreifenden Auswertung zu einer 15 Jahre alten Person in einem Seniorenhaushalt. Um solche Unplausibilitäten, die erst durch die Geheimhaltung entstanden sind, nicht auszuweisen, werden durch SAFE entstandene, im Originaldatenbestand nicht existierende Merkmalskombinationen in den Ergebnistabellen gesperrt dargestellt.

Des Weiteren können bei scheibenübergreifenden Auswertungen neue Geheimhaltungsfälle entstehen: Angenommen, es gibt in einer Gemeinde vier Männer im Alter zwischen 25 und 29 Jahren und 15 Reihenhäuser. Dann stellen diese Häufigkeiten in den beiden Merkmalsscheiben keine Geheimhaltungsfälle dar, da nur 1er- und 2er-Häufigkeiten durch SAFE beseitigt werden sollen. Kombiniert man nun die Datenbestände, kann es sein, dass nur einer dieser Männer in einem Reihenhaus lebt. Solche bei scheibenübergreifenden Auswertungen entstehende geheimhaltungsrelevante Häufigkeiten wurden durch die sogenannte 3er-Rundung entfernt, bei der 1er-Häufigkeiten zu 0 und 2er-Häufigkeiten zu 3 verändert wurden. Diese einfache Rundung ist deswegen gerechtfertigt, da die Häufigkeit von vier Männern im Alter zwischen 25 und 29 Jahren selbst und damit auch die Teilmenge derer in Reihenhäusern möglicherweise durch Geheimhaltung verändert worden sind.

3.5 Erreichte Genauigkeit

a) Vor-Nach-SAFE-Abweichungen bei kontrollierten Tabellen

Die hier zusammengestellten Kennzahlen beziehen sich auf alle statistischen Ergebnisse in den Tabellenfeldern der Auswertungstabellen innerhalb eines Datenbestandes, bei denen eine Geheimhaltung mit SAFE durchgeführt wurde: Dies umfasst etwa 214 Millionen Tabellenfelder, die Ergebnisse zur Bevölkerung ausweisen, beziehungsweise rund 164 Millionen Tabellenfelder mit Ergebnissen zu Gebäude-, Wohnungs-, Haushalts- und Familiendaten. Insgesamt wirkt sich die Geheimhaltung durch SAFE auf die Ergebnisse zu Gebäuden, Wohnungen, Haushalten und Familien etwas stärker aus – die mittlere absolute, durch SAFE bewirkte Veränderung der Originalhäufigkeiten beträgt hier 3,8. Bei den Bevölkerungsdaten hingegen liegen die absoluten Abweichungen im Mittel bei 2,5.

Tabelle 2 Vor-Nach-SAFE-Abweichungen bei der Gebäude-, Wohnungs-, Haushalts- und Familienscheibe und bei der Personenscheibe (kontrollierte Tabellen)

Gebäude, Wohnungen, Haushalte und Familien			Bevölkerung		
(absolute) Abweichung	Anzahl Tabellenfelder	Anteil Tabellenfelder kumuliert (mit Abweichung bis . . .) (in %)	(absolute) Abweichung	Anzahl Tabellenfelder	Anteil Tabellenfelder kumuliert (mit Abweichung bis . . .) (in %)
0	11 861 696	*7,2*	0	18 142 797	*8,4*
1	42 653 388	*33,2*	1	72 099 817	*42,0*
2	30 890 658	*52,0*	2	49 224 791	*65,0*
3	17 995 609	*63,0*	3	24 755 130	*76,5*
4	13 156 391	*71,0*	4	16 777 948	*84,3*
5	9 722 960	*76,9*	5	11 164 944	*89,5*
6	7 641 196	*81,6*	6	7 848 158	*93,1*
7	6 061 088	*85,3*	7	5 403 469	*95,7*
8	4 871 650	*88,3*	8	3 690 397	*97,4*
9	3 953 993	*90,7*	9	2 499 400	*98,5*
10	3 242 188	*92,6*	10	1 625 161	*99,3*
11	2 639 460	*94,2*	11	1 031 806	*99,8*
12	2 143 338	*95,5*	12 oder mehr	460 664	*100*
13	1 754 884	*96,6*			
14	1 438 963	*97,5*			
15	1 173 370	*98,2*			
16	955 428	*98,8*			
17	812 872	*99,3*			
18	706 897	*99,7*			
19	337 575	*99,9*			
20 oder mehr	126 044	*100*			
Zusammen	164 139 648		Zusammen	214 724 482	

In Tabelle 2 erkennt man, dass die durch SAFE bewirkte Veränderung der Originalhäufigkeiten bei der Mehrheit der Tabellenfelder bei bis zu +/– 2 liegt, nämlich bei 52 % der in den Tabellen ausgewiesenen Ergebnisse zu Gebäuden, Wohnungen, Haushalten und Familien und bei 65 % der ausgewiesenen Bevölkerungsergebnisse. Abweichungen von mindestens 8 sind schon deutlich seltener. Sie finden sich bei nur noch 4,3 % der Tabellenfelder mit Bevölkerungsergebnissen beziehungsweise bei weniger als 15 % der Tabellenfelder mit Ergebnissen zu Gebäuden, Wohnungen, Haushalten und Familien. Bei den Bevölkerungsdaten treten Abweichungen in den ausgewiesenen Ergebnissen von mehr als +/–11 sehr selten auf. Abweichungen in dieser Größenordnung kommen nur bei etwa 2 von 1 000 Ergebnissen vor. Bei den ausgewiesenen Ergebnissen zu Gebäuden, Wohnungen, Haushalten und Familien kommen – äußerst selten – auch Abweichungen von 20 oder mehr vor. Weniger als 8 von 10 000 (exakt: 126 044 von 164 Millionen) Ergebnissen weisen eine Abweichung dieser Größenordnung auf.

b) Vor-Nach-SAFE-Abweichungen bei nicht kontrollierten Tabellen

Werden Tabellen nicht kontrolliert, können die Vor-Nach-SAFE-Abweichungen durchaus hoch werden. Abschätzen lassen sich diese von vornherein generell nicht. Besonders bei scheibenübergreifenden Auswertungen können die Abweichungen hoch ausfallen, da sich dort die Abweichungen zweier Scheiben multiplizieren können. Unkontrollierte Auswertungen liegen auch immer dann vor, wenn eine Teilmenge des geheim gehaltenen Datenbestandes betrachtet wird. Selbst wenn eine personenbezogene Auswertungstabelle für den Gesamtbestand an Personen kontrolliert wurde und geringe Vor-Nach-SAFE-Abweichungen aufwies, gab es in der Regel bei vielen Personendatensätzen Änderungen der Merkmalskombinationen. Diese können sich im Ergebnis bezogen auf die Kontrolltabellen ausgeglichen haben, für eine (nicht durch Kontrolltabellen abgedeckte) Teilmenge muss dies aber nicht der Fall sein. Dieses Problem trat beim Zensus 2011 bei den scheibenübergreifenden Auswertungen zusätzlich auf: Während zwar auf der Personenscheibe Auswertungstabellen für den bundesweiten Personenbestand von 80 219 695 Bürgerinnen und Bürgern kontrolliert waren, wurden bei Kombinationen von Personen- mit Haushaltsmerkmalen nur die Personen ausgewiesen, denen ein mit einer Wohnung verknüpfter Haushalt zugewiesen werden konnte, nämlich 78 672 982 Personen. Für diese Teilmenge an Personen waren die Auswertungstabellen jedoch nicht kontrolliert.

Wie bereits beschrieben wurden die Kontrolltabellen auf räumlicher Ebene der Gemeinde beziehungsweise für Hamburg und Berlin auf Ebene der Stadtbezirke definiert. Dies bedeutet auch, dass Auswertungen unterhalb dieser räumlichen Einheit ebenfalls nicht kontrolliert wurden und somit hohe Abweichungen aufweisen können.

3.6 Geheimhaltung von Verhältniszahlen

Zur Berechnung von als Quotienten aus Zähler und Nenner gebildeten Verhältniszahlen, zum Beispiel der durchschnittlichen Wohnungsgröße, werden in der Zensusdatenbank die Originaldaten benutzt[11], weil ein Quotient von durch SAFE veränderten Zahlen in bestimmten Konstellationen

11 Allerdings werden im Zuge der Rundung auf das vorgesehene Darstellungsformat (zum Beispiel als Prozentwert mit einer Nachkommastelle) Ergebnisse gelegentlich aufgerundet, obwohl nach kaufmännischer Rundungsvorschrift abzurunden wäre, und umgekehrt. Dies geschieht zur Vermeidung von Konsistenzproblemen zu den durch SAFE geänderten Zählern oder Nennern.

erheblich vom Originalverhältniswert abweichen kann. Es kann sein, dass die Breite dieses Wertespektrums in keinem sinnvollen Verhältnis zur Darstellungsgenauigkeit (zum Beispiel als Prozentzahl mit einer Nachkommastelle bei Anteilswerten) steht.

Beispiel: In einer kleineren Gemeinde gibt es 20 Gebäude eines bestimmten Typs, von denen genau 50 % (also 10 Gebäude) in eine bestimmte Altersklasse des Baujahrs fallen. Wird diese Verhältniszahl aus den SAFE-Werten berechnet, können sich – je nach Ausprägung der SAFE-Veränderung – große Unterschiede ergeben. So könnte der Nenner durch SAFE beispielsweise um 4 auf 24 vergrößert und der Zähler um 5 auf 5 verkleinert worden sein – ausgewiesen würde ein Prozentsatz von 20,8 %. Umgekehrt, bei Verkleinerung des Nenners um 4 Fälle und Vergrößerung des Zählers um 5 Fälle durch SAFE, würde das Ergebnis als 93,8 % ausgewiesen.

Werden die Originaldaten benutzt, muss jedoch verhindert werden, dass aus den Verhältniswerten auf Originalwerte von Zählern oder Nennern zurückgeschlossen werden kann, da – selbst wenn dieser Originalwert selbst keinen Geheimhaltungsfall darstellt – möglicherweise durch Vergleich von Originalwerten mit durch SAFE geänderten Werten andere Geheimhaltungsfälle aufgedeckt werden könnten. Verhältniswerte werden deshalb nur dann ausgewiesen, wenn sie für ausreichend große Gruppen statistischer Einheiten gebildet werden. Bei der Bewertung der Gruppengröße spielt die Darstellungsgenauigkeit eine Rolle: Wenn beispielsweise in einer Gemeinde mit rund 1 000 Einwohnern nur eine Person mit einer bestimmten Staatsangehörigkeit lebt, kann der Anteil dieser Staatsangehörigengruppe für diese Gemeinde nicht als Prozentzahl mit einer Nachkommastelle als 0,1 % ausgewiesen werden. Denn multipliziert man diesen Anteil mit der Einwohnerzahl wird offensichtlich, dass es sich um genau eine Person handelt ($0{,}001 \cdot 1\,000 = 1$). In diesem Falle würde keine Verhältniszahl ausgewiesen werden. Dagegen wäre bei einem Ergebnisnachweis ohne Nachkommastellen kein exakter Rückschluss möglich: Das dann dargestellte Ergebnis 0 % kommt sowohl bei einer, als auch bei zwei, drei oder vier Personen mit der betreffenden Staatsangehörigkeit zustande. Bei einer größeren Gruppe statistischer Einheiten kann meist auch ein Wert mit Nachkommastelle ausgewiesen werden. In einer Gemeinde mit 10 000 Einwohnern beispielsweise kommt das Ergebnis 0,1 % bei jeder Staatsangehörigenanzahl zwischen 5 und 14 zustande. Ein exakter Rückschluss auf die Originalzahl ist nicht möglich.

3.7 Kennzeichnung von geheimhaltungsbedingten Ergebnisabweichungen

Um Fehlinterpretationen der Zensusergebnisse vorzubeugen, werden in den Ergebnistabellen Zahlen, bei denen sowohl die absolute als auch die relative Abweichung des veränderten Zahlenwerts vom Original-Zahlenwert deutlich erhöht sind, in Klammern ausgewiesen. Werte mit ungewöhnlich großen Abweichungen werden gesperrt und in den Ergebnistabellen mit einem Punkt dargestellt. Derart große Abweichungen kommen vor allem in kombinierten Auswertungen aus den beiden Merkmalsscheiben, bei Darstellungen nach nicht im SAFE-Auswertungsdatenbestand enthaltenen regionalen Gliederungen oder bei Auswertungen aus nicht in den Kontrolltabellen enthaltenen Merkmalen vor, denn hier gelten die Qualitätsaussagen aus Abschnitt 3.5 Teil a) zur Häufigkeitsverteilung der durch SAFE erzeugten Abweichungen der Tabellenfelder von ihren Originalwerten naturgemäß nicht. Das Sperrsymbol Punkt wird zudem bei Tabellenfeldern eingesetzt, bei denen in den geheim gehaltenen Tabellen eine Merkmalskombination ausgewiesen wird, die bei Berechnung auf den Originalwerten nicht vorkommt (zum Beispiel 15-jährige Personen im Haushaltstyp „Seniorenhaushalt“).

4 Fazit

Die Daten des Zensus 2011 sind durch das angewandte Geheimhaltungsverfahren SAFE sehr flexibel für die Nutzer über die Zensusdatenbank auswertbar, ohne dass dabei ein unverhältnismäßig hoher Informationsverlust in Kauf genommen werden müsste (wie es beim Einsatz von Zellsperrungsverfahren zu erwarten gewesen wäre) und ohne dass ein Aufdeckungsrisiko zu befürchten wäre. Trotz dieses Erfolges zeigten sich in der Testphase und auch in der Anwendung immer wieder Aspekte, die durch Anpassung der Datenzuschnitte (zum Beispiel Zuschnitt der Merkmalsscheiben) oder durch Erweiterung der zur Anwendung kommenden Methode (zum Beispiel Geheimhaltung der Verhältniszahlen) nötig waren. Die bei scheibenübergreifenden Auswertungen nicht von vornherein bestimmbare Genauigkeit erfordert bei der Interpretation der Daten erhöhte Aufmerksamkeit.

Für den kommenden Zensus 2021 ist geplant, weitere Qualitätsuntersuchungen durchzuführen, beispielsweise einen Vergleich mit anderen Geheimhaltungsverfahren, oder auch Kosten und Aufwand bei der Implementierung neuer Verfahren abzuschätzen. Ferner erfordern sowohl die Geheimhaltung georeferenzierter Zensusergebnisse als auch die Geheimhaltung von Zensusdatenangeboten der Forschungsdatenzentren noch weitere methodische Überlegungen.

Auszug aus Wirtschaft und Statistik

Herausgeber
Statistisches Bundesamt, Wiesbaden

www.destatis.de

Schriftleitung
Dieter Sarreither,
Vizepräsident des Statistischen Bundesamtes

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

Ihr Kontakt zu uns
www.destatis.de/kontakt

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

Abkürzungen

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

Zeichenerklärung

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

Abweichungen in den Summen ergeben sich durch Runden der Zahlen.