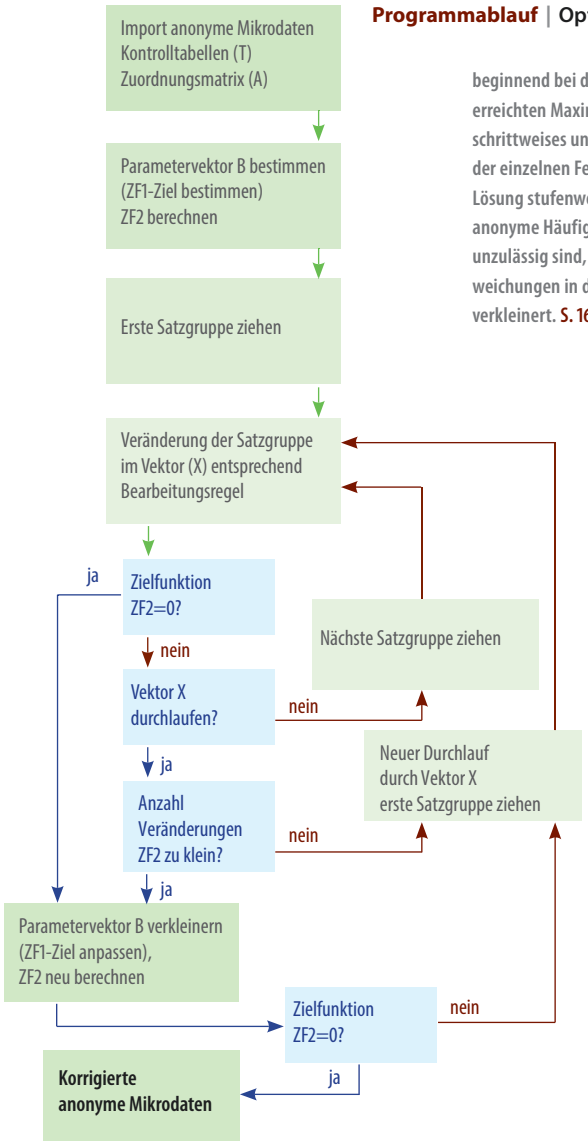
Die Qualität der ersten Lösung des Geheimhaltungsverfahrens SAFE wird durch eine Korrektur „Optimierung der Lösung“ noch verbessert. Das Programm versucht,

Programmablauf | **Optimierung der Lösung**



beginnend bei den im ersten Schritt erreichten Maximalabweichungen, durch schrittweises unabhängiges Verkleinern der einzelnen Fehlerschranken (bounds) die Lösung stufenweise zu korrigieren. Da nicht anonyme Häufigkeiten im Korrekturlauf unzulässig sind, werden nur noch die Abweichungen in den Auswertungstabellen verkleinert. S. 16

Impressum

Amt für Statistik Berlin-Brandenburg

Zeitschrift für amtliche Statistik Berlin Brandenburg
9. Jahrgang

Herausgeber
Amt für Statistik Berlin-Brandenburg
Behlertstraße 3a
14467 Potsdam
Tel.: 0331 8173-1777

Verantwortlicher Redakteur i. S. d. BbgPG
Hartmut Bömermann
Redaktion
Nicole Dombrowski,
Dr. Holger Leerhoff,
Anja Malchin, Dr. Thomas Troegel,
Ramona Voshage (Leitung)

Preis
Einzelheft EUR 6,00
ISSN 1864-5356

Satz und Gestaltung
Amt für Statistik Berlin-Brandenburg

Druck
TASTOMAT GmbH,
15345 Petershagen/Eggersdorf

Das Amt für Statistik Berlin-Brandenburg hat seinen Sitz in Potsdam und weitere Standorte in Berlin und Cottbus.

Auskunft und Beratung

Behlertstraße 3a
14467 Potsdam

Telefon: 0331 8173-1777
Fax: 030 9028-4091
info@statistik-bbb.de

Zeichenerklärung

0	weniger als die Hälfte von 1 in der letzten besetzten Stelle, jedoch mehr als nichts
–	nichts vorhanden
…	Angabe fällt später an
( )	Aussagewert ist eingeschränkt
/	Zahlenwert nicht sicher genug
•	Zahlenwert unbekannt oder geheim zu halten
x	Tabellenfach gesperrt, weil Aussage nicht sinnvoll
p	vorläufige Zahl
r	berichtigte Zahl
s	geschätzte Zahl

Abweichungen in der Summe können sich durch Schätzungen ergeben

Editorial

Liebe Leserinnen und Leser,

von Aktivisten, die die digitale Existenz radikal leben wollen, wird das Ende der Privatheit ausgerufen und alles Persönliche auf Netzplattformen gestellt, die großen Internetunternehmen bieten einen bis dahin kaum gekannten Komfort gegen die stillschweigende Überlassung der Daten ihrer Nutzer und die drei Buchstaben eines staatlichen Nachrichtendienstes sind zum Menetekel einer forcierten Überwachung geworden. Die Herausforderungen an den Datenschutz waren wohl noch nie so groß wie in der heutigen vernetzten und sich schnell weiter digitalisierenden Welt. Wie können Daten geschützt werden?

In der amtlichen Statistik hat man sich mit dieser Frage bereits frühzeitig beschäftigt und Regelungen und Verfahren entwickelt, die die Angaben der zu einer Statistik Berichtenden sicher schützen. Die strikte Trennung zwischen Verwaltung und Statistik mit dem Verbot der Rückübermittlung von Daten, die Anordnung einer Erhebung durch die Legislative, die Löschung aller Hilfsinformationen, die für den Erhebungsprozess erforderlich sind, und die statistische Geheimhaltung gehören zum Fundament der amtlichen Statistik.

Der statistischen Geheimhaltung widmen sich in dieser Ausgabe der *Zeitschrift für amtliche Statistik Berlin Brandenburg* drei Fachbeiträge und das Fachgespräch.

↳

Inhaltsübersicht

Im ersten Fachbeitrag stellt Dr. Jörg Höhne das von ihm entwickelte Verfahren **SAFE** vor. SAFE ist ein datenveränderndes Anonymisierungsverfahren. Datenveränderung klingt provozierend, da die Statistik doch genaue und verlässliche Ergebnisse liefern soll. Wie geht das zusammen? Und warum wird nicht auf Bewährtes vertraut? Der Autor führt in den Algorithmus und die strukturerhaltende Optimierungsstrategie ein, deren Ziel die Lösung des Konfliktes zwischen Genauigkeit und Schutzwirkung ist. SAFE bietet die Möglichkeit, einen anonymisierten Datenkubus zu erzeugen, der beliebig auswertbar ist und jeder denkbaren Auswertung anonymisierte Tabellen liefert – eine entscheidende Voraussetzung für flexible Auswertungsdatenbanken im Internet.

Julia Höninger behandelt die Enthüllungsrisiken in ihrem Beitrag **Mindestfallzahlregel versus Randwertregel**. Die 3er-Mindestfallzahlregel ist die bekannteste Anwendungsregel für die Tabellengeheimhaltung. Gefordert wird bei der Erörterung der Randwertproblematik ein Umdenken, das in der Konsequenz sogar die Sperrung kleiner Häufigkeiten verzichtbar machen könnte.

Für die wissenschaftliche Forschung sieht das Bundesstatistikgesetz einen privilegierten Zugang zu den Mikrodaten vor, den die Statistischen Ämter des Bundes und der Länder in ihren Forschungsdatenzentren (FDZ) ermöglichen. Bevor Ergebnisse den geschützten Bereich verlassen, ist eine Geheimhaltungsprüfung vorgeschaltet. Diese Prüfung ist aufwändig und verzögert die Übergabe an die Wissenschaftlerinnen und Wissenschaftler. Jakob Pohlisch, Julia Höninger und Ramona Voshage stellen in ihrem Beitrag **FiRe** einen Ansatz zur Teilautomatisierung vor. Sowohl die FDZ als auch die Wissenschaft profitieren von einer stärkeren Verlagerung von (Teil-) Prozessen auf technische Systeme.

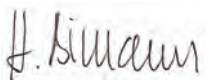
Das Fachgespräch mit Sarah Giessing (Statistisches Bundesamt) ergänzt den Themenblock als Orientierungshilfe zu den Entwicklungssträngen in der statistischen Geheimhaltung.

Ebenfalls methodische Fragen behandelt Dr. Thomas Heymann in seinem Beitrag **Unternehmen und Betriebe**. Der Schwerpunkt liegt auf der Entwicklung in den Ländern Berlin und Brandenburg seit Januar 2004, wie sie aus der Gewerbeanzeigen- und Insolvenzstatistik nachgezeichnet werden kann.

Eine neue Analyse von Zensusergebnissen stellen Verena Kutzki und Marco Schwarz vor. Sie untersuchen **Erstbezugseigentümer in Berlin und Brandenburg** in einem generationenbezogenen Ansatz, der Personen-, Haushalts-, Gebäude und Wohnungsmerkmale einbezieht. In ihrem Resümee entwickeln sie Forderungen zum Merkmalsumfang und der Auswertbarkeit für den künftigen Zensus 2021.

Ich hoffe, dass der eine oder andere Beitrag Ihr Interesse findet.

Eine anregende Lektüre wünscht Ihnen



Hartmut Bömermann
verantwortlicher Redakteur

Kurzbericht

Bevölkerungsvorausberechnung im Auftrag der Gemeinde Petershagen/Eggersdorf

von **Jürgen Paffhausen**

Zu den Aufgaben des Amtes für Statistik Berlin-Brandenburg (AfS) gehört es, maßgeblich an Bevölkerungsvorausberechnungen für die Bundesländer Berlin und Brandenburg mitzuwirken. Für Berlin werden Bevölkerungsprognosen unter der Federführung der Senatsverwaltung für Stadtentwicklung und Umwelt erstellt. Das AfS liefert dafür die nötige Datengrundlage und steht der Senatsverwaltung beratend zur Seite. Für das Land Brandenburg werden Bevölkerungsprognosen gemeinsam mit dem Landesamt für Bauen und Verkehr erarbeitet. Darüber hinaus fertigt das Amt auch Vorausberechnungen im Auftrag einzelner Städte und Gemeinden an.

Durch den Zensus 2011 sind die Ergebnisse der zuletzt durchgeführten Bevölkerungsvorausberechnung für das Land Brandenburg und seine Verwaltungsbezirke nur noch eingeschränkt verwendbar. Die Berechnung der Prognose erfolgte durch das AfS auf den Ausgangsdaten des Jahres 2010. Eine Aktualisierung des Rechenwerks war bislang noch nicht möglich, da die fortlaufende Ermittlung der Bevölkerungszahl (amtliche Bevölkerungsfortschreibung) noch auf vorläufigen Ausgangsdaten beruhte. Jetzt liegen die endgültigen Ergebnisse der Bevölkerungsfortschreibung bis zum Jahr 2013 vor, sodass mit der Planung einer neuen Bevölkerungsprognose begonnen werden kann.

In der Gemeinde Petershagen/Eggersdorf, die etwa 30 km östlich vom Berliner Stadtzentrum entfernt liegt und dem Berliner Umland (dem sogenannten Speckgürtel) zuzurechnen ist, gab es wegen anstehender Planungsvorhaben bereits vor der Fertigstellung einer neuen landesweiten Prognose Bedarf an Zahlen über die voraussichtliche Bevölkerungsentwicklung. So wurde das AfS von der Gemeindeverwaltung damit beauftragt, zeitnah eine Berechnung auf der neuen Datengrundlage des Zensus 2011 zu erarbeiten.

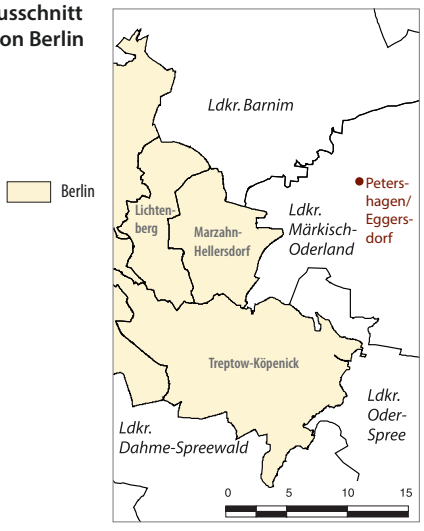
Nachdem die Bevölkerungszahl von Petershagen/Eggersdorf seit Beginn der 1990er Jahre stetig angestiegen ist, und zwar von gut 8 000 Einwohnern auf über 14 000 Einwohner (+ 68 %), sind beispielsweise die Fragen zu beantworten, ob sich diese Entwicklung weiter fortsetzen wird, mit welcher Zahl von Kindern im Vorschul- und Schulalter zu rechnen ist und wie viele Seniorinnen und Senioren es künftig geben wird.

Die für eine Vorausberechnung zu treffenden Annahmen über die zu erwartende Entwicklung der Einflussgrößen der Bevölkerungszahl (Geburten, Sterbefälle, Zu- und Fortzüge) wurden vom AfS gemeinsam mit Vertretern der Gemeinde getroffen. Es wurden drei Szenarien entwickelt, die zu drei Prognosevarianten führten: einer oberen, einer mittleren und einer unteren Variante. Die Varianten unterscheiden sich in der Höhe der erwarteten Zuzugsgewinne. Eine Entwicklung der Bevölkerungszahl zwischen der oberen und der unteren Variante wird für wahrscheinlich gehalten.

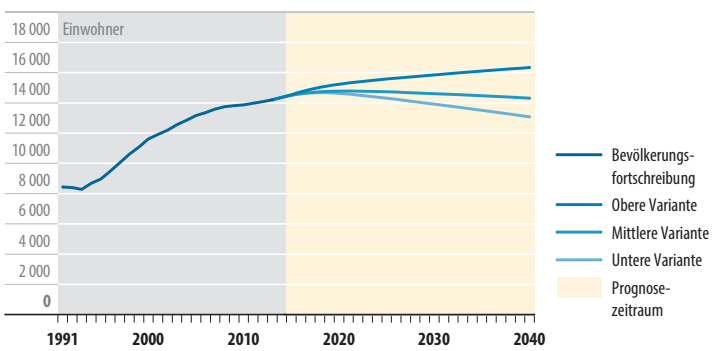
Wenn sich die getroffenen Annahmen erfüllen, dann wird die Gesamtbevölkerung von Petershagen/Eggersdorf bei allen drei Varianten vom Basisjahr 2013 zumindest bis zum Jahr 2018 weiter wachsen (Abbildung b). Bei der oberen Variante setzt sich das Wachstum bis zum Ende des Prognosezeitraumes 2040 weiter fort und liegt bei einer Zahl von über 16 000 Einwohnern. Bei der mittleren Variante wächst die Bevölkerungszahl noch bis zum Jahr 2021 an und geht dann bis 2040 auf das Ausgangsniveau des Basisjahres 2013 zurück. Beim Eintreffen der Annahmen der unteren Variante geht die Bevölkerungszahl nach 2018 kontinuierlich auf das Niveau von 2005 zurück.

Jürgen Paffhausen leitet das Referat *Bevölkerungs-, Kommunal- und Regionalstatistik* im Amt für Statistik Berlin-Brandenburg.

a | **Kartenausschnitt östlich von Berlin**



b | **Bevölkerungsentwicklung von Petershagen/Eggersdorf**



Kurzbericht

Tagungsbericht: Big Data – Big Brother oder Big Chances?

Symposium der Deutschen Arbeitsgemeinschaft Statistik an der Beuth Hochschule für Technik Berlin

von **Hartmut Bömermann**

Für den 24. April 2015 hatte die Deutsche Arbeitsgemeinschaft Statistik (DAGStat) zu einem Symposium an die Beuth Hochschule in Berlin-Mitte eingeladen. Begrüßt wurden die Teilnehmer vom Ersten Vizepräsidenten Professor Dr. Hans Gerber. An der Hochschule sind im Wintersemester 2014/15 insgesamt 12 519 Studierende immatrikuliert, darunter etwa ein Drittel Studentinnen. Die Hochschule hat zu „Data Science" einen Arbeitsverbund eingerichtet, an dem neben Lehrenden der Fachbereiche *Mathematik* und *Informatik und Medien* Studierende und Industriepartner beteiligt sind. Die Beuth Hochschule gehört auch zum Konsortium des „Big Data Center Berlin". Die Vorsitzende der DAGStat, Professorin Dr. Christine Müller, bedankte sich für die Gastfreundschaft der Beuth Hochschule und den zur Verfügung gestellten Tagungsraum, eine geräumige ehemalige Maschinenhalle.

In ihren Eingangsworten umriss die Vorsitzende der DAGStat die Aufgaben der Arbeitsgemeinschaft, die 2005 gegründet wurde, um die Wissenschaft mit Anwenderinnen und Anwendern zusammenzubringen und ihnen ein Forum zu bieten. Die DAGStat veranstaltet hierzu jährlich Symposien. Mit der diesjährigen Veranstaltung zu Big Data feierte die Arbeitsgemeinschaft zugleich ihr 10-jähriges Jubiläum.

© **Foto:** Peter Schaar

Podiumsdiskussion

v. l.: Prof. Dr. Klaus-Robert Müller (TU Berlin), Prof. Dr. Ralf Wagner (Universität Kassel), Peter Schaar (EAID), Dr. Susanne Schnorr-Bäcker (Destatis), Hartmut Bömermann (AfS) und Prof. Dr. Christine Müller (TU Dortmund)

Big Data wird in der Regel durch ein 3- oder 5-V-Modell charakterisiert; es und zeichnet sich aus durch: Datenmenge *(volume)*, Datenvielfältigkeit *(variety)*, Geschwindigkeit des Datenanfalls und der -verarbeitung *(velocity)*, Veränderlichkeit der Daten *(variability)* und die Qualität bzw. Richtigkeit dieser Daten *(veracity)*.

Für viele Beobachter scheint es noch nicht entschieden, ob Big Data nur ein Marketinghype oder eine echte Innovation sein wird. Und wenn es eine bedeutende Innovation sein sollte, was davon zu halten sei. Nicht zuletzt die NSA-Affäre hat dazu beigetragen, die schier unglaublichen Speicher- und Verarbeitungskapazitäten einer breiten Öffentlichkeit bewusst zu machen, die das bisher kaum Vorstellbare um Längen überbieten. Damit ist diese nächste Stufe der Datenspeicherung, Verarbeitung und Nutzbarmachung aber mit der Dystopie des „Big Brother" kontaminiert. Zuletzt berichtete das Institut für Demoskopie in Allensbach über die kognitive Dissonanz zwischen der Sorge um den Datenschutz und einer gleichzeitig steigenden Nutzung von Internetdiensten. Im Symposium wurde dieses Spannungsfeld aufgezogen, um Chancen und Risiken und deren Einhegung diskutieren zu können.

Im Vortrag „Maschinelles Lernen und ‚Big Data'" gab Professor Dr. Klaus-Robert Müller, Technische Universität Berlin, zunächst eine kurze Einführung in das Maschinelle Lernen und die unterschiedlichen Eigenschaften von Support Vektor Maschinen bzw. neuronalen Netzen bei der Verarbeitung sehr großer Datenmengen. Als Beispiel wurde das „Berlin Brain-Computer Interface" (BBCI) vorgestellt. Dieses Interface koppelt Aktivitäten der motorischen Rinde des Zentralen Nervensystems über EEG-Elektroden, die auf der Kopfhaut fixiert werden, mittels einer interpretierenden Verarbeitungslogik mit externen Geräten, z. B. einem Monitor, dessen Cursorbewegung gesteuert werden kann. Nach einer kurzen Lernphase kann die Probandin/der Proband das externe Gerät über die abgegriffenen Messwerte steuern. Vom EEG werden sehr hohe Datenflüsse erzeugt. Das gesuchte Signal – die an die Daten gerichtete Frage – wird mit der Hilfe von Verfahren des maschinellen Lernens im breiten und komplexen Datenstrom identifiziert. Das Besondere ist, dass das Lernen dabei auf die Maschine verlagert wird, wodurch die Trainingsphase sehr stark verkürzt werden kann. Naheliegende Anwendungsfälle sind Patienten nach einem Schlaganfall oder mit einem Locked-in-Syndrom.

Der Vortrag „Krebsepidemiologie – vom Kleingewerbe zu internationalen Konsortien" von Professor Dr. Rudolf Kaaks, Deutsches Krebsforschungszentrum (DKFZ), Nationale Kohorte, gab Einblicke in ein sehr komplexes Forschungsprojekt. Statt in Fall-Kontrollstudien die Wirkung von spezifischen Risikofaktoren auf einzelne Krebstypen zu untersuchen, werden bei diesem prospektiven Design für eine Untersuchungsgruppe Befragungsdaten, Ergebnisse medizinischer Untersuchungen, Analysen biologischer Einzelproben, Versicherungsdaten, Geotagging und andere Quellen miteinander kombiniert, um Effekte einer großen Anzahl von (interagierenden) Risikodeterminanten analysieren zu können. Diese sehr groß angelegte, populationsbezogene Studie wird als gemeinsame Wissenschaftsinfrastruktur mit klar definierten Regeln betrieben. Interne und externe Wissenschaftler haben Zugang zu den Forschungsdaten. Der Datenschutz und die Schutzanforderungen an die Datenhaltung sind sehr hoch. Studienteilnehmer geben ihr Einverständnis in einem transparenten Verfahren und haben die Möglichkeit, es jederzeit zu widerrufen.

Über den bisherigen Stand in der amtlichen Statistik berichtete Dr. Susanne Schnorr-Bäcker, Statistisches Bundesamt (Destatis), in ihrem Vortrag „Big Data in der amtlichen Statistik – Möglichkeiten und Grenzen". Traditionell gewinnt die amtliche Statistik ihre Daten durch Primärerhebungen und aus Verwaltungsdaten. Die Nutzung von sekundärstatistischen Quellen, wie Verwaltungsdaten und Registern, wird ausgebaut, da die verstärkte Nutzung sekundärstatistischer Quellen die Last für Respondenten reduziert und die Effizienz der Datengewinnung und deren Aktualität erhöhen kann. Big Data entwickelt diesen Ansatz über die bisher diskutierten Szenarien hinaus weiter. Als neue relevante Datenquellen dienen frei zugängliche webbasierte Angebotsplattformen (Güter, Jobs, öffentliche Dienstleistungen), verteilte Sensoren, Funkzellendaten von Mobiltelefonen, Verhaltensdaten von Web 2.0-Plattformen oder Satellitenfotos. Eine bereichsübergreifende Beschäftigung mit Big Data ist ein strategischer Arbeitsschwerpunkt von Destatis. In einer Machbarkeitsstudie soll die Eignung von Web-Scraping-Techniken für die Preisstatistik untersucht werden. Das Statistische Bundesamt arbeitet auf europäischer Ebene in einer Task Force Big Data mit und ist in weitere inter- und supranationale Aktivitäten eingebunden.

„‚Ja, wie für mich gemacht!' Targeting von Kunden im interaktiven Marketing" überschrieb Professor Dr. Ralf Wagner, Universität Kassel, seinen Vortrag. Die Nutzbarmachung der Daten über die Produkt- und Dienstleistungssuche, das Kaufverhalten und die Nutzung von Medien ermöglichen eine gezielte Ansprache potenzieller Kunden, zu denen so eine Kommunikation aufgebaut werden kann, die den Präferenzen entspricht. Im Gegenzug verlieren intime Details der Lebensführung ihren opaken Status. Vertrauen ist aber eine zentrale Voraussetzung für eine gelingende Kommunikation.

Peter Schaar, Europäische Akademie für Informationsfreiheit und Datenschutz (EAID), rekurrierte in seinem Vortrag „Big Data, Statistik und Datenschutz – Lösungen in Sicht?" auf das Urteil des Bundesverfassungsgerichts zur Volkszählung aus dem Jahr 1983. Das Bundesverfassungsgericht hat klare Grundsätze im Umgang mit personenbezogenen Daten formuliert. Im Zentrum steht die informationelle Selbstbestimmung jedes Bürgers/jeder Bürgerin. Ist der heutige Datenschutz, der ein Grundrechtsschutz ist, aber noch zeitgemäß? Ist Big Data mit den Grundsätzen Datenvermeidung, Datensparsamkeit, Zweckbindung überhaupt vereinbar? Peter Schaar präsentierte in seinem Vortrag einen Forderungskatalog, und zwar: Transparenz der Verarbeitung und Bewertung, Privacy by Design, keine Datenverarbeitung als Selbstzweck, Verwendung anonymisierter/pseudonymer Daten, beschränkter Zugriff auf Einzelangaben, Ausschluss sensibler Daten, keine Bildung von Persönlichkeitsprofilen, keine algorithmischen Einzelentscheidungen, keine Diskriminierung.

In der anschließenden Podiumsdiskussion, die von Professorin Dr. Christine Müller und Hartmut Bömermann vom Verband Deutscher Städtestatistiker/Amt für Statistik Berlin-Brandenburg moderiert wurde, konnten die aufgeworfenen Fragen unter Beteiligung des Publikums vertieft werden. Big Data ist in seiner Vielfältigkeit und raschen Entwicklung ein überaus spannendes Gebiet für die Statistik und ein herausforderndes für Privatheit und Grundrechtsschutz. Der Ansatz, die Komplexität des Themas durch eine Skandalisierung (Stichwort NSA) reduzieren zu wollen, um so Übersichtlichkeit zu gewinnen, trägt nicht dazu bei, die Chancen und Risiken dieses wichtigen Zukunftsthemas zu verstehen und zu gestalten.

Hartmut Bömermann leitet die Abteilung *Bevölkerung und Regionalstatistik* des Amtes für Statistik Berlin-Brandenburg. Zu seinen Arbeitsschwerpunkten gehören die Gebiete Sozialstrukturanalyse und Methoden raumbezogener Statistik.

Links

- Deutsche Arbeitsgemeinschaft Statistik (DAGStat): www.dagstat.de
- Beuth Hochschule für Technik Berlin: www.beuth-hochschule.de
- Machine Learning/Intelligent Data Analysis, Technische Universität Berlin: www.ml.tu-berlin.de
- Berlin Brain-Computer Interface: www.bbci.de
- Berlin Big Data Center: www.bbdc.berlin
- Deutsches Krebsforschungszentrum (DKFZ): www.dkfz.de
- Internationales Direktmarketing der Universität Kassel: www.uni-kassel.de/fb07/institute/ibwl/personen-fachgebiete/wagner-prof-dr/home.html
- Europäische Akademie für Informationsfreiheit und Datenschutz (EAID): www.eaid-berlin.de
- Statistisches Bundesamt: www.destatis.de
- FAZ (16.4.2015): „Abgehängt in der schönen neuen Welt", URL: www.faz.net/aktuell/politik/inland/leben-und-arbeiten-mit-dem-internet-in-deutschland-13540014.html

Kurzbericht

Klimaneutrales Berlin 2050

Das Berliner Energie- und Klimaschutzabkommen – amtliche Statistik als Datengrundlage

von **Mathias Geburek**

Der Berliner Senat ist sich seiner Verantwortung gegenüber dem Klimawandel bewusst und bereitet den Umstieg zu einer effizienten Energieversorgung aus erneuerbaren Energien vor. Zum einen soll der Energieverbrauch allgemein gesenkt und zum anderen sollen die CO_2-Emissionen auf ein niedrigeres Niveau reduziert werden. In der Machbarkeitsstudie „Klimaneutrales Berlin 2050", welche im April 2014 vorgestellt wurde, erfolgte bereits die Untersuchung verschiedener Möglichkeiten zu einer klimafreundlicheren Stadt. Im nächsten Schritt findet nun die Erarbeitung eines Berliner Energie- und Klimaschutzprogramms (BEK) statt, das auf den 16. Berliner Energietagen (27.–29. April 2015) von Staatssekretär Christian Gaebler (Senatsverwaltung für Stadtentwicklung und Umwelt) angekündigt wurde.

Damit die Stadt klimaneutral werden kann, wurden in der Machbarkeitsstudie Energie- und CO_2-Einsparziele gesetzt sowie Maßnahmen genannt, wie diese Ziele bis zum Jahr 2050 erreicht werden können. Zu den wichtigsten Kontrollwerten zählen die CO_2-Emmissionen pro Einwohner, da diese unabhängig von der Bevölkerungsentwicklung sind. „Klimaneutral" wurde für Berlin definiert als die Reduktion der CO_2-Emissionen pro Einwohner bis auf ein Niveau, das „das Weltklima unterhalb der gefährlichen Schwelle einer Erwärmung von 2 Grad halten kann" [1]. Laut Machbarkeitsstudie bedeutet dies in konkreten Zahlen einen Ausstoß von 2 Tonnen Treibhausgas (CO_2-Äquivalente[1]) pro Einwohner. Absolut würde das einer Absenkung der reinen CO_2-Emissionen nach Verursacherbilanz[2] auf 4,4 Mill. Tonnen bedeuten. Dies entspricht einer Reduzierung um 85 % gegenüber dem Jahr 1990 [2].

Zur Berechnung der CO_2-Emissionen in Berlin bietet die jährliche Energie- und CO_2-Bilanz des Amtes für Statistik Berlin-Brandenburg (AfS) die Datengrundlage, welche bis zum Jahr 1990 zurückreicht. Veröffentlicht werden darin Zahlen zum Energieverbrauch – unterteilt nach Energieträgern und Verbrauchergruppen –, Kennzahlen zur Strom- und Fernwärmeerzeugung sowie errechnete Werte für die Kohlenstoffdioxid-Emissionen nach Emittenten.

Auch der städtische Gasversorger, die GASAG AG, beschäftigt sich intensiv mit der Energiewende in Berlin und will ebenfalls einen Beitrag zur klimaneutralen Stadt leisten. Zu den Berliner Energietagen wurden hierfür Vertreter der Region Rhein-Neckar sowie der Stadt Bottrop eingeladen. Beide Regionen befinden sich seit einiger Zeit im energetischen Umbruch und planen eine Energieversorgung auf Basis von erneuerbaren Energien.

Für die GASAG AG ist die Modernisierung des Wärmemarktes in Berlin von zentraler Bedeutung. Wie aus der Berliner Energie- und CO_2-Bilanz 2012 hervorgeht, liegt der Anteil des Sektors Haushalte, GHD[3] und übrige Verbraucher am Endenergieverbrauch bei 68,4 %[4] [2]. „Der entscheidende Hebel für die Energiewende liegt im Wärmemarkt", hieß es auf der Veranstaltung. Einsparungen beim Heizenergieverbrauch werden sich letztlich im Sektor Haushalte, GHD und übrige Verbraucher zeigen. Das Einsparen von Endenergie führt direkt zu einem geringeren CO_2-Ausstoß und somit zu einer klimafreundlicheren Stadt. Bei der Podiumsdiskussion wurden anschließend Maßnahmen genannt, mit denen ein Umstieg auf erneuerbare Energien gelingen kann. Ein zentraler Ansatz in beiden Regionen ist ein Zusammenschluss wichtiger Akteure auf dem Erzeuger- und Verbrauchermarkt zu einem Informationsnetzwerk. Hierzu zählen die örtlichen Energieversorger, Planungsbüros, Handwerker, Hochschulen, Unternehmen sowie die Bürgerinnen und Bürger. Ebenso wurden sogenannte „Leuchtturmprojekte" angeführt, welche für die Bevölkerung als Vorreiter gelten sollen. Berlin verankert im neuen Energiewende-Gesetz ebenfalls solche Projekte mit Vorbildfunktion. Dazu heißt es im Entwurf des Energiewendegesetzes in § 7 Absatz 1: „Das Land Berlin setzt sich zum Ziel, den Kohlendioxidausstoß der Landesverwaltung … bis zum Jahr 2030 weitgehend auszugleichen und diese somit CO_2-neutral zu organisieren" [3].

Die zentrale Aussage dieser Veranstaltung lautete, dass sich ein Umstieg auf erneuerbare Energien nur im Dialog mit den Bürgerinnen und Bürgern sowie in der Zusammenarbeit aller Akteure verwirklichen lässt. Die gewonnenen Konzepte müssen jetzt in reale Projekte umgesetzt werden, damit Einsparungen in der Energie- und CO_2-Bilanz sichtbar werden.

1 CO_2-Äquivalente: Andere Treibhausgase (z. B. Methan [CH_4] oder Lachgas [Distickstoffmonoxid, N_2O]) werden gemäß ihrem spezifischen Beitrag zur globalen Erwärmung in das Erwärmungspotenzial von CO_2 umgerechnet.

2 Verursacherbilanz: eine auf den Endenergieverbrauch bezogene Darstellung der Treibhausgas-Emissionen.

3 GHD: Gewerbe, Handel, Dienstleistungssektor

4 Summe aller Energieträger, inklusive Stromverbrauch

Der Statistische Bericht „E IV 4 – j / 12 Energie- und CO_2-Bilanz im Land Berlin 2012" steht im Internetangebot des AfS im Excel- und PDF-Format zur Verfügung: www.statistik-berlin-brandenburg.de/publikationen/stat_berichte/2015/SB_E04-04-00_2012j01_BE.pdf

Mathias Geburek ist Sachbearbeiter im Referat *Verarbeitendes Gewerbe, Bergbau, Energie- und Wasserversorgung* des Amtes für Statistik Berlin-Brandenburg.

Literatur

[1] Senatsverwaltung für Stadtentwicklung und Umwelt (2014): Klimaneutrales Berlin 2050 – Ergebnisse der Machbarkeitsstudie. Berlin.

[2] Amt für Statistik Berlin-Brandenburg (2015): E IV 4 – j / 12 Energie- und CO_2-Bilanz im Land Berlin 2012. Potsdam.

[3] Entwurf Berliner Energiewendegesetz (2015): Entwurf eines Gesetzes zur Umsetzung der Energiewende und zur Förderung des Klimaschutzes in Berlin (Berliner Energiewendegesetz – EWG Bln), Stand 14.04.2014, abrufbar unter http://www.stadtentwicklung.berlin.de/umwelt/klimaschutz/energiewendegesetz/download/EnergiewendeG_Bln_GESETZESTEXT.pdf

Kurzbericht

Fachstatistische Veranstaltungen des AfS

von **Ricarda Nauenburg**

Das Amt für Statistik Berlin-Brandenburg (AfS) organisierte im Juni drei Veranstaltungen, die einen intensiven Austausch mit Datennutzerinnen und -nutzern aus Wissenschaft, Fachverwaltungen als auch anderen statistischen Ämtern zum Ziel hatten.

Am 2. Juni 2015 fand die diesjährige Fachtagung des AfS für die Brandenburger Statistikstellen statt. Themenschwerpunkte waren die Bevölkerungs- und Sozialstatistiken sowie die statistische Geheimhaltung. Jörg Fidorra, Vertreter des Vorstandes, begrüßte die Teilnehmerinnen und Teilnehmer. Auf der Tagesordnung standen in diesem Jahr Beiträge zum Stand der Bevölkerungsstatistik nach dem Zensus 2011, zur Reliabilität von Kreisergebnissen aus dem Mikrozensus, zu Möglichkeiten und Grenzen der Sozialstatistiken und zu neuen Entwicklungen bei der Geheimhaltung statistischer Daten. Eine Präsentation zum Migrationsmonitoring der Stadt Potsdam von Dr. Matthias Förster (Statistikstelle Potsdam) rundete die Fachtagung ab. Die Veranstaltungsreihe hat eine langjährige Tradition und dient dem Zweck, einem Fachpublikum das regionale Datenangebot des AfS für das Land Brandenburg bekanntzumachen, Datenwünsche entgegenzunehmen und Informationen zwischen Statistikproduktion und Statistiknutzerinnen und -nutzern auszutauschen. Während der intensiven Diskussion der Tagungsbesucher ergaben sich bereits inhaltliche Anregungen für die Vorbereitung der nächsten Fachtagung im Frühjahr 2016. Gleichzeitig wurde angestoßen, das Format auch für Berlin anzubieten.

In der vom AfS organisierten Veranstaltungsreihe „Messung der Preise" fand die 19. Konferenz als gemeinsame Veranstaltung mit dem Statistischen Landesamt Mecklenburg-Vorpommern am 16. und 17. Juni 2015 in Schwerin statt. Hier trafen sich Preisstatistiker mit den Nutzern von Preisstatistiken aus Banken, Wirtschaft und Wissenschaft, um sich über neue Entwicklungen der theoretischen Grundlagen sowie Fragen der Durchführung der Preisstatistik als auch über Ergebnisse und Erfahrungen bei der Nutzung der Preisstatistiken in Wirtschaft und Wissenschaft auszutauschen. Das Themenspektrum der diesjährigen Veranstaltung reichte dabei von Wohnimmobilienpreisen über Modelle zu regionalen Preisvergleichen bis hin zu praktischen Erörterungen wie z. B. den Einfluss des Mindestlohns auf die Verbraucherpreise.

Informationen zur Konferenz sowie die Vorträge sind abrufbar unter: https://www.statistik-berlin-brandenburg.de/home/messung-der-preise.asp

Auch die Teilnehmerinnen und Teilnehmer des „8. Berliner VGR-Kolloquiums" am 18. und 19. Juni 2015 in Berlin kamen aus verschiedensten nationalen und internationalen statistischen Ämtern, Behörden und wissenschaftlichen Institutionen. Diese Veranstaltungsreihe widmet sich bereits seit den 1990er Jahren den theoretischen Voraussetzungen und konzeptionellen Grundlagen der Systeme Volkswirtschaftlicher Gesamtrechnungen (VGR). Schwerpunkt der diesjährigen Veranstaltung war die Finanzierungsrechnung in der VGR. Neuerungen in der Finanzierungsrechnung wurden in Vorträgen von Vertretern der Europäischen Zentralbank, der Österreichischen Nationalbank sowie der Bundesbank vorgestellt.

Informationen zum Kolloquium sowie die Vorträge sind abrufbar unter: www.statistik-berlin-brandenburg.de/home/vgr-kolloquium.asp

Ricarda Nauenburg ist Leiterin des Referates *Mikrozensus, Sozialberichte* des Amtes für Statistik Berlin-Brandenburg.

Entwicklungen in der amtlichen Statistik

Atlas der Wirtschaftseinheiten

von **Thomas Heymann**

Seit einigen Monaten bietet das Amt für Statistik Berlin-Brandenburg (AfS) auf seiner Internetpräsenz (www.statistik-berlin-brandenburg.de) unter der Rubrik „Interaktive Karten" interaktive Atlanten auf der Grundlage regionaler Daten des Zensus 2011 und des Sozialberichts 2013 für Berlin und Brandenburg an. Kürzlich wurde das Angebot durch den Atlas der Wirtschaftseinheiten 2014 ergänzt. Dieser Atlas beinhaltet Karten der Gewerbeanzeigenstatistik, des Unternehmensregisters und der Insolvenzstatistik auf Ebene der Landkreise und der kreisfreien Städte für Brandenburg sowie auf Ebene der Bezirke für Berlin.

Die einzelnen Karten können aus Drop-Down-Menüs der Rubrik „Thema" im Kopf der Seite ausgewählt werden (Abbildung a). Der Seitenaufbau des Atlasses ist immer gleich gestaltet: Neben der Ansicht der ausgewählten Karte ist die Kartenlegende und darunter die Erläuterung über ihre Inhalte zu finden. Rechts davon wird die statistische Übersicht über die Werte der Gebietseinheiten angezeigt. Das untere Ende dieser Hälfte der Atlasseite füllt eine Grafik. Hier kann zwischen einem Säulendiagramm mit Größenverhältnissen der jeweiligen Merkmale der Gebietseinheiten und einer Kurvendarstellung im Falle von Zeitreihen für Berlin und Brandenburg gewählt werden.

Die Gewerbeanzeigenstatistik liefert monatliche Ergebnisse. Für den Atlas der Wirtschaftseinheiten werden die Jahresergebnisse seit 2008 ausgewertet. Abbildung b zeigt die in der Gewerbeanzeigenstatistik ausgewiesenen Betriebsgründungen. Dieses Merkmal ist eine Untermenge der Neugründungen und soll mit seinem Kontrapart der Betriebsaufgaben das unternehmerische Gründungsgeschehen wiedergeben. Abbildung c zeigt die Betriebsaufgaben in der Zeitreihenansicht.

Wenn auf der Karte eine Gebietseinheit (hier als Beispiele der Bezirk „Mitte" von Berlin und der Landkreis „Prignitz" von Brandenburg) markiert ist, wird mit einer Kurvengrafik die Entwicklung des jeweiligen Merkmals (hier die Betriebsaufgaben)

a | Inhaltsangabe der Karten unter der Rubrik „Thema"

angezeigt. Es können zusätzlich weitere Kreise oder Bezirke ausgewählt und markiert werden. Eine weitere Funktion ist das interaktive „Abspielen" der Zeitreihe, indem auf das Pfeilsymbol ▸ am linken unteren Rand der Karte geklickt wird. Gleichzeitig baut sich der Kurvenverlauf vom Startjahr ausgehend in der Grafik neu auf.

Ein Vergleich ausgewählter Gebietseinheiten für unterschiedliche Indikatoren und Jahresstände ist möglich, da die Auswahl über alle Karten und Statistiken des Atlasses erhalten bleibt. Weiterhin kann eine Zoom-Funktion oder eine Filterung nach Gebietseinheiten aktiviert werden.

Die in Abbildung d aufgezeigte Variante ist ein Kartenbeispiel aus dem Fundus des statistischen Unternehmensregisters. Neun thematische Karten stellen die Verteilung von Unternehmen und Betrieben sowie der sozialversicherungspflichtig Beschäftigten in Berlin und Brandenburg vor.

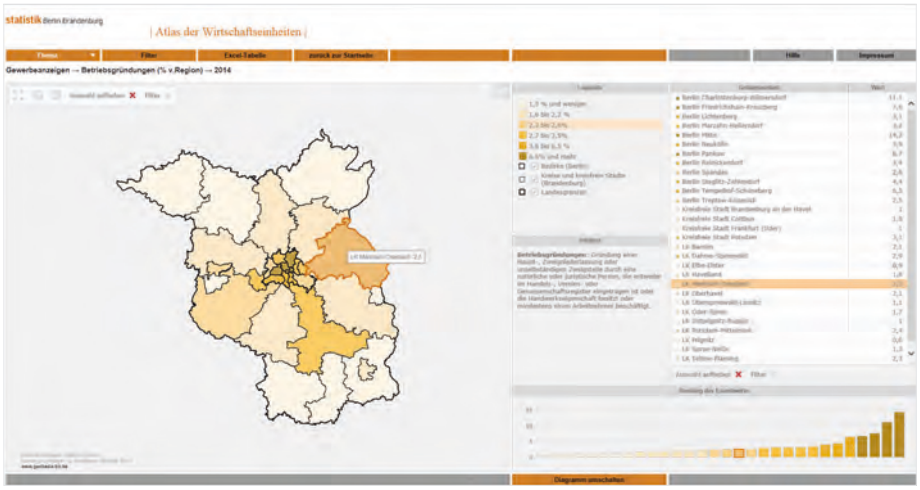
Außerdem werden für sieben zusammengefasste Wirtschaftsbereiche vom „Produzierenden Gewerbe" über „Verkehr, Information und Kommunikation" bis „Kultur, Freizeit, Sport, sonstige Dienstleistungen" mit jeweils drei thematischen Karten Informationen zu Betriebsgröße, Betriebsdichte und sozialversicherungspflichtig Beschäftigten je km^2 vorgestellt. Grundlage für die Unterscheidung der wirtschaftlichen Aktivitäten ist die „Klassifikation der Wirtschaftszweige 2008".

Als letzte Statistik dieses Atlasses kann die Verteilung von eröffneten Insolvenzverfahren von verschuldeten Unternehmen und Verbrauchern und ihre Veränderung seit 2006 betrachtet werden (Abbildung e).

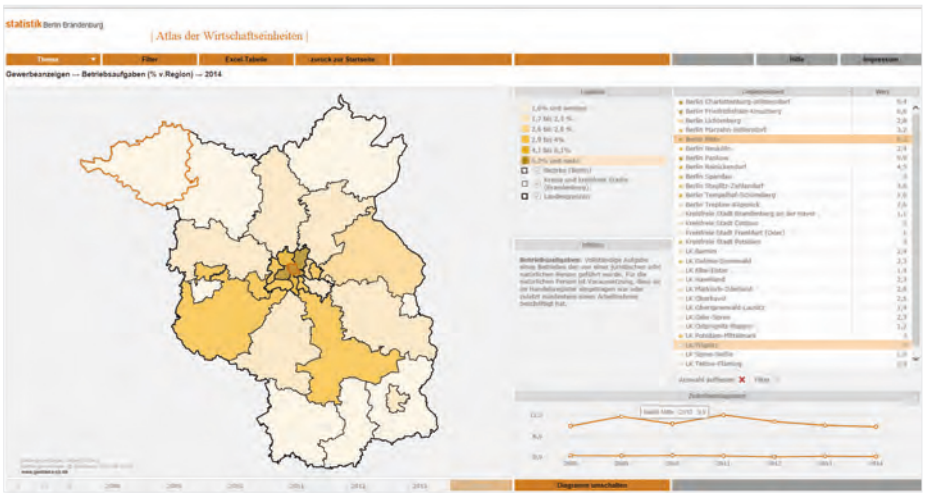
Interessierte Nutzerinnen und Nutzer können sich alle Daten im Excel-Format herunterladen.

Mit den interaktiven Karten erweitert das Amt für Statistik Berlin-Brandenburg sein Angebot, die

b | **Betriebsgründungen in Berlin und Brandenburg 2014**



c | **Betriebsaufgaben in Berlin und Brandenburg seit 2008**

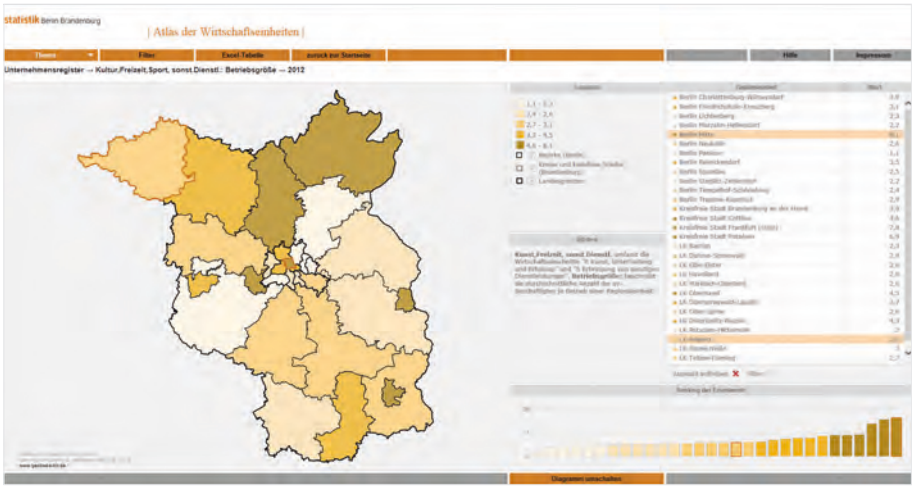


amtliche Statistik für regionale Betrachtungen zu verwenden. Wie die Zensusatlanten für Berlin und Brandenburg demonstrieren, können auch kleinere Gebietseinheiten (z. B. Statistische Gebiete, LOR, Mittelbereiche) in den Karten dargestellt werden, sofern die Erfordernisse der Geheimhaltung es zulassen. Der interaktive Atlas ist ausbaufähig und wird in Zukunft um weitere Statistiken und regionale Betrachtungen ergänzt.

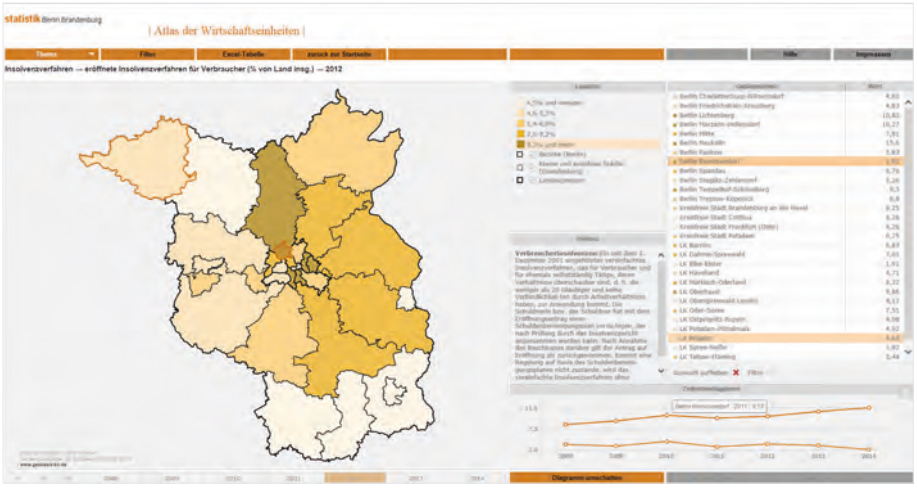
Dr. Thomas Heymann leitet das Referat *Unternehmensregister, Gewerbeanzeigen, Insolvenzen* des Amtes für Statistik Berlin-Brandenburg.

Der Atlas der Wirtschaftseinheiten 2014 steht zur Verfügung unter:
https://www.statistik-berlin-brandenburg.de/instantatlas/interaktive-karten.asp

d | Unternehmensregister: Betriebe aus dem Wirtschaftsbereich „Kultur, Freizeit, Sport, sonstige Dienstleistungen" in Berlin und Brandenburg 2012 nach Betriebsgröße



e | Eröffnete Insolvenzverfahren für Verbraucher in Berlin und Brandenburg seit 2008



Entwicklungen in der amtlichen Statistik

Berufsqualifikationsfeststellungsgesetz soll Anerkennungen erleichtern

von **Andreas May-Wachowius**

Zum 1. April 2012 trat das Berufsqualifikationsfeststellungsgesetz (BQFG)[1] in Kraft. Ziel des Gesetzes ist es, Menschen mit im Ausland erworbenen Berufsqualifikationen den Zugang zum deutschen Arbeitsmarkt zu erleichtern beziehungsweise sogar erst zu ermöglichen. Nicht nur den Migrantinnen und Migranten, die einen Großteil der Antragsteller darstellen, soll dieses Gesetz dienlich sein, sondern auch das Problem des mit dem demographischen Wandel einhergehenden Fachkräftemangels in Deutschland mildern. Die Implementierung einheitlicher Bewertungs- und Anerkennungsverfahren soll unter Berücksichtigung der Besonderheiten einzelner Berufsgruppen für eine Ausweitung, Vereinfachung und Verbesserung dieser Verfahren sorgen. Die Verabschiedung dieses Gesetzes war die Umsetzung der EU-Berufsanerkennungsrichtlinie[2] aus dem Jahr 2005 in nationales Recht.

Seit April 2012 erheben die Statistischen Ämter der Länder die bundesrechtlich geregelten Berufe. In Berlin und Brandenburg wurden 871 entsprechende Verfahren im ersten vollständig erhobenen Berichtsjahr 2013 abgeschlossen. Bei 83 % davon wurde eine volle Gleichwertigkeit der Berufsqualifikation festgestellt. Bundesweit wurden im selben Zeitraum 13 344 Verfahren abgeschlossen, 89 % davon mit Anerkennung der vollen Gleichwertigkeit.

Zu den in dieser Statistik erhobenen Merkmalen gehören neben den Angaben zum Referenzberuf und zur Entscheidung bezüglich der Anerkennung auch Angaben zu Herkunft und Ausbildungsstaat des Antragstellers sowie zum zeitlichen Verlauf des Verfahrens und gegebenenfalls zu gegen die Entscheidung eingelegtem Rechtsbehelf.

Da es aber auch nach Landesrecht geregelte Berufe gibt, waren die Bundesländer angehalten, mit entsprechenden Landesgesetzen nachzuziehen. Das Inkrafttreten der 16 Landesgesetze erstreckte sich über den Zeitraum von August 2012 bis Juli 2014.

Für die amtliche Statistik besteht die Herausforderung darin, die 16 Landesstatistiken zu einer koordinierten Länderstatistik zusammenzuführen. Leider sind die notwendigen Voraussetzungen noch nicht in allen Bundesländern geschaffen: So wird das Amt für Statistik Berlin-Brandenburg zwar die Daten der Brandenburger Landesstatistik an das Statistische Bundesamt liefern, die Berliner Daten mangels gesetzlicher Grundlage jedoch nicht.

Wer seine im Ausland erworbene Berufsqualifikation anerkennen lassen möchte, findet im „Anerkennungsfinder" des Bundesministeriums für Bildung und Forschung unter https://www.anerkennung-in-deutschland.de fundierte Informationen zur Umsetzung des Vorhabens.

Berliner und Brandenburger Daten des BQFG können auch in der StatIS-Datenbank des Amtes für Statistik Berlin-Brandenburg unter https://www.statistik-berlin-brandenburg.de/datenbank/inhalt-datenbank.asp im Sachgebiet Bildung und Kultur abgerufen werden. Dieses Internetangebot ermöglicht interessierten Nutzerinnen und Nutzern die Erstellung flexibler Tabellen mit den BQFG-Mikrodaten.

Andreas May-Wachowius ist Sachgebietsleiter im Referat *Schule Berlin, Bildungsanalysen* im Amt für Statistik Berlin-Brandenburg.

1 Gesetz über die Feststellung der Gleichwertigkeit von Berufsqualifikationen vom 6. Dezember 2011 (BGBl. I S. 2515), geändert durch Artikel 23 des Gesetzes vom 25. Juli 2013 (BGBl. I S. 2749).

2 Richtlinie 2005/36/EG des Europäischen Parlaments und des Rates vom 7. September 2005 über die Anerkennung von Berufsqualifikationen (ABl. L 255, S. 22), zuletzt geändert durch die Richtlinie 2013/55/EU des Europäischen Parlaments und des Rates vom 20. November 2013 (ABl. 354, S. 132).

Entwicklungen in der amtlichen Statistik

Einsatz von Rasterkarten im Rahmen des Zensus 2011

von **Tobias Kirchner**

Für Datenauswertungen in der amtlichen Statistik stehen in zunehmendem Maße nicht nur administrative Einheiten, wie Landkreise oder Gemeinden, sondern auch Gitterzelleninformationen zur Verfügung. Die räumliche Verortung des Erhebungsgegenstandes und somit die Zuordnung zu den Gitterzellen erfolgt über georeferenzierte Koordinatenpaare. Die so gewonnenen Informationen sollen nicht nur in Form von Tabellen und statischen Karten verarbeitet, sondern auch für interaktive Auswertungszwecke einem erweiterten Nutzerkreis zur Verfügung gestellt werden. Dem folgend konnte im Jahr 2014 der Agraratlas, in dem u. a. Betriebsgrößen und Flächenanteile angebauter Nutzpflanzen aus Erhebungsdaten der Landwirtschaftszählung 2010 visualisiert werden, mit einer Rasterweite von fünf Kilometern als erste bundesweite Anwendung auf Gitterzellenebene online gehen.[1]

Ende April 2015 wurde eine Onlineanwendung mit Ergebnissen des Zensus 2011 durch die Statistischen Ämter des Bundes und der Länder veröffentlicht.[2] Die dem Zensusatlas zugrunde liegenden Gitterzellen sind INSPIRE-konform und weisen eine Auflösung von 1 km² auf.

Um eine zeitnahe Verfügbarkeit der Daten zu gewährleisten, wurde zur Erstellung des Zensusatlas weitgehend auf die programmtechnische Lösung des Agraratlas zurückgegriffen. Erweitert wurde die Applikation um eine regionale Suchfunktion sowie um die Hintergrundkarte WebAtlasDE des Bundesamtes für Kartographie und Geodäsie, die ab einer Auflösung von 1:150 000 dargestellt wird. Beide Funktionen dienen einer erhöhten Übersichtlichkeit und besseren Navigation in der Karte.

Zu den zehn im Rahmen einer Bund-Länder-Arbeitsgruppe (AG) abgestimmten Indikatoren gehören neben der Bevölkerungsanzahl und dem Durchschnittsalter auch die Leerstandsquoten der Wohnungen. Neben der Prämisse, für die Öffentlichkeit interessante Indikatoren darzustellen, wurde insbesondere Wert auf eine sinnvolle Klasseneinteilung bei geringstmöglichem Informationsverlust gelegt. Für die Ausweisung der Einwohnerzahl sind im Hinblick auf die Geheimhaltung marginale Veränderungen der Originalwerte vorgenommen worden. So wurden Gitterzellen, in denen nur eine Person wohnt, als unbewohnt klassifiziert und somit der Kategorie „unbewohnt oder geheim zu halten" zugeordnet. Gitterzellen mit zwei Personen laut Zensus 2011 werden mit drei Personen ausgewiesen und fallen somit in die Kategorie „3 bis unter 250 Einwohner pro km²". Die anderen Indikatoren weisen bundesweit Geheimhaltungsquoten zwischen 0,1 % aller Gitterzellen für den Anteil der Bevölkerungsgruppe unter 18 Jahren an der Gesamtbevölkerung und 1,8 % für die Leerstandsquo-

a | **Variante 1 der Klassenbesetzung für den Indikator „Altersdurchschnitt in Jahren" im Raster 1x1 km im Bundesgebiet** (Diskussionsgrundlage Destatis vom 27.01.2015)

Klasse	Absolut	Relativ (in %)
Wert geheim	1 338	0,4
Unbewohnt	146 845	40,6
0 - < 30	5 743	1,6
30 - < 40	46 130	12,8
40 - < 47	108 454	30,0
47 - < 57	43 228	12,0
57 - 100	9 740	2,7

b | **Variante 2 der Klassenbesetzung für den Indikator „Altersdurchschnitt in Jahren" im Raster 1x1 km im Bundesgebiet** (Diskussionsgrundlage Destatis vom 27.01.2015)

Klasse	Absolut	Relativ (in %)
Wert geheim	4 219	1,2
Unbewohnt	146 845	40,6
0 - < 40	51 679	14,3
40 - < 42	29 197	8,1
42 - < 44	34 693	9,6
44 - < 47	44 084	12,2
47 - 100	50 761	14,0

1 http://www.atlas-agrarstatistik.nrw.de/
2 https://atlas.zensus2011.de/

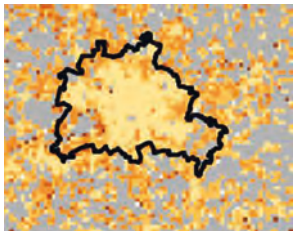
te auf. Dabei hat sowohl die Klassenanzahl als auch die Wahl der Klassengrenzen Einfluss auf die Menge der Geheimhaltungsfälle und auf die kartographische Darstellung. Abbildungen a und b zeigen unterschiedliche Klassifizierungen des Indikators „Altersdurchschnitt in Jahren". Aus Abbildung a wird ersichtlich, dass bei einer Geheimhaltungsquote von lediglich 0,4 % ca. die Hälfte aller bewohnten Gitterzellen auf die Klasse 40 bis unter 47 entfällt. Dies führt zu einer relativ homogenen kartographischen Darstellung mit fünf Werteklassen, in der lediglich Extremwerte sichtbar sind. Um eine differenzierte kleinräumige Verteilung aufzuzeigen, eignet sich die Klassifizierung in Abbildung b besser, auch wenn hierbei eine höhere Anzahl an Werten geheim zu halten ist. Diese Klassifizierung wurde für den Zensusatlas verwendet.

Bei der Einteilung der Klassen wurde darauf geachtet, fachlich sinnvolle Abgrenzungen vorzunehmen, die für das gesamte Bundesgebiet aussagekräftig sind. Hierzu wurden für jeden Indikator mehrere Vorschläge erarbeitet und die daraus resultierenden Klasseneinteilungen und Kartenentwürfe in der AG abgestimmt. Hierbei zeigte sich, dass, je inhomogener das in einer Rasterung darzustellende Gebiet bezüglich der Ausprägungen eines Indikators war, desto schwieriger gestaltete sich die Klasseneinteilung. So erscheint für den Indikator der durchschnittlichen Wohnfläche pro Wohnung eine untere Klasse „unter 80 m^2" für die Darstellung des gesamten Bundesgebietes in der Karte zwar sinnvoll. Auch Klassenbesetzung und Geheimhaltungsfälle sind hier nicht als problematisch anzusehen. Jedoch ist die Aussagekraft der genannten Klasse in Ballungsgebieten, wie Hamburg oder Berlin (siehe Abbildung c), stark eingeschränkt, da hier jeweils mehr als 50 % aller Zellen der jeweiligen Stadtgebiete in die untere Kategorie fallen würden. Abbildung d hingegen zeigt die Klasseneinteilung mit einer unteren Klasse „unter 60 m^2", bei der für Berlin ein deutlich differenzierteres Bild gezeichnet wird.

Von den mehr als 360 000 Gitterzellen, die das Bundesgebiet mit einem 1x1 km-Raster in der Kartenprojektion ETRS89-LAEA abdecken, sind laut Zensuserhebung ca. 41 % unbewohnt bzw. ca. 44 % ohne Wohnraum. Diese sind in den Karten und Legenden der einzelnen Indikatoren jeweils als separate Klasse ausgewiesen. Um den Wiedererkennungswert von unbewohnten Rasterzellen bzw. Rasterzellen ohne Wohnraum zu erhöhen, wurde für diese durchgängig die Farbe Grau gewählt. Aufgrund der hohen Anzahl dieser Zellen ohne Zensusergebnis überwiegt allerdings die graue Farbgebung bei der Darstellung der Gesamtausdehnung der Karte (siehe Abbildung e). Für kleinräumige Betrachtungen lassen sich hingegen bewohnte und unbewohnte Gebiete bzw. Gebiete mit und ohne Ergebnisse des Zensus 2011 mit Hilfe dieser Darstellungsvariante gut erkennen, da sich das Grau von den im Zensus genutzten Farbspektren der Indikatoren gut unterscheiden lässt. Beispielhaft hierzu zeigt Abbildung f zwei Gitterzellen des ehemaligen Flughafens Tempelhof, die als „ohne Wohnraum" klassifiziert sind.

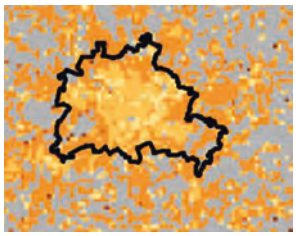
Für Gitterzellen, die nur teilweise in Deutschland liegen, sind im Zensusatlas jeweils Werte ausgewiesen, die sich lediglich aus dem deutschen Anteil der Gitterzelle berechnen. Für länderübergreifende Analysen ist folgerichtig zu beachten, dass ein und dieselbe Gitterzelle unterschiedliche Werte aufweisen kann. Die Problematik kann aufgrund des

c | Wohnfläche pro Wohnung in Berlin und Umland in m^2



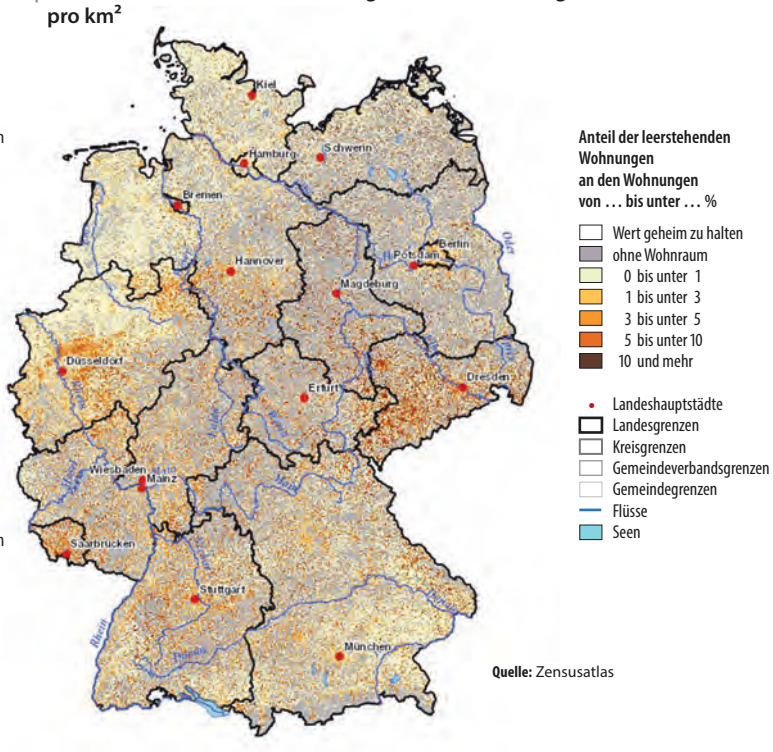
- Wert geheim zu halten
- ohne Wohnraum
- 0 bis unter 80
- 80 bis unter 100
- 100 bis unter 120
- 120 bis unter 140
- 140 und mehr

d | Wohnfläche pro Wohnung in Berlin und Umland pro m^2



- Wert geheim zu halten
- ohne Wohnraum
- 0 bis unter 60
- 60 bis unter 80
- 80 bis unter 140
- 140 bis unter 160
- 160 und mehr

e | Anteil der leerstehenden Wohnungen an den Wohnungen am 09.05.2011 pro km^2



Quelle: Zensusatlas

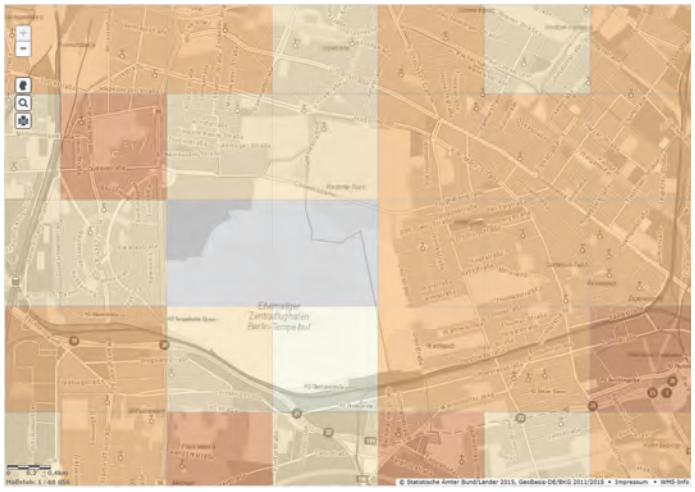
fehlenden Zugriffs auf Zensusdaten anderer Länder im Zensusatlas nicht gezeigt werden. Aus diesem Grund veranschaulicht Abbildung g die differierende Klassenzuweisung anhand des Grenzgebietes zwischen Berlin und Brandenburg. So ergeben sich für die beiden markierten Grids im Grenzgebiet von Berlin und Brandenburg drei unterschiedliche Werte, je nachdem ob eine Auswertung mit den Daten für Berlin (Abbildung g-1), Brandenburg (Abbildung g-2) oder mit einem kombinierten Datensatz beider Länder (Abbildung g-3) durchgeführt wird.

Neben der Variante der Darstellung sämtlicher von der Landesgrenze angeschnittenen Gitterzellen (alle farbigen Grids in Abbildung h), bestehen auch weitere Möglichkeiten, um Gitterzellenwerte in Grenzgebieten auszuweisen. So können Grids abgeschnitten werden (blaue Gitterzellen in Abbildung h) oder lediglich zur Darstellung gelangen, wenn ein Flächenanteil von mehr als 50 % einer anderen Raumeinheit – etwa zweier Bundesländer – erreicht ist.[3] Die in Abbildung h lila dargestellten Gitterzellen liegen komplett innerhalb der Berliner Stadtgrenze und stellen somit eine vierte Zuweisungsvariante dar.

Für die Zuweisung der Gitterzelle zu administrativen Einheiten, ist – die korrekte Kartenprojektion vorausgesetzt – der Generalisierungsgrad der Grenzgeometrien ein wichtiges Kriterium, da dieser darüber entscheiden kann, ob eine Gitterzelle einer Gemeinde, einem Bundesland o. Ä. zuzuordnen ist.

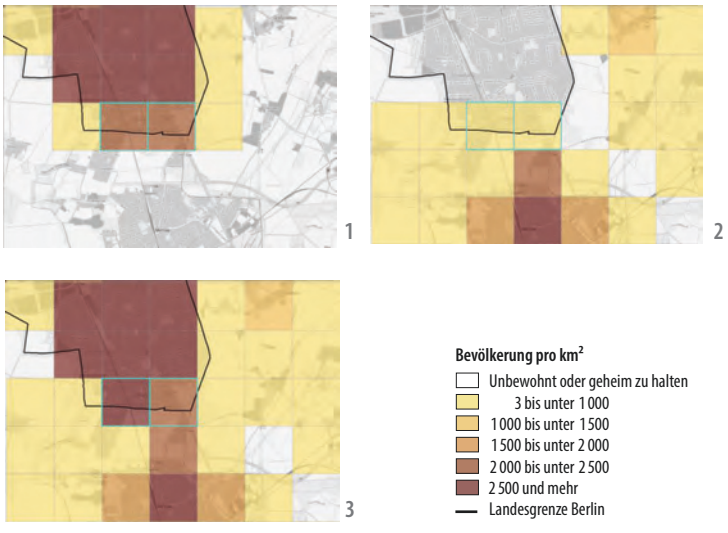
Da seit der Novellierung des Bundesstatistikgesetzes im Jahr 2013 die regionale Zuordnung von Erhebungsmerkmalen an geographische Gitterzellen erfolgen kann, ist künftig mit einer zunehmenden Anzahl an Indikatoren auf Gitterzellenebene aus der amtlichen Statistik zu rechnen.[4] Auch durch die Ausweitung der Open-Data-Portale[5] im Zuge der INSPIRE-Richtlinie wird sich in nächster Zeit nicht nur die Anzahl der Online-Anwendungen durch die datenführenden Institutionen erhöhen, sondern auch jene durch Drittanbieter und inte-

f | Anteil der leerstehenden Wohnungen an den Wohnungen in % am 09.05.2011 pro km² (Ausschnitt von Berlin)



Quelle: Zensusatlas

g | Differierende Klassenzuweisung für die Bevölkerungsdichte aus Daten des Zensus 2011 im 1x1 km-Raster an der Landesgrenze zwischen Berlin und Mahlow (Brandenburg)



3 Weist eine Gitterzelle Anteile an mehr als zwei Raumeinheiten auf, ist die Zelle entsprechend der Raumeinheit zuzuweisen, die den höchsten Flächenanteil in der Gitterzelle belegt.

4 Siehe Artikel 13 (Änderung des Bundesstatistikgesetzes) des Gesetzes zur Förderung der elektronischen Verwaltung sowie zur Änderung weiteren Vorschriften vom 25. Juli 2013 (BGBl. I S. 2749).

5 Zu nennen ist in diesem Zusammenhang insbesondere das Portal des Bundes https://www.govdata.de/. Landesspezifische offene Daten für Berlin finden sich zudem unter http://daten.berlin.de/.

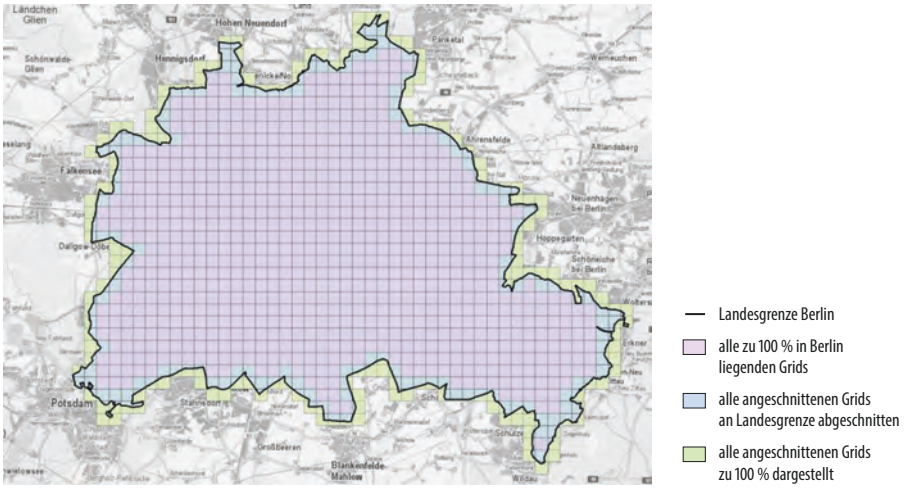
ressierte Bürgerinnen und Bürger. Die Daten der Anwendungen müssen dabei sowohl maschinenlesbar downloadbar, in entsprechenden Open-Data-Portalen auffindbar und als Dienst – etwa als wms- oder wfs-Dienst – zur Verwendung in eigenen GIS- und Kartenanwendungen verfügbar sein.[6] Durch den somit vereinfachten Zugang zu statistischen Informationen kann in der amtlichen Statistik mit geringerer Nachfrage nach Standardveröffentlichungen bei gleichzeitig wachsender Anzahl komplexer Auswertungen gerechnet werden.

Dieser Argumentation folgend, stehen neben der Applikation für die dargestellten zehn Indikatoren des Zensusatlas auch Downloadtabellen mit den dargestellten klassifizierten Werten sowie mit den separat geheim gehaltenen spitzen Werten[7], als auch die Möglichkeit der Einbindung der Applikation als wms-Dienst zur Verfügung. Somit können die visualisierten Daten unter den genannten Lizenzbedingungen mittels GIS-Systemen verarbeitet und in weitere Anwendungen integriert werden.[8]

Durch ein zunehmendes Angebot an Daten für INSPIRE-konforme Gitterzellen ist zudem zu erwarten, dass auch in der Bevölkerung die Akzeptanz und damit die Les- und Interpretierbarkeit von Rasterkarten erhöht werden kann. Der Zensusatlas soll dies entscheidend unterstützen. Auch nicht statistikaffinen Personen wird hiermit ein leicht zu bedienendes und somit niedrigschwelliges, interaktives Datenangebot unterbreitet.

Tobias Kirchner, Diplom-Geograph, ist seit 2011 im Referat *Zensus* des Amtes für Statistik Berlin-Brandenburg tätig, aktuell im Bereich Gebäude- und Wohnungszählung mit den Schwerpunkten Aufbereitung und Auswertung raumbezogener Daten sowie thematische Kartographie. Vorher war er Mitarbeiter bei der Gesellschaft für Markt- und Absatzforschung, Ludwigsburg.

h | Zuweisungen von 1 x 1 km-Gitterzellen zum Bundesland Berlin



6 Richtlinie 2007/2/EG des Europäischen Parlaments und des Rates vom 14. März 2007 zur Schaffung einer Geodateninfrastruktur in der Europäischen Gemeinschaft (INSPIRE), (ABl. L 108 vom 25.4.2007, S. 1).

7 https://www.zensus2011.de/SharedDocs/Aktuelles/Ergebnisse/DemografischeGrunddaten.html?nn=3065474

8 Zu Vor- und Nachteilen von Gitterzellendarstellungen siehe auch Zeitschrift für amtliche Statistik Berlin Brandenburg, 8. Jahrgang, Ausgabe 3/2014, S. 28ff.

Geheimhaltung

Das Geheimhaltungsverfahren SAFE

von **Jörg Höhne**

Der vorliegende Beitrag ist eine Methodenbeschreibung des Anonymisierungsverfahrens SAFE. Mit dem Verfahren SAFE kann ein anonymer Datenbestand erzeugt werden. Das kann einerseits mit dem Ziel erfolgen, einen anonymisierten Einzeldatenbestand über die Forschungsdatenzentren herauszugeben, beispielsweise als sogenanntes Scientific-Use-File. Andererseits kann SAFE als pre-tabulares Geheimhaltungsverfahren eingesetzt werden. Anstatt nach der Tabellenerzeugung alle Angaben in Tabellenfeldern zu prüfen und einzelne zu sperren, werden bei pre-tabularen Geheimhaltungsverfahren alle Tabellen aus dem anonymen Datenbestand berechnet und schützen so die Einzelangaben der Befragten. Der Beitrag beschreibt den mathematischen Hintergrund und die Lösungsalgorithmen.

1. Einleitung

Die Statistischen Ämter des Bundes und der Länder erheben Daten aus allen Bereichen des gesellschaftlichen Lebens – für über 250 Bundes- und Landesstatistiken, aber auch für Wahlen und Volksabstimmungen. Sie bereiten diese Daten auf und werten sie aus. Die Einzelangaben von Personen, wirtschaftlichen Einheiten und anderen Merkmalsträgern werden von der amtlichen Statistik geschützt und bleiben den Nutzerinnen und Nutzern der Statistiken daher verborgen. Das soll einen Missbrauch der Einzelangaben verhindern und so die Bereitschaft zur wahrheitsgemäßen Auskunft erhalten. In den Statistikgesetzen, z. B. § 16 des Bundesstatistikgesetzes[1], ist dieses Vorgehen gesetzlich geregelt.

Traditionell gewährleisten Geheimhaltungsverfahren den Schutz der Einzelangaben durch Informationsreduktion. Um die Geheimhaltung zu realisieren, werden in statistischen Ergebnissen (post-tabular) die Einzelangaben einerseits nach bestimmten Kriterien zusammengefasst (z. B. Wirtschaftszweige, Betriebsgrößenklassen, Altersgruppen, Regionen) und andererseits Angaben in noch vorhandenen sensiblen Tabellenfeldern unterdrückt oder durch weitere Vergröberung von Gliederungen unsichtbar gemacht. Außerdem besteht die Möglichkeit, durch Vergröberung der Angaben (Rundung) den Nutzen der Tabellenfelder zur Aufdeckung von Einzelangaben bei einem Missbrauch zu reduzieren.

Ein anderer Weg der statistischen Geheimhaltung besteht darin, sicherzustellen, dass bereits die Einzeldaten nicht mehr ihren Merkmalsträgern zugeordnet werden können. Die Geheimhaltungsverfahren werden dabei bereits vor der Auswertung/Tabellierung angewandt (pre-tabulare Verfahren). Das erfolgt beispielsweise durch das Entfernen von Informationen, die für die Reidentifikation besonders kritisch sind oder durch das gezielte Verändern einzelner Merkmale. Das Verfahren SAFE ist ein pre-tabulares Verfahren, bei dem eine anonyme Version des Datenkörpers über Datenveränderung der Einzelangaben erstellt wird. Aus diesem Datenkörper können dann alle potenziellen Auswertungen erstellt werden. In keiner Auswertung tritt dann mehr ein Geheimhaltungsfall auf. Da alle Auswertungen aus derselben anonymen Quelle erfolgen, sind sie außerdem untereinander konsistent.

a | Klassifikation der Geheimhaltungsverfahren

	Informationsreduzierende Verfahren	Datenverändernde Verfahren
Pre-tabulare Verfahren	• Vergröberung (Zusammenfassen von Kategorien) • Entfernen von Merkmalen	• Mikroaggregation, z. B. SAFE • Swapping • Stochastische Überlagerung auf Mikrodatenebene
Post-tabulare Verfahren	• Zellsperrung • Zusammenfassung	• Deterministische (konventionelle) Rundung • Zufällige Rundung • Kontrollierte Rundung • Stochastische Überlagerung auf Tabellenfeldebene
	↓ Löschen oder unterdrücken Information (auch unkritische Felder bei Sekundärsperrungen)	↓ Schutz entsteht durch Unsicherheit (auch bei unkritischen Feldern)

1 Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 13 des Gesetzes vom 25. Juli 2013 (BGBl. I S. 2 749).

Jedes Geheimhaltungsverfahren wird daran gemessen, dass es einen ausreichenden Schutz der Einzelangaben bietet und dabei mit einem möglichst geringen Informationsverlust einhergeht. Dass es bei Verfahren der Zellsperrung einen Informationsverlust gibt, ist für die Nutzerinnen und Nutzer von Tabellen leicht ersichtlich, da unterdrückte Felder keine Information mehr enthalten. Der Schutz der Einzelangaben durch datenverändernde Verfahren beruht auf der Änderung von Werten, die die Verwertbarkeit beim Missbrauch einschränkt. Die Datenveränderungen sind aber auch in Auswertungen enthalten und stellen dort ebenfalls einen Informationsverlust dar. Der Umfang dieses Informationsverlustes kann durch Qualitätsmaße angegeben werden. Er wird im Beitrag an mehreren Stellen angesprochen.

SAFE ist ein Verfahren der Mikroaggregation. Bei Mikroaggregationsverfahren werden einzelne sich unterscheidende Datensätze einer Mikrodatendatei durch gezielte Auswahl und Gruppenbildung so vereinheitlicht, dass jeder Datensatz in der Basisdatei mit mindestens zwei weiteren Sätzen in der Datei identisch ist. Die hier beschriebene Version des Verfahrens ist zur Behandlung von kategorialen Merkmalen geeignet. Der vorliegende Beitrag ist die überarbeitete Version von Höhne (2003), der die aktuellen Weiterentwicklungen bei der Anonymisierung von kategorialen Merkmalen einschließt. Das in der Programmversion von 2003 enthaltene SAFE-Modul zur Anonymisierung quantitativer Merkmale ist in der aktuellen Version nicht mehr enthalten. Das liegt darin begründet, dass im Rahmen vergleichender Untersuchungen Mikroaggregationsverfahren bei stetigen Werten im direkten Vergleich zu anderen datenverändernden Verfahren nicht gleich gut überzeugten, sodass die Entwicklungsaktivitäten in diesem Bereich auf andere Verfahren konzentriert wurden (siehe z. B. Ronning et al. 2005).[2] Der Beginn der Arbeiten zum SAFE-Verfahren liegt in den frühen 1990er-Jahren (vgl. Appel et al. 1993). Der Name entstand als Akronym für Sichere Anonymisierung für Einzeldaten (SAFE).

Bei Mikroaggregationsverfahren werden die zu vereinheitlichenden Gruppen meist durch die Minimierung eines Abstandsmaßes zwischen den Einheiten gebildet.[3] Die diversen Mikroaggregationsverfahren unterscheiden sich dabei in der Wahl des Abstandsmaßes, in der Gruppengröße und in der Art, wie die Gruppen nach der Gruppenbildung vereinheitlicht werden. Das Verfahren SAFE zählt zu den Mikroaggregationsverfahren, da auch hier Gruppen vereinheitlicht werden. Allerdings werden die Gruppen durch ein numerisches Optimierungsverfahren so gebildet, dass ein Satz an vorgegebenen Auswertungstabellen möglichst exakt – jedoch ohne Geheimhaltungsfälle – aus dem anonymen Material wieder erzeugt werden kann.

Das dargestellte Verfahren erfüllt das Kriterium der k-Anonymität mit k = 3, es ist somit 3-anonym. Ein Datenbestand ist k-anonym, wenn jede Merkmalskombination mindestens als k-Tupel auftritt. Originale Beobachtungseinheiten können Datensätzen des k-anonymen Datenbestandes nicht mehr eindeutig zugeordnet werden, da mindesten k-1 Datensätze genauso wahrscheinlich zum Original passen (vgl. Sweeney 2002).

2. Begriffsbestimmungen

a) Die Mikrodatendatei

Eine Mikrodatendatei ist eine Datei, in der jedes statistische Objekt (Merkmalsträger) durch einen einzelnen Datensatz (Zeile) repräsentiert wird. Sie wird deshalb auch als Einzeldaten- oder Basisdatei bezeichnet. Sie bildet den Ausgangspunkt für alle möglichen Auswertungen des Datenbestandes. Die Merkmale in Datensätzen unterscheiden sich hinsichtlich ihres Skalenniveaus, es gibt qualitative (kategoriale) Merkmale und quantitative (stetige) Merkmale. Da die aktuelle Version von SAFE nur kategoriale Merkmale anonymisiert, sollte die Mikrodatendatei nur die kategorialen Merkmale des zu anonymisierenden Datenbestandes enthalten. Die übrigen Merkmale der Mikrodatendatei (Identifikationsmerkmale oder quantitative Merkmale) werden durch das Verfahren nicht behandelt. Für Analysezwecke können sie in der Datei enthalten bleiben, um beispielsweise die Datenänderungen im Verfahren zu quantifizieren. Außerdem kann durch das nachträgliche Anwenden von Anonymisierungsverfahren für quantitative Merkmale eine Mikrodatendatei mit anonymen qualitativen und quantitativen Merkmalen erzeugt werden.

Bei qualitativen Merkmalen, auch kategoriale Merkmale genannt, handelt es sich um Merkmale, die eine diskrete, feste Anzahl an Ausprägungen haben. Die möglichen Ausprägungen sind in einer Schlüsseltabelle zusammengefasst. Handelt es sich um hierarchische Schlüssel, so können qualitative Merkmale durch Umschlüsselung auf höhere Aggregationsebenen umgesetzt werden, beispielsweise können Regionalschlüssel, wie Gemeinde, auch in Kreis, Regierungsbezirk oder Land umgeschlüsselt werden. Analog kann der Wirtschaftszweig in Branchen oder die einzelne Nationalität in Deutsch/Nichtdeutsch umgeschlüsselt werden.

Quantitative Merkmale sind in der amtlichen Statistik, bedingt durch die Messgenauigkeit, meist ganzzahlig. Es gibt keine endliche vorher festgelegte Schlüsselmenge, die zulässig ist. Beispiele sind Umsatz und Beschäftigte. Aus quantitativen Merkmalen lassen sich durch Gruppierung wieder qualitative Klassen erzeugen (z. B. Betriebe mit unter 20 Beschäftigten, 20 bis unter 50 Beschäftigten usw.). Mit der aktuellen Version von SAFE können keine quantitativen Merkmale anonymisiert werden.

Identifikationsmerkmale (Ident-Nummern, Betriebsnummern, Adress-IDs usw.) sind Schlüsselmerkmale, die eine eindeutige Zuordnung des

2 Im Bereich der Anonymisierung von wirtschaftsstatistischen Daten lag der Schwerpunkt der methodischen Forschung des Autors und seiner Kolleginnen im Amt für Statistik Berlin-Brandenburg bei der „kontrollierten stochastischen Überlagerung".

3 Eine Ausnahme bilden die Verfahren der stochastischen Mikroaggregation (siehe Lechner und Pohlmeier 2003).

Datensatzes zum statistischen Objekt ermöglichen. Diese werden im Rahmen der Anonymisierung nicht betrachtet, da sie vor der Nutzung anonymer Daten in jedem Fall entfernt werden.

Eine Mikrodatendatei (Basisdatei) B ist eine Menge aus n Objekten. Jedes Objekt wird durch ein Tupel von k Werten beschrieben. Dabei beschreiben die Werte $b_{i1}, \ldots, b_{ik}$ die qualitativen (kategorialen) Merkmale des i-ten Objektes.

B – Basisdatei mit $B = \{b_{ij}\}$ - $i = 1,2,\ldots,n$
- $j = 1,2,\ldots,k$

n – Anzahl der statistischen Objekte
k – Anzahl der qualitativen/kategorialen Merkmale
b_{ij} – Wert des Merkmals j beim Objekt i

b) Die Kontrolltabellen

Das Verfahren SAFE optimiert bei der Anonymisierung die Lösung an einem vorgegebenen Kanon an Auswertungstabellen. Auswertungstabellen sind dabei Häufigkeitstabellen, die dadurch gebildet werden, dass die Basisdatei über eine bestimmte Merkmalskombination aggregiert wird. Es erfolgt eine Gruppenbildung (Aufsummierung von Sätzen mit gleicher Ausprägungskombination). Diese geplanten Auswertungen werden Kontrolltabellen genannt, da die Qualität der Anonymisierung am möglichst guten Erhalt dieser Auswertungen kontrolliert wird.

Die Kontrolltabellen können einerseits automatisch erzeugt werden (alle möglichen Tabellen bis zur Dimension d). Dabei werden alle Tabellen aus Kreuzkombinationen der qualitativen Merkmale mit beliebigen Schlüsselstufen bis zu einer festgelegten Tabellendimension (Anzahl zusammen tabellierter Merkmale) verstanden. Bei einer Festlegung auf die maximale Dimension $d = 3$ werden beispielweise alle möglichen eindimensionalen Häufigkeitstabellen, alle zweidimensionalen Tabellierungen und alle möglichen Kreuztabellierungen aus drei qualitativen Merkmalen als Kontrolltabellen erstellt.

Alternativ zur automatischen Generierung besteht andererseits die Möglichkeit, Kontrolltabellen per Liste dem Programm zuzusteuern. Beide Varianten können auch kombiniert werden. Dadurch besteht auch die Möglichkeit, ausgewählte höherdimensionale (z. B. vier- und fünfdimensionale) Tabellen zusätzlich zu allen ein- bis dreidimensionalen Tabellen zu kontrollieren. Da aufgrund der hohen Anzahl der theoretisch möglichen höherdimensionalen Tabellen es meist nicht möglich ist, z. B. alle vier- und fünfdimensionalen Tabellen als Kontrolltabellen zu behandeln, ist diese Mischung der Kontrolltabellenfestlegung erforderlich, wenn auch einzelne hochdimensionale Tabellen später ausgewertet werden sollen.

3. Enthüllungsrisiken und der Lösungsansatz von SAFE

a) Enthüllungsrisiken in Tabellen

Aus § 16 BstatG ergibt sich die Verpflichtung, „Einzelangaben über persönliche und sachliche Verhältnisse […] geheimzuhalten". Jede Tabelle muss von der amtlichen Statistik daher vor der Veröffentlichung dahin gehend geprüft werden, dass kein Tabellenfeld (auch Tabellenzelle genannt) dazu geeignet ist, auf Einzelangaben zurück zu schließen. Von einem Enthüllungsrisiko oder einem Geheimhaltungsproblem spricht man, wenn aus einem Tabellenfeld Rückschlüsse auf ein einzelnes statistisches Objekt (Unternehmen, Bürger usw.) gezogen werden könnten und so Informationen über das statistische Objekt nur aufgrund der statistischen Veröffentlichung zugänglich werden.

Folgende Enthüllungsrisiken können bei der Veröffentlichung von Tabellen entstehen (vgl. Hundepool et al. 2012):

- Fallzahlprobleme
  Enthüllungsrisiken durch zu kleine Fallzahlen treten vor allem bei Wertetabellen auf. Wenn nur ein oder zwei Merkmalsträger zu einem Tabellenwert, beispielsweise einer Summe, beitragen, besteht das Risiko einer exakten Enthüllung von Einzelwerten. Trägt nur ein Merkmalsträger zu einem Tabellenfeld bei, so entspricht der Tabellenwert der Einzelangabe des Merkmalsträgers. Bei zwei statistischen Objekten besteht das Risiko darin, dass eines der beiden Objekte durch Differenzbildung die Information über das andere Objekt (aufgrund der gleichen Merkmale in der Regel der Konkurrent) problemlos generieren kann.
  Bei Häufigkeitstabellen zeigen kleine Fallzahlen seltene oder einzigartige Merkmalskombinationen an. Direkte Rückschlüsse auf Einzelangaben sind im Allgemeinen nicht möglich. Die Information, dass die Merkmalskombination selten oder einzigartig ist, kann eventuell trotzdem als problematisch eingestuft werden:
  – Die Identifikation könnte mithilfe von Zusatzwissen möglich sein.
  – Durch die Verknüpfung mit Informationen aus anderen Datenquellen/anderen Tabellen wird es möglich, direkt Rückschlüsse auf andere Merkmale der betreffenden Individuen zu ziehen.
- Randwerte/Randsummenprobleme
  Randsummenprobleme entstehen, wenn innerhalb einer Tabelle in einer Zeile oder Spalte nur eine Zelle belegt ist. In diesem Fall können, auch wenn mehr als zwei Objekte zur konkreten Ausprägung des Tabellenwertes beitragen, Attribute für alle Beitragenden enthüllt werden. Ein Beispiel in der Todesursachenstatistik für einen Randwert und ein daraus entstehendes Randsummenproblem ist das Folgende: Innerhalb einer Region und Altersgruppe sterben alle Personen an der gleichen Krankheit. Allein die Information über das Alter und die Region einer gestorbenen Person ermöglicht es dann, die Todesursache anhand der Statistik eindeutig zuzuordnen. Randsummenprobleme sind immer inhaltlich zu betrachten, d. h. es ist zu entscheiden, ob das Merkmal geheimhaltungskritisch ist. Es ist

offensichtlich kein Geheimhaltungsproblem, wenn innerhalb der Gruppe der unter 6-Jährigen einer Region alle Kinder nicht erwerbstätig sind.

Neben den Problemen aus der Darstellung bei Tabellen existieren bei der Herausgabe von Mikrodaten weitere Reidentifikationsprobleme, die sich aus sogenannten „Matching"-Versuchen ergeben. Bei diesen wird versucht, Informationen, die man über statistische Objekte aus externen Quellen gewonnen hat, gegen Sätze der anonymen Basisdatei anzuspielen und so bei einer eineindeutigen Übereinstimmung zusätzliche Eigenschaften aus der Basisdatei abzulesen. Es wird üblicherweise unterschieden zwischen:

- Einzelangriffen – Datenangriffe, bei denen versucht wird, durch ein Matching für einzelne Objekte (z. B. Unternehmen) Informationen zu erhalten – und
- Massenfischzügen – hier wird ein Datenbestand an einen anderen gematcht und so möglichst viele Sätze zugeordnet (vgl. Lenz 2010, Ronning et al. 2005, Höhne 2010).

Allen vier Deanonymisierungsrisiken wird im Rahmen des SAFE-Verfahrens Rechnung getragen.

b) Lösungsansatz von SAFE

Das Verfahren SAFE erzeugt einen anonymen Datenbestand, der das Kriterium der k-Anonymität (vgl. Sweeney 2002) mit $k = 3$ erfüllt. Jede Ausprägungskombination tritt mindestens als Dreier-Tupel auf. Damit ergeben sich für die einzelnen Deanonymisierungsrisiken folgende Sicherheiten:

- Fallzahlprobleme können in den Tabellen nicht mehr auftreten, da mindestens drei Sätze zu einem Tabellenwert beitragen. Das bedeutet, dass entweder die in der Realität auftretenden kritischen Merkmalskombinationen entfernt wurden oder durch die Aggregation die Häufigkeit der Ausprägungskombination auf mindestens 3 erhöht wurde.
- Matching-Algorithmen, egal ob Einzelangriffe oder Massenfischzüge, können nur zu einer mehrdeutigen Zuordnung führen. Wenn ein Satz mehrere Entsprechungen in der anonymisierten Basisdatei hat, die wiederum durch Gruppenbildung entstanden sind, so kann auch nicht daraus geschlossen werden, dass die zusätzlich ablesbaren Eigenschaften für das Original gelten. Die k-Anonymität (vgl. Sweeney 2002) schützt vor Matching, da jede Ausprägungskombination dreifach vorhanden ist.
- Randsummenprobleme können bei Auswertungstabellen entstehen, aber auch hier ist es durchaus möglich, dass diese Probleme nur das Ergebnis der Anonymisierungstechnik sind. Künstliche Randsummenprobleme werden zusätzlich erzeugt, wenn durch das Gruppieren Objekte mit qualitativ verschiedenen Eigenschaften, aber geringen Häufigkeiten aus dem Datenbestand entfernt wurden. Es ist somit kein sicherer Rückschluss auf die Eigenschaften der Basisdatei mehr möglich.

Mit der oben eingeführten Notation lässt sich das allgemeine SAFE-Geheimhaltungsproblem („allgemein" bedeutet auch für Datenbestände mit stetigen Merkmalen) folgendermaßen darstellen:

Die originale Mikrodatendatei sei die Matrix B^o. Für diese Datei lässt sich für alle vereinbarten Schlüsselstufen und Aggregationsvorschriften die Menge aller vorgegebenen Auswertungstabellen bilden.

$T^o = F(B^o)$ – Matrix der Ergebnisse aller Auswertungstabellen.

$t^o_{p,q,m} = f_{p,q,m}(B^o)$ – Für jeden Tabellenwert $t^o_{p,q,m}$ des Merkmals m für die q-te Ausprägungskombination in der Auswertungstabelle p existiert eine Berechnungsfunktion $f^o_{p,q,m}$, mit der sich der Tabellenwert aus der Matrix der Mikrodaten bestimmen lässt. Übliche Funktionen sind die Summenfunktion zur Bildung von Aggregaten, aber auch Funktionen zur Berechnung von Durchschnitten oder Anteilswerten. Nach Bestimmung aller Geheimhaltungsfälle in den Auswertungstabellen lassen sich für die Geheimhaltungsfälle eine untere Grenze ($z^u_{p,q,m}$) und eine obere Grenze ($z^o_{p,q,m}$) bestimmen, die ein Unzulässigkeitsintervall um den geheim zu haltenden Wert beschreiben. Handelt es sich bei den Auswertungstabellen um Wertetabellen, so können die Unzulässigkeitsintervalle beispielsweise in Anlehnung an die Dominanzregeln für Wertetabellen abgeleitet werden, die unterstellen, dass ein beitragender Einzelwert nicht genauer als mit einem Fehler von x % rückschließbar sein darf. Handelt es sich bei den Auswertungstabellen um Fallzahltabellen, so darf keine eindeutige Zuordnung möglich sein. Damit sind für diese Tabellen die Fallzahlen 1 und 2 unzulässig.

Für eine anonymisierte Basisdatei B^a muss gelten:

$$T^a = F(B^a) \quad (1)$$

mit

$$t^a_{p,q,m} = f_{p,q,m}(B^a)$$

und

$$z^u_{p,q,m} \geq f_{p,q,m}(B^a) \quad \vee \quad z^o_{p,q,m} \leq f_{p,q,m}(B^a)$$

mit:

T^a – Matrix der Ergebnisse aller anonymen Auswertungstabellen.

B^a – Matrix der anonymisierten Basisdatei. Die anonymisierte Basisdatei ist dadurch gekennzeichnet, dass jede Zeile mindestens dreimal identisch in der Matrix enthalten ist.

$t^a_{p,q,m} = f_{p,q,m}(B^a)$ – In der Auswertungstabelle p wird für die Auswertung der Ausprägungskombination q beim Merkmal m der anonyme Tabellenwert $t^a_{p,q,m}$ bei Auswertung der anonymen Basisdatei über diese Funktion ermittelt.

$z^u_{p,q,m}, z^o_{p,q,m}$ – untere und obere Schranke des Unzulässigkeitsintervalls im Tabellenfeld. Wenn ein Geheimhaltungsfall beim Merkmal m in der Ausprägungskombination q der Tabelle p existiert, dann beschreiben diese Schranken Grenzen, ab denen der veröffentlichte Wert nicht mehr für einen Datenmissbrauch als nutzbar angesehen werden kann. Wenn kein Geheimhaltungsfall in diesem Tabellenfeld existiert, gilt:
$z^u_{p,q,m} = z^o_{p,q,m} = t^o_{p,q,m}$,
sodass der Unzulässigkeitsbereich leer ist.

Gesucht ist eine anonyme Basisdatei, deren Auswertungstabellen den originalen möglichst ähnlich sind, d. h. der Abstand zwischen T^o und T^a sollte minimal sein. Die Ausgestaltung der Funktion zur Messung des Abstandes zwischen T^o und T^a hängt dabei von der konkreten Ausgestaltung des Begriffes der „Tabellenqualität" ab, die sich an dem Bedarf der Datennutzerinnen und -nutzer orientieren sollte.

Für den im Folgenden näher untersuchten Fall der Anonymisierung von ausschließlich kategorialen Merkmalen und damit auch ausschließlich Häufigkeitstabellen als zu kontrollierende Tabellen lässt sich das obige allgemeine SAFE-Problem vereinfachen. Da für jeden einzelnen Datensatz der Mikrodaten in Häufigkeitstabellen nur die Möglichkeit existiert, dass er in einem Tabellenfeld mitgezählt wird oder nicht, lässt sich eine Zuordnungsmatrix A (nur aus 0 und 1 Elementen) bilden. Der Zusammenhang zwischen der Häufigkeit der Sätze in einer Mikrodatendatei und dem Ergebnis der Auswertung in kontrollierten Tabellierungen ist dann:

$$T = AX$$

mit:

X – ist der Häufigkeitsvektor, der angibt, wie oft Objekte mit diesen Merkmalsausprägungen im Datenbestand vorhanden sind. Üblicherweise gilt bei originalen Mikrodaten $x_j = 1$. Werden identische Datensätze bereits vorher zusammengefasst, gilt $x_j > 1$ ($\sum x_j$ = Anzahl der Objekte).

T – Vektor aller Ergebnisse in Auswertungstabellen, also der Vektor aller Tabellenfelder von zu kontrollierenden Häufigkeitstabellen t_i; i = 1, 2, …, k (k-Anzahl der Häufigkeitsfelder in allen zu kontrollierenden Tabellen). Der Vektor hat eine Blockstruktur, wobei jeder Block die möglichen Tabellenfelder genau einer Häufigkeitstabelle der τ Häufigkeitstabellen enthält.

$$T = \begin{Bmatrix} T_1 \\ T_2 \\ \hline T_\tau \end{Bmatrix}$$

A – Zuordnungsmatrix mit $a_{ij} = 1$, wenn das Objekt j im Tabellenfeld i tabelliert wird, sonst $a_{ij} = 0$.

$$A = \begin{Bmatrix} A_1 \\ A_2 \\ \hline A_\tau \end{Bmatrix}$$

mit

$$A_1 = \left\{ \begin{array}{ccc} a_{1i1} = 1 & a_{1ij} = 0 & a_{1in} = 0 \\ a_{1i1} = 0 & a_{1ij} = 1 & a_{1in} = 0 \\ \hline a_{1(m-1)1} = 0 & a_{1(m-1)j} = 0 & a_{1(m-1)n} = 0 \\ a_{1m1} = 0 & a_{1mj} = 0 & a_{1mn} = 1 \end{array} \right\}$$

Wegen der Blockstruktur im Vektor der Tabellenfelder, die durch die Aneinanderreihung der einzelnen Auswertungstabellen entsteht, gilt auch eine Blockstruktur für A. Die Höhe der Blöcke innerhalb A ist identisch mit der in T und in jedem Block bilden die Spalten Einheitsvektoren, da jedes Objekt nur in genau einem Tabellenfeld einer Auswertungstabelle gezählt wird.

Die in der obigen allgemeinen Schreibweise (1) formulierten Ausschlussintervalle von nicht zulässigen Tabellierungswerten und die Existenz von mindestens drei identischen Einheiten können bei ausschließlich Häufigkeitstabellen einfach durch die folgende Bedingung abgebildet werden:

$$x^a_j = 0, 3, 4, \ldots; \text{ für alle j.}$$

Damit erfüllt dieser Häufigkeitsvektor x^a die Kriterien der k-Anonymität

(d. h. x^a_j ganzzahlig und $x^a_j \neq 1, 2$).

4. Mathematisches Modell und Optimierungsaufgabe

Der verfolgte Ansatz geht von der Bestimmung einer „optimalen Teilmenge" aus dem bereitgestellten Mikrodatenbestand aus. Zur Erhöhung der Schutzwirkung des Verfahrens kann dieser Mikrodatenbestand auch um fiktive Sätze erweitert werden.[4] Für diese Sätze wird die Häufigkeit 0 im originalen Häufigkeitsvektor hinterlegt.

Nach Festlegung der zu kontrollierenden Tabellen lassen sich für den originalen Datenbestand die Vektoren X und T sowie die Zuordnungsmatrix A bestimmen.

Beschreiben:

X^o – Vektor der originalen Häufigkeiten x_i^o der statistischen Objekte der Ausprägungskombination i; i = 1, 2, …, n im Originalbestand (n-Anzahl der Zeilen der Mikrodaten).

T – Vektor aller originalen Tabellenfelder (Häufigkeiten der Objekte) über alle zu kontrollierenden Tabellen t_j; j = 1, 2, …, k (k-Anzahl der Häufigkeitsfelder über alle zu kontrollierenden Randsummentabellen).

A – Zuordnungsmatrix – Blockmatrix mit Einheitsvektoren in den einzelnen zu kontrollierenden Blöcken A_t; t = 1, 2, …, τ (τ-Anzahl der zu kontrollierenden Häufigkeitstabellen). Wenn die Zeile der Mikrodaten j die Ausprägungen so besitzt, das diese Zeile im Tabellenfeld i gezählt wird, gilt $a_{tij} = 1$ sonst $a_{tij} = 0$.

Dann lässt sich der Zusammenhang zwischen den Mikrodaten zu den originalen Tabellenfeldern darstellen als:

$$T^o = AX^o.$$

Bei einer anonymen Datei muss gelten $x_i^a \in \{0, 3, 4, 5, \ldots\}$. Alle vorhandenen Objekte haben eine Häufigkeit von mindestens 3. Die Häufigkeit 0 bewirkt, dass diese Objekte in der anonymen Lösung nicht mehr vorhanden sind.

4 Die Notwendigkeit sollte auf der Grundlage der Gesamtanzahl an Einheiten und der geplanten Auswertungstiefe entschieden werden. Mit diesem Vorgehen wird verhindert, dass man darauf zurückschließen kann, dass jeder Satz der Mikrodaten auch real existieren müsste. Wäre das problematisch, werden die Daten um plausible, aber nicht vorkommende Merkmalskombinationen erweitert.

Für sehr tief gegliederte Auswertungen lässt sich leicht ein Datenbeispiel finden, für das keine Lösung als Gleichungssystem existiert.

Beispiel:

In einer feinen regionalen Gliederung existieren drei Einheiten (z. B. drei Betriebe in einer kleinen Gemeinde).

WZ	Betriebe original	Betriebe anonym	Abweichung
D.10	3	3	±0
D.10.1	1	0	–1
D.10.2	1	3	2
D.10.3	1	0	–1

Diese haben einerseits verschiedene Ausprägungen in den Wirtschaftszweiggruppen, müssen aber auch zusammen als eine Wirtschaftsabteilung (D.10) in Tabellen präsentiert werden können.

Unabhängig davon, wie man die anonyme Lösung gestalten würde, wäre eine absolute Abweichung <2 für alle Tabellenfelder nicht erreichbar. Es lassen sich auch weitere analoge Beispiele zeigen, bei denen eine Abweichung <2 für die Lösung bei mindestens zwei hierarchischen Auswertungstabellen nicht möglich ist. Das obige Gleichungssystem hat somit unter bestimmten Konstellationen keine Lösung $AX^a = T^o$ unter der geforderten Nebenbedingung $x_i^a \in \{0, 3, 4, 5, \ldots\}$. Deshalb ist ein Fehlervektor F (f_j – Fehler im Tabellenfeld j; j = 1, 2, …, k) einzufügen, der die Abweichungen zwischen der Tabellierung der Originaldaten und der anonymen Lösung beschreibt.

Die Menge aller möglichen anonymen Lösungen beschreibt sich dann als

$$\begin{aligned} AX^a + F &= T^o \\ \sum_{i=1}^{n} x_i^a &= O \\ x_i^a &\in \{0,3,4,5,\ldots\} \\ i &= 1,2,\ldots,n \\ j &= 1,2,\ldots,k \end{aligned}$$

mit:

F – Vektor der Tabellierungsfehler, f_j ist die Abweichung des anonymen Ergebnisses zum Original bei der Tabellierung des Tabellenfeldes j,

O – als Gesamtanzahl der statistischen Objekte.

Für die Bestimmung einer eindeutigen Lösung ist zusätzlich eine Zielfunktion Z einzuführen. Die Definition der Funktion orientiert sich an mehreren Zielen:

1. Die Funktion sollte möglichst transparent für die späteren Datennutzerinnen und -nutzer sein. Das Funktionsergebnis sollte für die Interpretation der Datenqualität gut anwendbar sein.
2. Die Funktion sollte innerhalb des Lösungsalgorithmus gut handhabbar sein.
3. Die Funktion sollte sowohl für die in den Tabellen nebeneinander auftretenden großen als auch kleinen Häufigkeiten sinnvolle Optimierungsziele vorgeben.

Vor diesem Hintergrund ist der maximale relative Fehler unbrauchbar, da bei Fallzahlproblemen (Unikaten im Datenbestand, die in einzelnen Tabellenfeldern allein dargestellt werden) ein relativer Fehler von –100 % bzw. +200 % unumgänglich ist. Dieser relative Fehler würde bereits beim Ändern eines geheim zu haltenden Tabellenfeldes von 1 (Unikat) zu 0 bzw. 3 entstehen. Dieser dort mindestens notwendige relative Fehler würde aber zu unbrauchbaren Ergebnissen führen, wenn man ihn für alle Tabellenfelder akzeptieren würde. Eine minimierte Summe der absoluten Abweichungen oder die Summe der Quadrate der Abweichungen erwiesen sich ebenfalls als ungünstig, da bei Testrechnungen einzelne sehr starke Ausreißer nicht verhindert werden konnten. Da für die Datennutzerinnen und -nutzer die Bewertung der Qualität der anonymen Daten für genau eine, ihre jetzt aktuell interessierende Datenabfrage relevant ist, ist die Aussage der mittleren Abweichung oder der mittleren quadratischen Abweichung nur schwer vermittelbar, wenn keine sicheren maximalen Schranken zusätzlich existieren. Der Maximalfehler in den Randsummentabellen erwies sich deshalb als brauchbares Kriterium für die Bestimmung der Optimalität. Dieser zulässige Maximalfehler kann dabei in Abhängigkeit von der Größe des Tabellenfeldes nochmals gestaffelt werden. Mögliche Varianten zur Bestimmung des Vektors g werden in Abschnitt 6 dargestellt.

Es ist somit für die verschiedenen möglichen Lösungsvektoren X die Lösung mit dem kleinsten Maximalfehler gesucht.

$$\begin{aligned} Z &= \min_{X} \left(\max_{j} (|\, f_j \,| - g_j) \right) \qquad (2) \\ AX + F &= T \\ \sum_{i=1}^{n} x_i &= O \\ x_i &\in \{0,3,4,5,\ldots\} \\ i &= 1,2,\ldots,n \\ j &= 1,2,\ldots,k \end{aligned}$$

mit:

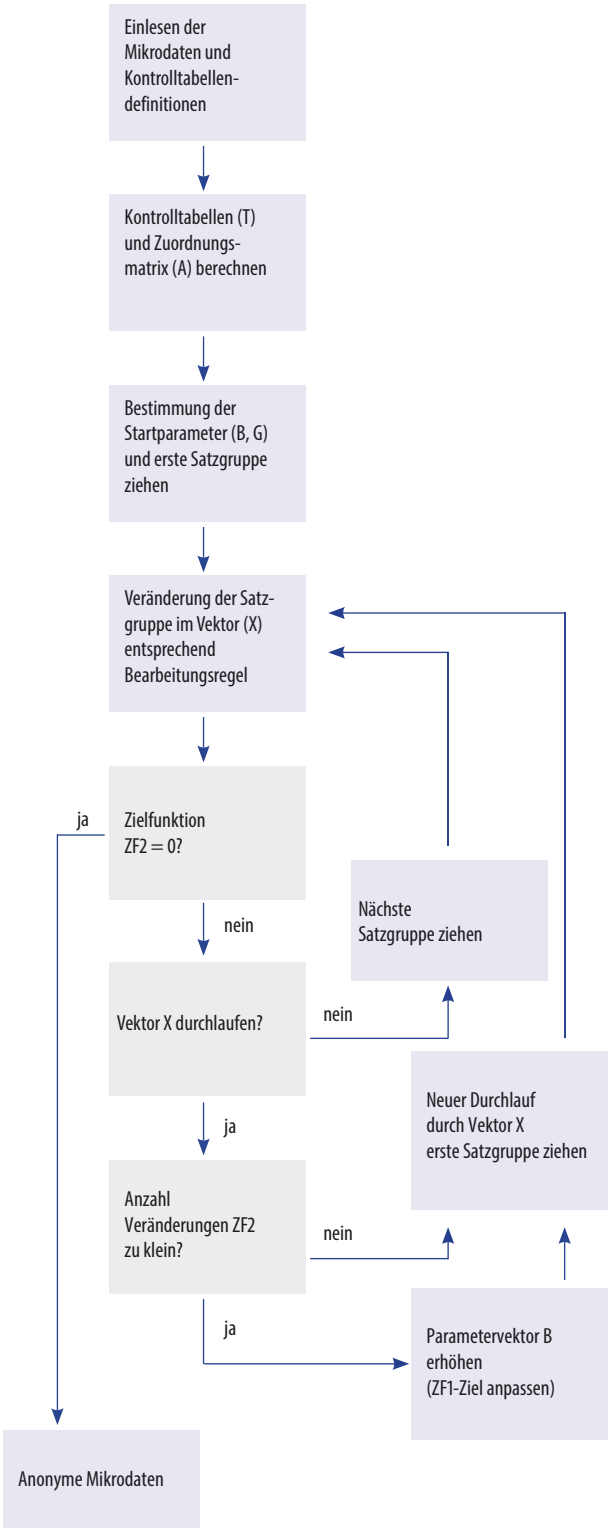
g_j – zusätzlich zulässige Abweichung im Tabellenfeld j, diese zusätzliche Abweichung wird in Abhängigkeit von der Größe des Tabellenfeldes bestimmt, $g_j = G(t_j)$.

Die Aufgabe besteht darin, eine Lösung zu finden, in der keine Merkmalskombinationen mehr auftreten, die bei Auswertungen Geheimhaltungsfälle erzeugen könnten. Deshalb müssen die unerwünschten kleinen Fallzahlen von 1 oder 2 geändert werden (z. B. auf eine Häufigkeit von 3 oder größer bzw. auf die Häufigkeit 0). Die Bedingungen „Erhalt der Gesamtanzahl der Objekte“ und „Minimierung der maximalen Abweichung der Tabellenwerte zwischen Original und Anonymisiert“ sind dabei zu berücksichtigen. Dieses Modell ist somit eine Optimierungsaufgabe mit n+k Unbekannten (Häufigkeiten und Tabellierungsfehler) und k+n+1 Nebenbedingungen (Anzahl Tabellengleichungen, die Mengeneinschränkungen für x_i und die Gesamtanzahl der Objekte). Aufgrund der n Nebenbedingungen zu x_i ist die Aufgabe nichtlinear und ganzzahlig, es können auch mehrere Lösungen unter diesen Bedingungen existieren.

Für reale Datenbestände hat diese Aufgabe eine Dimension, die mit heutiger Rechentechnik nicht explizit lösbar ist. Selbst für ein relativ kleines Beispiel

– 100 000 statistische Objekte mit sieben Merkmalen mit zusammen 16 verschiedenen Schlüsselstufen und 149 817 zu kontrollierenden Tabellenfeldern – ergibt sich ein Speicherbedarf von $(n+k+1)^2$ 4 Byte = 233 GB. Dieser müsste für einen schnellen Zugriff als Hauptspeicher verfügbar sein. Deshalb wurde als Alternative ein performanter numerischer Algorithmus gewählt.

b | Programmablauf „Finden der zulässigen Lösung“



5. Lösungsweg/Numerischer Algorithmus

Numerische Algorithmen zeichnen sich durch eine schrittweise Annäherung an das Optimierungsziel aus. Es wird dabei ein Verfahren festgelegt, das von einer bekannten schlechten Lösung zu einer besseren führt. Die bessere Lösung ist dadurch gekennzeichnet, dass sie näher an der gesuchten Lösung liegt. Das wiederholte Anwenden der Verfahrensregeln führt dann zum Auffinden der gesuchten Lösung.

a) Bestimmung der Startlösung

Da zu Beginn keine anonyme Startlösung bekannt ist, muss diese als erstes bestimmt werden. Dazu gibt es zwei Möglichkeiten:

1. Entweder man bestimmt eine anonyme Startlösung durch einen einfachen Algorithmus. Hier könnte man beispielsweise jeden dritten Satz der Mikrodatendatei mit der Häufigkeit 3 verwenden oder auch andere Algorithmen wählen.
2. Alternativ beschreibt man die Bestimmung der Startlösung als eine separate Optimierungsaufgabe. Dazu „erweitert man die Lösungsmenge“, d. h. es werden auch nicht vollständig anonymisierte Lösungen (für das Ziel eigentlich nicht zulässige Lösungen) in die Lösungsmenge mit aufgenommen. Damit sind die Originalhäufigkeiten des Datenbestandes bereits eine Startlösung. Der Algorithmus muss im ersten Schritt jedoch die Minimierung der Anzahl der noch vorhandenen Geheimhaltungsfälle als primäres Ziel mit in die Zielkriterien aufnehmen, damit der Iterationsalgorithmus das Auffinden der zulässigen Lösung ermöglicht. Der numerische Algorithmus teilt sich dann in die zwei Schritte „Finden der zulässigen Lösung“ und „Optimierung der Lösung“.

Von dieser gefundenen Lösung ausgehend muss dann bei den weiteren Schritten iterativ eine Verbesserung der Qualität der Tabellenfelder erfolgen, wobei jede Veränderung dann so erfolgt, dass keine Geheimhaltungsfälle ($x_j = 1$ oder $x_j = 2$) mehr zugelassen werden.

Nach verschiedenen Tests wurde der zweite Weg verwendet, weil so der Vorteil besteht, dass der Tabellierungsfehler in der gefundenen Startlösung bereits sehr klein ist. Bei Variante 1 sind die entstehenden Tabellierungsfehler zufällig normalverteilt, d. h. es existieren einzelne starke Ausreißer (sehr große Tabellierungsfehler). Bei Variante 2 kann bereits das angestrebte Ziel einer minimalen Maximalabweichung der Tabellierungsfehler als untergeordnetes Zielkriterium mit aufgenommen werden. Die Lösung der Aufgabe wurde damit in zwei Optimierungsaufgaben „Finden der zulässigen Lösung“ und „Optimierung der Lösung“ geteilt, die in zwei Programmen umgesetzt sind.

b) Zielfunktion/Entscheidungsregel/Algorithmus

Der Ablaufplan zur Umsetzung des Algorithmus „Finden der zulässigen Lösung“ ist in Abbildung b dargestellt. Die Zielkriterien für das Verfahren zum „Finden der zulässigen Lösung“ sind die Nachfolgenden, wobei die Nummerierung der Anwendungsreihenfolge entspricht:

1. Minimierung des maximalen Fehlers in den Tabellierungen,
2. Minimierung der Anzahl der verbleibenden Geheimhaltungsfälle in der Datei,
3. Maximierung der Möglichkeiten für weitere Veränderungen,
4. Minimierung des mittleren Fehlers in allen kontrollierten Tabellierungen.

Zwischen dem 1. und dem 2. Ziel besteht ein Widerspruch. Da Originaldaten natürlich einen Tabellierungsfehler von 0 in allen Auswertungen haben, wird ein Verändern der Daten, um Geheimhaltungsfälle zu verringern, zu neuen Tabellierungsfehlern führen. Auch danach ist es nicht immer auszuschließen, dass neue Tabellierungsfehler auftreten. Um diesen Widerspruch aufzulösen, wird das erste Zielkriterium durch die Vorgabe einer fixen Fehlerschranke verfolgt. Diese eröffnet einen Entscheidungsspielraum, der eine Veränderung der Daten zur Verbesserung des zweiten Zielkriteriums erlaubt. Gleichzeitig wird diese Fehlerschranke nur bei unbedingtem Bedarf angepasst. Beginnend mit einem festen Startfehler (in der Regel ±2) wird die Realisierung der Teilziele 2 bis 4 angestrebt. Die Fehlerschranke (bound) von ±2 ist die kleinste mögliche Fehlerschranke (siehe auch Beispiel im Abschnitt 4). Wird mit diesem Startwert für die Fehlerschranke keine Lösung gefunden, so wird sie um 1 erhöht und die Suche fortgesetzt. Wenn Erfahrungen durch die Lösung gleichartiger Beispiele vorliegen, können auch größere Schranken als Startwert vorgegeben werden.

Beim schrittweisen Durchlaufen der Datei werden alle diejenigen potenziellen Veränderungen durchgeführt, die die Ziele 2 bis 4 verbessern. Das bedeutet, es wird folgende Optimierungsaufgabe „Finden der zulässigen Lösung" behandelt:

$$
\begin{aligned}
&ZF1: \quad \min(\max(b)) \qquad (3)\\
&ZF2: \quad \min\left(\sum_{i=1}^{n} c_i\right)\\
&ZF3: \quad \min\left(\sum_{i=1}^{k} S\left(\left|t_i^z - t_i^o\right|\right)\right)\\
&ZF4: \quad \min\left(\sum_{i=1}^{k} \left|t_i^z - t_i^o\right|\right)\\
&\left|T^z - T^o\right| \le G + B\\
&T^z = AX^z
\end{aligned}
$$

mit:

$$
\begin{aligned}
&x_j^z = 0,1,2,3,4,\ldots\\
&c_j = 1 \quad ; \text{gdw.} \quad x_j \in \{1,2\}\\
&c_j = 0 \quad ; \text{gdw.} \quad x_j \notin \{1,2\}
\end{aligned}
$$

mit:

X^z – Vektor der Häufigkeiten möglicher Sätze in den Zwischenschritten im Datenbestand
- als Startlösung $X^z = X^o$, wobei
- einmalige Originalsätze ($x_j^o = 1$)
- völlig identische Originalsätze zusammengefasst ($x_j^o > 1$)
- ggf. künstliche Sätze ($x_j^o = 0$)

T^o – Vektor der Häufigkeiten aller Tabellenfelder in den zu kontrollierenden Tabellen bei Auswertung des Originaldatenbestandes, ermittelt als $T^o = AX^o$

A – Zuordnungsmatrix ($a_{ij} = 1$, wenn Objekt j im Tabellenfeld i gezählt wird, sonst $a_{ij} = 0$)

$|T^z - T^o|$ – Vektor aller Tabellierungsfehler

B – Vektor der vorgegebenen (maximal zulässigen/„akzeptierten") Fehlerschranke (bound) für die einzelnen Tabellenfelder. b_i ist der zulässige Fehler im Feld i[5].

G – Vektor eventuell erlaubter zusätzlicher Fehler aufgrund der Größe der Tabellenfelder[6]

C – Vektor der Geheimhaltungsfälle $c_i = 1$, wenn x_i eine Häufigkeit von 1 oder 2 hat, sonst $c_i = 0$.

$S()$ – Straffunktion für zu kleine Randabweichungen. Entspricht die Abweichung eines Tabellenfeldes der erlaubten Maximalabweichung, so kann kein Satz mehr in diese Richtung geändert werden, ohne dass eine sofortige Kompensation innerhalb der Satzgruppe erfolgen muss (bei $t_i = 3$ und $b = 3$ ist ein weiteres Erhöhen blockiert, da dann $t_i = 4$ die bounds b verletzt). Es wird versucht, diese Konstellation zu vermeiden. Deshalb sind Tabellenfelder mit Abweichungen, kleiner als der zulässige Maximalfehler, günstiger und werden durch diese Straffunktion bevorzugt. Für die genaue Ausgestaltung dieser Funktion siehe Abschnitt 6a.

Veränderungen werden nur bei Einhaltung der gegebenen bounds b (ZF1) so vorgenommen, dass sie die Anzahl an Geheimhaltungsfällen minimieren (ZF2). Die Lösung der Aufgabe ist erreicht, wenn die Summe der $c_i = 0$ ist, also kein x_i mehr den Wert 1 oder 2 besitzt. Sollten bezüglich der Anzahl der Geheimhaltungsfälle neutrale Veränderungsmöglichkeiten existieren, so wird das Qualitätsziel 3 (ZF3) und bei Neutralität von 2 und 3 das Qualitätsziel 4 (ZF4) für die Entscheidung berücksichtigt.

Vor der Bearbeitung der Mikrodatendatei werden die Sätze so sortiert, dass möglichst wenige Schlüsseländerungen zwischen benachbarten Zeilen anzutreffen sind. Durch die Sortierung des Datenbestan-

5 Die Werte b_i können im Programm getrennt für ein- und mehrdimensionale Tabellenfelder verwaltet werden, wobei für mehrdimensionale Tabellenfelder nur ein fester zusätzlicher Abweichungswert für alle Felder gilt. Damit sind die b_i zwei einheitliche Werte Parameter für alle i (zwei Programmparameter).

6 Die Gewichtung der Tabellenfelder erfolgt nach der Größe des Tabellenwertes $g_i = G(t_i)$. Großen Tabellenfeldern wird so eine größere zulässige Abweichung erlaubt. Ein Beispiel implementierter Gewichtung ist $g_i = \text{int}(\log_{10} t_i)$. Damit werden folgende zusätzliche Abweichungen zulässig:

Tabellenwerte	1 bis 9	10 bis 99	100 bis 999	1 000 bis 9 999	…
zusätzliche Abweichung	0	1	2	3	

Es sind aber auch andere Gewichtungsfunktionen möglich (siehe Abschnitt 6f).

des befinden sich Datensätze nebeneinander, die in vielen Auswertungen im gleichen Tabellenfeld tabelliert werden. Bei der gleichzeitigen Behandlung von drei Unikaten würde durch die Veränderung von zwei Datensätzen zu 0 und einem zu 3 in diesen Tabellenfeldern eine Kompensation der Veränderungen auftreten. Um diese Kompensationseffekte bei der Veränderung benachbarter Sätze mit ausnutzen zu können, werden sequenziell gleitende Gruppen von drei, ggf. vier Sätzen aus der sortierten Datei gezogen. Für diese Gruppen werden alle möglichen Veränderungskombinationen ±1, ±2 und ggf. ±3 gebildet. Die Auswahl einer durchzuführenden Veränderung erfolgt dann nach folgendem Schema:

1. Ermittlung aller Veränderungskombinationen ±1, ±2 und ggf. ±3, für die die Häufigkeit der einzelnen Merkmalskombinationen nichtnegativ bleibt ($x_j >= 0$). Die Veränderung +2 kommt bei $x_j = 1$ und die Veränderung –2 bei $x_j = 2$ zur Anwendung, damit eine nicht anonyme Häufigkeit in beide Richtungen direkt auf anonym (d. h. 0 oder 3) geändert werden kann. Die Veränderung +3 kommt bei $x_j = 0$ und die Veränderung –3 bei $x_j = 3$ zur Anwendung, damit keine anonyme Häufigkeit auf 1 oder 2 (d. h. nicht anonym) zurückgeändert werden muss.
2. Für diese Veränderungskombinationen wird die Verletzung des vorgegebenen Maximalfehlers getestet und die Veränderung der Ziele 2–4 ausgewertet. Es werden dann nur noch die Veränderungskombinationen ausgewählt, die die Maximalfehlerbeschränkung nicht verletzen (keine Verletzung der bounds in ZF1 zulassen).
3. Die Menge aller Veränderungskombinationen wird reduziert um die Kombinationen, die nach ihrer Realisierung mehr Geheimhaltungsfälle erzeugen würden, als vorher vorhanden waren (keine Verschlechterung von ZF2 zulassen).
4. Die durchzuführende Veränderung wird bestimmt, indem man die Kombinationen auswählt, die das Ziel 2 (die Geheimhaltungsfälle zu beseitigen) am meisten verbessert. Verbleiben mehrere Kombinationen in der Auswahl, so wird das nächste Teilziel (erst Ziel 3, dann Ziel 4) für die Entscheidung herangezogen. Es werden ggf. auch die bezüglich höherwertiger Teilziele neutralen Kombinationen ausgewählt, wenn keine Kombination das höherwertige Teilziel verbessert. Eine ausgewählte Kombination verbessert somit mindestens ein Teilziel. Verschlechterungen eines Teilzieles sind nur bei gleichzeitiger Verbesserung eines höherwertigen Zieles möglich.

Ist die Ergebnismenge leer, findet keine Veränderung statt. Ansonsten wird die erste Kombination aus der Menge durchgeführt.

Beispiel:

Es ist eine Datei mit den qualitativen Merkmalen Wirtschaftszweigklassifikation (WZ) und Region zu anonymisieren. Als zu kontrollierende Randsummentabellen seien nur folgende eindimensionale Randsummen zu testen: die eindimensionalen Tabellierungen der Wirtschaftszweigklassifikation als 2-, 3- und 5-Steller und die Region für Berlin nach Stadtteil (Berlin-West, Berlin-Ost) und Bezirk. Mehrdimensionale Tabellen seien vernachlässigt. Die unteren Tabellen enthalten die Abweichungen in der Anzahl der Betriebe, die sich im Verlauf der bisherigen Anonymisierung ergeben haben.

Tabellenfelder, in denen keiner der Sätze der ausgewählten Gruppe tabelliert wird, können durch die Kombinationen auch nicht verändert werden. Deshalb sind nur die grau hinterlegten Tabellenfelder der folgenden Tabellen zu betrachten.

WZ 2-Steller	Abw.
…	…
CA	–1
CB	2
DA	2
DB	–1
DC	±0
…	…

WZ 3-Steller	Abw.
…	…
DA1	3
DA2	–1
DB1	±0
DB2	–1
DC1	±0
…	…

WZ 5-Steller	Abw.
…	…
DA189	1
DA191	1
DA196	–1
DA197	1
DA200	±0
…	…

Stadtteil	Abw.
Berlin-Ost	1
Berlin-West	–1

Bezirke	Abw.
…	…
03	1
04	–2
05	2
06	–1
…	…

Die aktuell zulässige Maximalfehlerschranke (bound b) sei ±3. Eine weitere Unterscheidung der Tabellenabweichung g_i sei vernachlässigt ($g_i = 0$ für alle Tabellenfelder). Folgende Satzgruppe des Datenbestandes wird aktuell untersucht:

Satz in der Satzgruppe	WZ	Bezirk	Stadtteil Berlin-…	Anzahl Betriebe	Veränderungsvarianten der Zeile
…	…	…	…	…	
1	DA191	03	Ost	3	–1, +1, –3
2	DA197	05	West	1	–1, +1, +2
3	DA200	06	West	1	–1, +1, +2
…	…	…	…	…	

Damit ergeben sich folgende Veränderungsvarianten:

Zielkriterien/Entscheidungskriterien:
1*) Bleibt Veränderung innerhalb der maximalen Fehlerschranke?
2*) Veränderung der Anzahl an Geheimhaltungsfällen,
3*) Veränderung des Randabstandes der Lösung (Bei diesem Beispiel sei vereinfacht die Anzahl der Tabellenfelder mit $|t_i| = b$ als ZF3 betrachtet),
4*) Veränderung der Summe der absoluten Randsummenfehler.

Nach Entfernung der Veränderungskombinationen, bei deren Realisierung die zulässigen Tabellenabweichungen überschritten werden (ZF1), werden auch die Kombinationen entfernt, deren Umsetzung mehr Geheimhaltungsfälle als vorher erzeugen würde (Spalte ZF2). Beide sind in Tabelle 1 grau unterlegt. Aus den verbleibenden Möglichkeiten wird nach folgender Auswahlregel gewählt (jeweils fett dargestellt):

1. Die meisten Geheimhaltungsfälle beseitigen die Kombinationen 17, 33 und 35 (jeweils 2).
2. Die größte Verbesserung des Randabstandes (oder geringste Verschlechterung) erzeugt darunter die Kombination 35. Wären hier immer noch mehrere Kombinationen gleichwertig, entscheidet der Einfluss auf den mittleren Randsummenfehler (Spalte ZF4 möglichst klein).

Die vorgeschlagenen Veränderungen der Kombination 35 werden durchgeführt. Bei diesem Beispiel ist erkennbar, dass ggf. auch nicht geheim zu haltende Fälle mit verändert werden, wenn es für das Gesamtproblem nützlich ist. Diese Veränderung von nicht geheim zu haltenden Fällen ermöglicht beispielsweise die Revision von bereits getroffenen Änderungsentscheidungen, aber auch die Bildung größerer anonymer Gruppen als 3. Sie werden aber nur eingeschränkt geprüft, z. B. wenn sonst keine Lösung möglich ist (siehe weiter unten „Auswahltechniken der zu testenden Satzgruppen").

Danach wird eine neue Satzgruppe von x Sätzen aus der Datei gezogen, die nach den gleichen Entscheidungsregeln getestet und bearbeitet wird. Dieser Algorithmus „Auswahl einer Gruppe von Sätzen und Entscheidung der Veränderung" wird ständig wiederholt. Da eine Veränderung nur bei Annäherung an das Ziel (siehe Zielfunktionen) durchgeführt wird, können keine Schleifen auftreten. Die Veränderung zu einer alten Zwischenlösung wäre nur bei einer Verletzung der obigen Entscheidungsregeln möglich.

Aus kombinatorischer Sicht wäre es auch möglich, den ersten Satz unverändert zu lassen und nur den zweiten und/oder dritten Satz zu ändern. Diese Kombinationen werden aber nicht geprüft, da sie in der folgenden Satzgruppe mit enthalten sind und dort aber zusätzlich die Kompensation mit den nachfolgenden Sätzen geprüft wird.

1 | Beispiel für Entscheidungsregeln

Variante	Veränderung Satz ... um …			Veränderung Zielkriterien				Bemerkung
	1	2	3	ZF1*)	ZF2*)	ZF3*)	ZF4*)	
1	–1	–1	-1	nein	–1	0	–3	
2	–1	–1	1	nein	0	–1	–10	
3	–1	–1	2	nein	–1	–1	–6	
4	–1	–1	0	nein	0	–1	–8	
5	–1	1	–1	nein	0	1	1	
6	–1	1	1	nein	1			
7	–1	1	2	ja				
8	–1	1	0	nein	1			
9	–1	2	–1	ja				
10	–1	2	1	ja				
11	–1	2	2	ja				
12	–1	2	0	ja				
13	–1	0	–1	nein	0	–1	–2	
14	–1	0	1	nein	1			
15	–1	0	2	nein	0	0	–1	
16	–1	0	0	nein	1			
17	1	–1	–1	nein	**–2**	**1**	5	
18	1	–1	1	nein	–1	1	1	
19	1	–1	2	ja				
20	1	–1	0	nein	–1	0	2	
21	1	1	–1	ja				
22	1	1	1	ja				
23	1	1	2	ja				
24	1	1	0	ja				
25	1	2	–1	ja				
26	1	2	1	ja				
27	1	2	2	ja				
28	1	2	0	ja				
29	1	0	–1	ja				
30	1	0	1	ja				
31	1	0	2	ja				
32	1	0	0	ja				
33	–3	–1	–1	nein	**–2**	**1**	4	
34	–3	–1	1	nein	–1	–1	–4	
35	–3	–1	2	nein	**–2**	**–1**	**–3**	zu wählende Kombination
36	–3	–1	0	nein	–1	–1	–1	
37	–3	1	–1	nein	–1	0	4	
38	–3	1	1	nein	0	0	0	
39	–3	1	2	nein	–1	0	5	
40	–3	1	0	nein	0	0	–1	
41	–3	2	–1	ja				
42	–3	2	1	ja				
43	–3	2	2	ja				
44	–3	2	0	ja				
45	–3	0	–1	nein	–1	–1	3	
46	–3	0	1	nein	0	–1	–5	
47	–3	0	2	nein	–1	–1	0	
48	–3	0	0	nein	0	–1	–2	

6. Aspekte des Algorithmus

a) Maximierung der Möglichkeiten für weitere Veränderungen (Teilziel ZF3)

Bei einer Auswahl aus mehreren Veränderungsmöglichkeiten, die die gleiche Anzahl an Geheimhaltungsfällen beseitigen, wird zuerst der Einfluss auf die Möglichkeiten für weitere Veränderungen berücksichtigt. Innerhalb der Satzgruppe können sich in der Regel nicht alle Veränderungen in den Tabellierungsfehlern gegenseitig kompensieren, da die Sätze nicht vollständig identisch sind. Eine einzelne Datensatzveränderung ist nur dann gefährdet, wenn die Fehler in den Tabellenfeldern sich auf beiden Seiten (positive und negative Abweichungen) gleichzeitig zu dicht am Rand der zulässigen Abweichung befinden. Für die Beseitigung eines Geheimhaltungsfalles durch Verringerung der Häufigkeit um 1 darf kein Tabellenfeld i einen negativen Fehler haben, der gleich dem zulässigen Maximalfehler $-(b_i+g_i)$ ist. Analog ist die Beseitigung eines Geheimhaltungsfalles durch Erhöhung der Häufigkeit von 1 auf 3 nur dann möglich, wenn die Abweichung des Tabellenfeldes nach oben weder (b_i+g_i) noch $(b_i+g_i)-1$ beträgt. Weiterhin verhindert eine Tabellenfeldabweichung von $>=(b_i+g_i)-2$ bzw. $<=2-(b_i+g_i)$ die Korrektur einer Geheimhaltungsentscheidung in Form eines Wechsels der Häufigkeit von 0 auf 3 oder umgekehrt. Deshalb werden diese Abweichungen in Tabellenfeldern im Teilziel ZF3 als „kritische Abweichungen" berücksichtigt, wenn eine Auswahlmöglichkeit unter mehreren Kombinationen besteht, die die gleiche Anzahl an Geheimhaltungsfällen beseitigen.

Eine einfache Fehlerfunktion zum „mittleren" Tabellierungsfehler (wie ZF4) kann dem Ziel „Maximierung der Möglichkeiten für weitere Veränderungen im Datenbestand" nicht gerecht werden, vor allem dann nicht, wenn auch noch versucht wird, mit zwei getrennten Maximalfehlerschranken für ein- und mehrdimensionale Tabellen zu arbeiten. Als Regel wird nebenstehende Straffunktion S zur Messung des Randabstandes verwendet.

$$S=\sum_{i=1}^{k} s_i$$

$$s_i=\begin{cases} 9 & ;\forall\,|f_i|=b_i+g_i \\ 4 & ;\forall\,|f_i|=b_i+g_i-1 \\ 1 & ;\forall\,|f_i|=b_i+g_i-2 \\ 0 & ;\forall\,|f_i|<b_i+g_i-2 \end{cases}$$

Die Minimierung dieser Straffunktion steht als Ziel somit vor der allgemeinen Verbesserung der Summe der Tabellierungsfehler, weil sie sich positiv auf die Wahrscheinlichkeit auswirkt, dass weitere Veränderungen möglich sind.

b) Numerische Probleme

Bei numerischen Algorithmen können mehrere Probleme auftreten:

1. Schleifen

   Um zu verhindern, dass der Algorithmus sich in einer endlosen Schleife aufhängt, bestehen einige Anforderungen an die Auswahlregeln. Die Funktionen, mit deren Hilfe die Auswahl getroffen wird, müssen folgende beiden Bedingungen erfüllen:

- Für zwei beliebige Lösungen X_1 und X_2 des Lösungsraumes gilt: Der Abstand von X_1 nach X_2 ist gleich dem negativen Abstand X_2 nach X_1

$$\overline{X_1X_2}=-\overline{X_2X_1}$$

Diese Bedingung würde beispielsweise dann vernachlässigt, wenn man versuchte, sich bei der Straffunktion nur auf die störenden Tabellenfelder eines gerade in der Auswahlgruppe betrachteten Geheimhaltungsfalls zu konzentrieren. Eine „Bevorzugung" eines aktuellen Geheimhaltungsfalles führt bei einem Wechsel zum nächsten Geheimhaltungsfall automatisch dazu, dass die obige Regel verletzt ist und somit Schleifen nicht mehr ausgeschlossen werden können. Im obigen Fall hat jede Lösung für alle drei Zielkriterien immer genau einen Funktionswert (unabhängig von der Veränderungsrichtung). Damit gilt:

$$\overline{X_1X_2}=f(X_2)-f(X_1)=-(f(X_1)-f(X_2))=-\overline{X_2X_1}$$

- Die Zielkriterien müssen einerseits in Optimierungsrichtung beschränkt sein und andererseits sicherstellen, dass es eine Lösungsmenge gibt, die die gesuchte Lösung enthält. Die zweite Teilfunktion (ZF2 – Anzahl der vorhandenen Geheimhaltungsfälle) ist in Optimierungsrichtung (nach unten) beschränkt, denn es können nicht mehr Geheimhaltungsfälle entfernt werden, als vorhanden sind. Gleichzeitig ist das Minimum (0 Geheimhaltungsfälle) identisch mit der gesuchten Lösung. Die dritte Teilfunktion (ZF3 – Straffunktion für in Nähe des Maximalfehlers liegende Tabellenfelder) ist ebenfalls nach unten beschränkt. Sie ist 0, wenn kein Tabellenfeld einen Wert im Bereich $f_{max} \geq abs(f_i) \geq f_{max}-2$ besitzt. Die vierte Teilfunktion (ZF4 – Summe der Fehler in Tabellenfeldern) hat ihr Minimum, wenn alle Tabellenfelder fehlerfrei sind. Da für die dritte und vierte Teilfunktion keine Einschränkungen bezüglich Zulässigkeit gestellt werden, sind alle Kombinationen bezüglich dieser Teilfunktionen zulässige Lösungen. Die Zielfunktion ZF1 ist ebenfalls nach unten beschränkt. Der kleinste absolute Fehler kann nur 0 sein. Es besteht aber das Problem, dass nicht sicher ist, ob bei einem Fehler von 0 eine Lösung existiert. Die Bestimmung einer Lösungsmenge, die die Lösung enthält, wird im Rahmen der im folgenden dargestellten Stagnation gelöst.

2. Stagnation

   Es kann vorkommen, dass zwar einerseits der Algorithmus zum Ziel konvergiert, andererseits aber die Geschwindigkeit so langsam ist, dass man nicht in einer vertretbaren Zeit zum Ergebnis kommt. In diesem Fall muss der Algorithmus die Situation erkennen können und reagieren. Das wird bei diesem Verfahren mit dem „Gashebel" zulässiger Maximalfehler (bound) geregelt. Wird bei einem Durchlauf nicht ein erwarteter Anteil von Geheimhaltungsfällen beseitigt, so erfolgt eine Vergrößerung der bounds. Damit entstehen wieder größere Freiräume, wodurch beim erneuten Durchlauf weitere Geheimhaltungsfälle beseitigt werden können. Die Geschwindigkeit des Verfahrens lässt sich somit über den Parameter „Anteil der mindestens zu beseitigenden Geheimhaltungsfälle" regeln. Eine zu große Geschwindigkeit geht hierbei jedoch zu Lasten der Qualität des Ergebnisses (höherer Maximalfehler in der Lösung).

Dass es bounds geben muss, für die eine Lösung existiert, kann man mit folgendem Beispiel zeigen. Verwendet man eine vereinfacht generierte Startlösung wie beispielsweise jeden dritten Satz des Datenbestandes mit der Häufigkeit 3, so lassen sich auch aus diesem Bestand alle Kontrolltabellen berechnen. Verwendet man die aus diesen Kontrolltabellen bestimmbare Maximalabweichung als bound, so ergibt sich ein Lösungsraum an möglichen Datenbeständen, für den mindestens eine zulässige Lösung existiert (die generierte Triviallösung). Auch wenn diese Lösung mit Sicherheit nicht die gesuchte Lösung ist, so genügt sie dem Existenzbeweis einer Lösung, der für den Nachweis der Lösbarkeit erforderlich ist.

c) Auswahltechniken der zu testenden Satzgruppen

Die Entscheidung zur Beseitigung der Geheimhaltungsfälle durch Verändern der Häufigkeit wird nur dann verhindert, wenn die Veränderung der Häufigkeit mit einer Verletzung der Schranke der erlaubten Maximalabweichung (b_i+g_i) in einem Tabellenfeld einhergehen würde. Dann wird die Entscheidung verschoben. Durch nachfolgende Veränderungen kann es durchaus sein, dass bei einem erneuten Testen der gleichen Satzgruppe die dann vorhandenen Abweichungen im Tabellenfeld die Beseitigung ermöglichen, da sich die konkreten Abweichungen mit jeder Veränderung im Datenbestand ebenfalls verändern können und auf die Schaffung von entsprechenden Freiräumen Wert gelegt wurde. Außerdem können ggf. Teile der Satzgruppe gruppiert mit nachfolgenden Sätzen anonymisierbar sein.

Durch gezielte Auswahltechniken sind folgende Probleme zu lösen:

- Gruppierung zu Häufigkeiten größer 3,
- automatische Erkennung der erforderlichen bounds,
- Kontrolle und ggf. Korrektur von „alten" Geheimhaltungsentscheidungen.

Zu Beginn eines Anonymisierungslaufs ist die erforderliche Maximalabweichung in der Regel nicht bekannt. Bei mehrfachen Läufen für den gleichen Datenbestand (z. B. Monatsreihen) könnten ggf. Erfahrungen vorliegen. Es muss aber ein Startwert der bounds vorgegeben werden. Problematisch sind dabei zu eng gestellte bounds. Das würde dazu führen, dass zuerst nur die homogenen Satzgruppen anonymisiert (zusammengefasst) werden, und im Nachhinein immer inhomogenere Sätze übrig bleiben, denn die gleichartigen Satzgruppen dazwischen wurden entfernt. Bei einem ständigen Durchsuchen der Gesamtdatei würden mit großem Testaufwand die homogen Satzgruppen anonymisiert, während die Anonymisierung der inhomogenen Sätze immer weiter verschoben werden. Um die erforderlichen bounds automatisch und schnell durch den Algorithmus zu erkennen, wird folgendes Verfahren verwendet:

Es wird nur ein kleiner Teil der Datei bearbeitet. Für diesen Teil wird die Anonymisierung durchgeführt. Erst, wenn dieser Teil fast vollständig abgearbeitet ist, wird ein weiterer Teil der Gesamtdatei dazu genommen. Das „fast vollständig" ist erforderlich, um strukturelle Ausreißer mit den größeren kombinatorischen Möglichkeiten der Gesamtdatei entscheiden zu können. Ist die Anonymisierung für diesen Teil nicht möglich, wird bereits der bound-Parameter im Modell höher gesetzt. Eine Erhöhung des Parameters schafft in jedem Fall neue Freiräume, um weitere Veränderungen an geheim zu haltenden Sätzen vorzunehmen. Somit kann auch bei sehr großen Dateien schnell der notwendige bound-Parameter erkannt werden und das Verfahren bedeutend schneller laufen. Die verbleibenden Reste des bereits bearbeiteten Teilbestandes werden regelmäßig wieder mitgetestet, da sich durch jede Veränderung auch die Tabellierungsabweichungen ändern und somit ständig andere Möglichkeiten existieren.

Beim Durchlaufen der Datei werden zuerst standardmäßig nur die noch existierenden Geheimhaltungsfälle betrachtet. Es werden immer drei benachbarte Geheimhaltungsfälle getestet. Tritt trotzdem eine Stagnation des Verfahrens ein, so werden nacheinander andere Auswahltechniken durchgeführt, wobei nach jedem Lauf getestet wird, ob die Stagnation noch vorhanden ist. Die weiteren Auswahltechniken sind:

1. Es werden alle noch vorhandenen Sätze (Häufigkeit > 0) getestet. Da es auch vorkommen kann, dass in der Nähe eines Geheimhaltungsfalles (zwei Sätze davor und dahinter) alle Sätze entfernt wurden, werden im ersten Versuch alle nicht mehr vorhandenen Sätze (Häufigkeit = 0) vernachlässigt und die Auswahlgruppen aus der verbleibenden Menge gebildet. Dieses Verfahren zeichnet sich auch dadurch aus, dass mit kleinen Veränderungen der Häufigkeit bei bereits anonymisierten Sätzen (±1) auch die Teilziele 3 und 4 verbessert werden können. Durch diese Gruppenauswahl wird beispielsweise die Möglichkeit, einen Geheimhaltungsfall zu einer bereits vorhandenen 3er-Gruppe mit hinzuzunehmen, getestet. Es werden dann vier identische Einheiten gebildet.
2. Die noch vorhandenen Geheimhaltungsfälle werden zusammen mit ihren im sortierten Bestand physisch benachbarten Sätzen getestet. Das können auch bereits durch die Anonymisierung entfernte Sätze sein, wobei die Entscheidung für diese benachbarten Sätze revidiert werden kann.
3. Als letzte Variante kann auch ein Gesamtdurchlauf der Datei durchgeführt werden. Dann könnte auch das Entfernen eines Satzes bei allen Sätzen noch einmal revidiert werden, wenn es dem Gesamtziel (siehe Entscheidungsregeln) dient. Dieses Auswahlverfahren ist am rechenaufwändigsten und wird deshalb nur dann angewendet, wenn bereits der Gesamtbestand in den Anonymisierungslauf einbezogen ist und nur noch sehr vereinzelt Geheimhaltungsfälle existieren.

Da die Anzahl der zu ziehenden Satzgruppen und auch die Anzahl der Änderungsmöglichkeiten in den Satzgruppen bei den verschiedenen Auswahltechniken unterschiedlich ist, wird auch die Erwartung bezüglich der Anzahl der zu beseitigenden Geheimhaltungsfälle für die Auswahltechniken verschieden

angesetzt. Auswahltechniken, bei denen sehr viele Satzgruppen gebildet und getestet werden, müssen auch entsprechend mehr Geheimhaltungsfälle beseitigen, um noch als performant zu gelten.

d) Zeitverlauf/Zeitbedarf

Das Verfahren zeichnet sich durch einen hyperbolischen Zeitverlauf aus. Einer sehr schnellen Anonymisierung der ersten Hälfte folgt eine sehr langsame Anonymisierung der zweiten Hälfte der Merkmalsträger. Nach ca. 50 % der Gesamtlaufzeit sind in der Regel nur noch weniger als 5 % der Geheimhaltungsfälle zu anonymisieren. Für diese wird wegen der Nutzung von aufwändigen Auswahlalgorithmen mehr Zeit benötigt. Es wäre durch Anpassung des Stagnationsmaßes auch möglich, den Zeitverlauf linearer zu gestalten. Das geht aber mit einer Erhöhung der Maximalabweichung im Ergebnis einher (bei Tests ergab sich eine ca. 25 % größere Maximalabweichung).

e) Eindimensionale Tabellierungen besser erhalten

Die Wertigkeit eindimensionaler Tabellenfelder wird meist höher eingeschätzt als die mehrdimensionaler Tabellenfelder. So ist Datennutzerinnen und -nutzern beispielsweise die Anzahl der Einwohner insgesamt in einer Region wichtiger als die Häufigkeiten in komplexeren mehrdimensionalen Auswertungen. Es wurde von Datenproduzentinnen und -produzenten sowie Nutzerinnen und Nutzern die externe Forderung postuliert, dass diese Tabellenfelder einen geringeren Fehler im Ergebnis aufweisen müssen. Dabei wurde als Zielvorstellung für eindimensionale Auswertungen der theoretische Minimalfehler von ±2 angestrebt. Für weniger als 2 wurden zuvor bereits Gegenbeispiele gezeigt, bei denen die Forderung nicht realisierbar ist.

Beim SAFE-Programm können deshalb unterschiedliche Fehlerschranken ein- und mehrdimensional vorgegeben werden. Da eine ständige Restriktion von 2 jedoch für das numerische Verfahren zu eng ist, wurde folgende Regel eingebaut. Es können dem Programm drei Parameter übergeben werden. Die Parameter haben folgende Bedeutung:

1. Startfehler eindimensionaler Tabellenfelder,
2. Startfehler mehrdimensionaler Tabellenfelder,
3. maximaler Abstand der beiden Fehlerschranken.

Bei Stagnation und der Entscheidung, die bounds b zu vergrößern, werden so lange nur die b_i mehrdimensionaler Tabellenfelder um 1 erhöht, bis der maximale Abstand erreicht ist, danach erfolgt ein gleichzeitiges Erhöhen aller b_i um 1.

f) Zusätzliche Gewichtung bei großen Zellwerten

In den Tabellen sind bei großen Zellwerten größere Abweichungen in den Tabellenfeldern durch den eingeführten Fehlervektor g zulässig. Eine größere Abweichung bedeutet bei einem großen Tabellenwert eine kleinere relative Veränderung und kann deshalb dort eher akzeptiert werden. In der Anwendung von SAFE werden verschiedene Regeln für Größenklassen verwendet. Für alle gilt: Je größer die Größenklasse, desto größer ist der zulässige zusätzliche Fehler.

Während in der ersten Version (Variante 1) die Gewichte im Fehlervektor g nach dem dekadischen Logarithmus gebildet wurden, ist aktuell auch eine zweite Gewichtungsversion im Einsatz. Bei beiden aktuellen Versionen gibt es zehn Größenklassen. Die dekadischen Größenklassen entsprechend dem dekadischen Logarithmus, werden also als $g_i = \text{int}(\log_{10} t_i)$ gebildet (Tabelle 2, Variante 1). Die größte Größenklasse enthält Tabellenfelder mit einer Besetzung von $>= 1000\,000\,000$. Bei der zweiten Version gibt es im Bereich der Zellbesetzungen zwischen 10 und 1000 Einheiten noch weitere Untergliederungen (Tabelle 2, Variante 2). Beide Gewichtungsversionen wurden bereits mehrfach getestet, die Version mit den zusätzlichen Größenklassen bei Besetzungszahlen unter 1000 wird derzeit bevorzugt, da in den „kleineren" Tabellenfeldern, bei denen eine Abweichung um 1 noch größere Auswirkungen auf den relativen Fehler hat, die Abweichungen noch stärker minimiert werden.

7. Korrektur der Lösung

Um dem Wunsch nach einem möglichst kleinen Maximalfehler stärker nachzukommen, wird im Anschluss an die Anonymisierung noch folgende Korrekturaufgabe gelöst. Ziel ist es, sowohl die eindimensionalen Tabellenfelder stärker zu verbessern als auch die unterschiedliche Gewichtung schrittweise aus der Lösung zu entfernen. Der Ablaufplan zur Softwareumsetzung der Aufgabe ist in Abbildung c dargestellt.

Dazu wird die Optimierungsaufgabe „Optimierung der Lösung" gelöst:

2 | Tabellenwerte der Gewichtungsversionen

zusätzliche Abweichung	Variante 1		Variante 2	
0	1 bis	9	1 bis	9
1	10 bis	99	10 bis	19
2	100 bis	999	20 bis	49
3	1 000 bis	9 999	50 bis	99
4	10 000 bis	99 999	100 bis	199
5	100 000 bis	999 999	200 bis	999
6	1 000 000 bis	9 999 999	1 000 bis	9 999
7	10 000 000 bis	99 999 999	10 000 bis	99 999
8	100 000 000 bis	999 999 999	100 000 bis	999 999
9	1 000 000 000 und mehr		1 000 000 und mehr	

$$
\begin{aligned}
&\text{ZF 1:} \quad \min\left(\max(b)\right) \\
&\text{ZF 2:} \quad \min\left(\sum_{i=1}^{n} c_i\right) \\
&\text{ZF 3:} \quad \min\left(\sum_{i=1}^{k} S_K\left(\left|t_i^z - t_i^o\right|\right)\right) \\
&\text{ZF 4:} \quad \min\left(\sum_{i=1}^{k} \left|t_i^z - t_i^o\right|\right) \\
&\left|T^z - T^o\right| \le F \\
&T^z = AX^z
\end{aligned}
\qquad \text{mit:} \quad
\begin{aligned}
&x_j^z = 0,3,4,\ldots \\
&c_j = 1 \;\; ;\text{gdw.} \;\; t_j^z > f_j \\
&c_j = 0 \;\; ;\text{gdw.} \;\; t_j^z \le f_j
\end{aligned}
\tag{4}
$$

mit:

C – Vektor der Schrankenverletzungen
$c_i = 1$, wenn im Tabellenfeld t^z_i die Abweichung den zulässigen Wert f_i übersteigt, sonst $c_i = 0$.

Alle anderen Variablen behalten ihre Bedeutung.

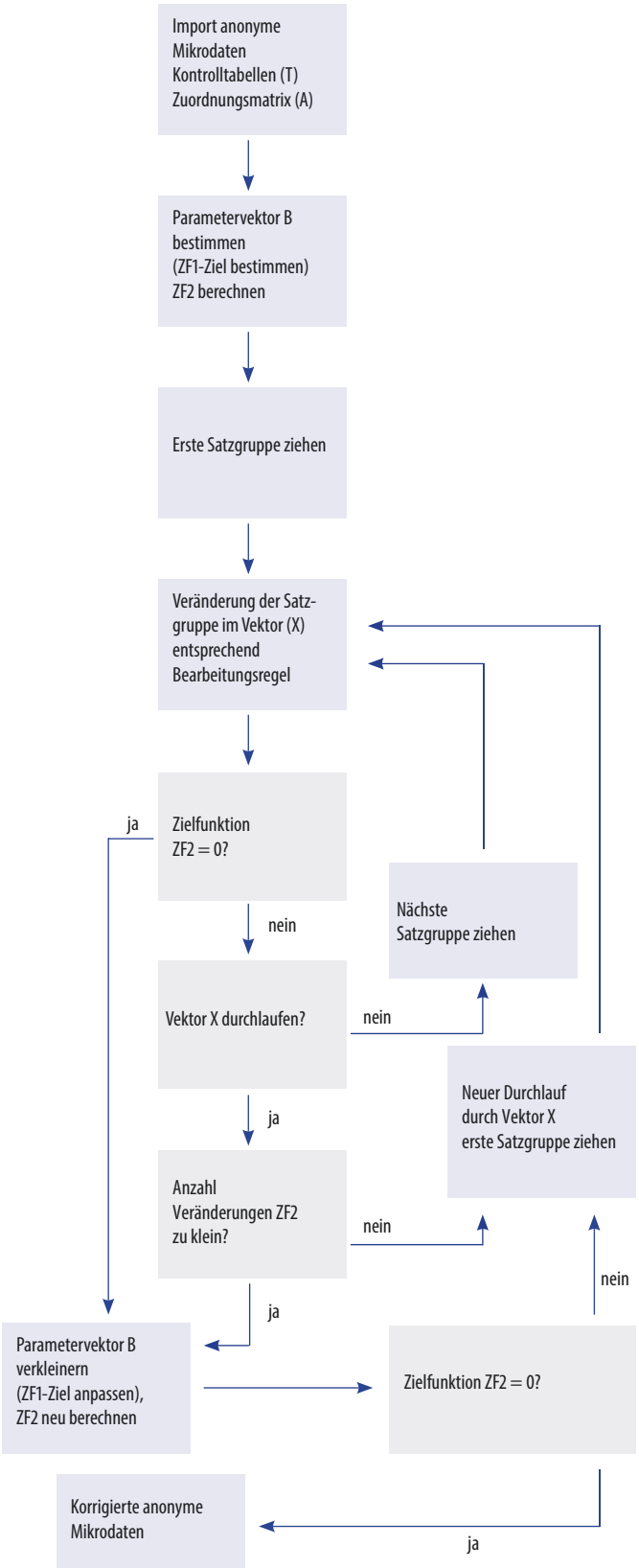
Im Unterschied zur Aufgabe 3, bei der keine Schrankenverletzungen erlaubt waren, wird in dieser Aufgabe d gezielt die Verletzung einzelner Schrankenwerte beseitigt. Diese Schrankenverletzung kann erzeugt werden, indem die zulässigen Abweichungen in Tabellenfeldern schrittweise verkleinert werden. Die zulässigen Abweichungen ergeben sich einerseits aus dem Typ des Tabellenfeldes (ein- oder mehrdimensionale Auswertung) sowie andererseits aus der Größe des Wertes im Tabellenfeld, d. h. $f_i = b_i + g_i$. Da diese Werte sowohl von dem Typ der Auswertung und der Größe des Tabellenfeldes abhängen, werden sie als entsprechende zweidimensionale Liste behandelt (Tabelle 3). Bei mehrdimensionalen Tabellenfeldern wird eine um l erhöhte Abweichung erlaubt als bei den eindimensionalen.

Für die Korrektur der Lösung werden dem Algorithmus zwei Parameter übergeben, die die Zielparameter der durch die Korrektur zu erreichenden Abweichungsschranken für ein- und mehrdimensionale Tabellenfelder beschreiben. Das Programm versucht, beginnend bei den durch die Lösung der Aufgabe 1 erreichten Maximalabweichungen, durch schrittweises unabhängiges Verkleinern der einzelnen gewichteten bounds (Summe aus bound-Parameter und Gewicht) die Korrektur stufenweise durchzuführen. Es werden die in Tabelle 3 erhaltenen gewichteten bounds von links oben beginnend um 1 reduziert, bis eine weitere Reduzierung die Anzahl der so entstehenden Schrankenverletzungen eine vorgegebene Maximalzahl übersteigen würde. Beim Reduzieren wird die Regel beachtet, dass gewichtete bounds für ein höheres Tabellengewicht nicht kleiner sein dürfen als für ein kleineres Tabellengewicht (siehe folgendes Beispiel). Danach wird der Datenbestand wie bei Aufgabe 1 iterativ bearbeitet und durch die Änderung der Häufigkeiten im Bereich der anonymen Lösungen diese Anzahl verkleinert. Nicht anonyme Häufigkeiten sind im Korrekturlauf nicht mehr zulässig. Ist die Anzahl der bounds-Verletzungen 0, werden die bounds für die Tabellenfelder neu berechnet, die noch nicht die Zielparameter erreicht haben, und die Aufgabe 2 neu gelöst. Sind die Zielbounds erreicht oder beim weiteren Verkleinern eines einzelnen bounds wird die Anzahl der Schrankenverletzungen so groß, dass die Lösbarkeit als unwahrscheinlich eingeschätzt wird, wird das Programm beendet.

3 | Abweichungen nach Größenklassen und Dimensionen der Tabellenfelder

Typ der Auswertung	Gewicht g_i des Tabellenfeldes t_i					
	0	1	2	3	4	5
Eindimensional	k	k+1	k+2	k+3	k+4	k+5
Mehrdimensional	k+l	k+l+1	k+l+2	k+l+3	k+l+4	k+l+5

c | Programmablauf „Optimierung der Lösung“



Beispiel:
Die Abweichung der Tabellenfehler sei das Ergebnis der Anonymisierung. Als Ergebnis der Korrektur wird eine Maximalabweichung von 8 angestrebt:

Abweichung	Gewichtung des Tabellenfeldes g_i								
	0	1	2	3	4	5	6	7	8
Ergebnis der Anonymisierung									
0	82 527	20 772	18 737	10 368	8 654	18 169	14 005	1 885	86
1	267 988	40 163	35 643	19 349	15 911	34 570	26 519	3 417	180
2	143 137	28 356	27 313	16 814	13 937	29 883	23 773	3 115	196
3	31 107	16 021	17 078	11 028	11 468	24 385	19 265	2 524	141
4	6 625	5 827	11 621	6 926	6 592	19 210	14 501	1 932	122
5	522	1 550	3 254	5 302	4 118	8 986	10 208	1 405	116
6	15	154	885	1 150	3 521	5 260	4 255	992	90
7	0	11	88	265	527	5 413	2 310	480	67
8	0	0	4	25	102	431	2 609	274	30
9	0	0	0	0	7	58	84	251	17
10	0	0	0	0	0	3	14	23	20
11	0	0	0	0	0	1	1	1	2
12	0	0	0	0	0	0	1	0	1
13	0	0	0	0	0	0	0	0	1
14	0	0	0	0	0	0	0	0	0
>= 15	0	0	0	0	0	0	0	0	0
Maximal-abweichung	6	7	8	8	9	11	12	11	13
Parameter der bounds	6	7	8	8	8	10	11	11	12
Ergebnis des 1. Durchlaufs der Korrektur									
0	82 238	20 660	18 684	10 302	8 512	18 317	14 071	1 866	95
1	267 800	40 168	35 493	19 453	16 102	34 613	26 840	3 599	196
2	143 012	28 319	27 301	16 724	14 142	30 242	24 100	3 175	195
3	31 602	15 878	17 591	11 179	11 190	24 469	19 378	2 611	164
4	6 700	6 030	11 215	7 305	7 217	18 019	13 906	1 900	103
5	554	1 602	3 313	4 783	4 567	9 787	9 286	1 218	111
6	15	186	964	1 153	2 514	6 599	5 230	828	70
7	0	11	61	318	572	3 688	3 368	590	70
8	0	0	1	10	21	539	1 189	389	45
9	0	0	0	0	0	96	155	101	18
10	0	0	0	0	0	0	22	19	2
11	0	0	0	0	0	0	0	3	0
>= 12	0	0	0	0	0	0	0	0	0
Maximal-abweichung	6	7	8	8	8	9	10	11	11
Parameter der bounds	6	7	8	8	8	8	9	10	10
Ergebnis des 17. Korrekturdurchlaufs zur Parameterneubestimmung									
0	81 778	20 212	18 706	10 030	8 368	18 783	15 665	2 099	150
1	267 313	39 945	34 415	19 276	16 048	35 275	28 726	4 030	215
2	142 585	28 529	27 365	16 393	13 926	30 858	24 709	3 387	201
3	32 530	15 818	18 164	11 463	11 010	24 447	19 414	2 676	172
4	7 041	6 317	10 788	7 491	75 22	18 163	14 861	1 949	107
5	654	1 802	3 852	4 564	4 821	10 813	8 440	11 61	91
6	20	225	1 242	1 517	2 259	5 170	3 765	595	74
7	0	6	90	464	827	2 670	1 842	371	53
8	0	0	1	29	56	190	123	31	6
>= 9	0	0	0	0	0	0	0	0	0

Die Tabellenfelder, die die gesetzten bounds überschreiten, sind grau hinterlegt.

Nach 17 Korrekturdurchläufen (mit bounds-Parameterbestimmungen) ist das endgültige Korrekturergebnis erreicht.

Es erfolgt der Abbruch der Korrektur, da das Ziel einer Maximalabweichung von 8 erreicht ist.

Innerhalb der Aufgabe 2 wird eine neue Straffunktion für die Lösungskorrektur verwendet (S_K). Diese Straffunktion versucht Tabellenfelder mit der gerade zulässigen Maximalabweichung verstärkt zu verändern. Diese Abweichungen erhalten einen deutlich erhöhten Strafterm in der neuen Straffunktion. Es gilt wieder die gleiche Abstufung zwischen den Zielfunktionen. Die erste Zielfunktion sind die gesetzten Abweichungsschranken. Es werden wieder Veränderungskombinationen in Satzgruppen getestet. Werden dabei Sätze verändert, die in einem Tabellenfeld die bounds überschritten haben, so sind Veränderungen nur möglich, wenn sie diese Überschreitung danach beseitigen (ZF2). Dadurch wird verhindert, dass eine Überschreitung von ±1 durch die Änderung vergrößert wird. Danach werden die Veränderungen an der Veränderung in ZF3 und bei dort neutralen Veränderungen an ZF4 ausgewählt:

$$S_K = \sum_{i=1}^{k} s_{K\,i}$$

$$s_{Ki} = \begin{cases} 500 & ;\forall(\left|t_i^z - t_i^o\right| = f_i) \\ 9 & ;\forall(\left|t_i^z - t_i^o\right| = f_i - 1) \\ 4 & ;\forall(\left|t_i^z - t_i^o\right| = f_i - 2) \\ 1 & ;\forall(\left|t_i^z - t_i^o\right| = f_i - 3) \\ 0 & ;\forall(\left|t_i^z - t_i^o\right| < f_i - 3) \end{cases}$$

8. Rematching – Zuordnung zu Identifikatoren

Als separater Programmbaustein steht eine Routine zum Rematching zur Verfügung. Ergebnis des SAFE-Laufs ist eine Datei, in der die Merkmalskombinationen mit ihrer originalen und anonymen Häufigkeit stehen.

Das Programm nimmt eine Zuordnung der mit der SAFE-Anonymisierung erhaltenen anonymen Lösung zu den Sätzen des originalen Datenbestandes vor. Damit kann auch ein Zurückspielen von anonymen Daten anstelle der originalen in Auswertungsdatenbanken erfolgen. Diese müssen dann nicht als unabhängige Daten ausgewertet werden, sondern können auch die DB-Verknüpfungen zu anderen Tabellen oder quantitativen Werten weiterhin verwenden.

Dazu wird für die Merkmale des Datenbestandes eine Prioritätsreihenfolge festgelegt, die angibt, welche Merkmale/Merkmalsstufen wann aus dem Vergleich ausgeschlossen werden sollen. Die Anzahl der Merkmalsstufen sei *m*.

Für die Zuordnung wird folgender Algorithmus abgearbeitet:

1. Sortiere den originalen und den anonymen Datenbestand in der gewählten Prioritätsreihenfolge der Merkmale/Merkmalsstufen.
2. Setze die Anzahl der aktuell in den Vergleich einzubeziehenden Merkmale i auf i = m.
3. Gruppiere den noch nicht zugeordneten Datenbestand der anonymen Lösung in Gruppen mit identischen Ausprägungen in den ersten i-Merkmalen. Gruppiere den originalen Datenbestand analog. Wähle die erste Gruppe des Datenbestandes aus.
4. Für die gewählte Gruppe des anonymen Datenbestandes (Menge A) suche die analoge Gruppe aus dem noch nicht zugeordneten originalen Datenbestand (Menge B), die in den ersten i-Merkmalen die gleichen Ausprägungen hat.
5. Solange die Menge B nicht leer ist, suche für den ersten Satz der Menge A aus der Menge B den Satz heraus, der in den meisten Merkmalen (auch den nicht mehr in die Gruppierung einbezogenen) mit dem gewählten Satz aus A übereinstimmt. Kennzeichne diese beiden Sätze als Zuordnungspaar und entferne den Satz aus der Menge A (der „noch nicht zugeordneten" Sätze) und den Partner der aus der Menge B der originalen nicht zugeordneten Sätze (Kennzeichnung als nicht mehr „für Zuordnungen verfügbar"). Sind die Mengen A und B nicht leer, bearbeite die verbleibenden Sätze der Menge A durch Wiederholung des Schrittes 4.
6. Ist die Menge B leer (keine möglichen Partner vorhanden) oder die Menge A komplett zugeordnet, so bearbeite die nächste Gruppe des anonymen Datenbestandes (neue Menge A). Dazu wird diese Gruppe wieder ab Schritt 4 bearbeitet. Ist keine nächste Gruppe im anonymen Datenbestand vorhanden, so gehe zu Schritt 7.
7. Wenn alle Sätze zugeordnet sind (spätestens bei Durchlauf mit i = 0), dann gehe zu Schritt 8. Wenn noch nicht alle Sätze zugeordnet sind, so reduziere die zum Vergleich einzubeziehenden Merkmale um 1 (i = i – 1) und beginne wieder mit Schritt 3.
8. Exportiere alle gespeicherten Satzpaare in dem gewählten Speicherformat als Ausgabedatei.

9. Realisierung

Programmtechnischer Schwerpunkt des Verfahrens ist ein schnelles Testen der einzelnen Satzgruppen mit ihren möglichen Veränderungen und deren Auswirkungen auf die Randsummentabellen. Das erfordert, dass alle Randsummentabellen verfügbar sind und zu jedem Satz der Basisdatei schnell alle zugehörigen Tabellenfelder und die aktuellen Fehler gelesen werden können. In der Aufgabenstellung ist es zwar eine klassische Datenbankanwendung, da die Abfragen sich für jeden Satz einer Satzgruppe und jede Tabelle unterscheiden, entstehen jedoch sehr viele unterschiedliche Abfragen, die mit einer Datenbankanwendung nicht performant realisiert werden können. Deshalb wurde das Programm in C geschrieben. Um die Suchalgorithmen zu beschleunigen, wurde ein spezieller Indextyp entwickelt. Dieser ähnelt einem graphischen Index, wie er für hierarchische Datenbanken verwendet wird. Er hat neben dem schnelleren Zugriff auch den Vorteil, dass er weniger Speicherplatz erfordert und somit auch größere Probleme vollständig im Hauptspeicher realisierbar sind. Das Programm wird inzwischen als 32- oder 64-Bit-Mehrprozessor-Anwendung gewartet und weiterentwickelt.

10. Sicherheit und Qualität des Verfahrens

a) Sicherheit

Die Datensicherheit lässt sich beim Verfahren SAFE klar beurteilen. SAFE ist ein datenveränderndes Verfahren. Die Sicherheit entsteht hier nicht durch das Unterdrücken oder Löschen von sensiblen Informationen, sondern durch die Unsicherheit, ob Informationen verändert wurden. Jede Kombination an Merkmalsausprägungen ist dreifach vorhanden, wobei dies durch die Veränderung von exotischen Merkmalskombinationen entstanden sein kann. Das Nichtvorhandensein von Merkmalskombinationen bedeutet nicht, dass es im Originaldatenbestand diese Kombination nicht gegeben hat. Randwerte in Tabellen können daher nicht mehr als Angriffswissen verwendet werden. Die Information, dass alle Merkmalsträger bei einer Variablen die gleiche Ausprägung aufweisen, kann auch durch die Veränderungen in SAFE entstanden sein. Ein „Datenangreifer" kann daher nicht mit Sicherheit ableiten, dass alle Merkmalsträger die ausgewiesene Ausprägung gemeldet haben.

Das Verfahren hat den Vorteil, dass der Datenbestand nur einmal anonymisiert werden muss und dann alle Auswertungen aus diesem anonymen Datenbestand erzeugt werden können. Da der anonyme Datenbestand eine 3-anonyme Datei erzeugt hat, sind alle Auswertungstabellen sicher. Eine manuelle Prüfung jeder einzelnen Auswertungstabelle auf die Einhaltung des Statistikgeheimnisses entfällt. Kleine Fallzahlen werden in keiner Tabelle mehr ausgewiesen.

b) Qualität bei Anwendung als Tabellengeheimhaltungsverfahren

Das SAFE-Verfahren erzeugt bei jedem Durchlauf einige Kennzahlen zur Qualität des Anonymisierungslaufs. Mit diesen Kennzahlen kann die Qualität des Tabellengeheimhaltungsverfahrens beurteilt werden. Es können auch verschiedene Varianten, die sich beispielweise durch unterschiedliche Kontrolltabellensets auszeichnen, verglichen werden. Es werden Häufigkeitstabellen erstellt, wie oft welche absoluten Abweichungen sowohl in ein- als auch in zwei- und höherdimensionalen Tabellenfeldern je nach Größenklasse auftreten. In diesen Dateien kann der größte Fehler je Größenklasse abgelesen werden. Die Häufigkeitstabellen können als Ausgangspunkt genommen werden, um durchschnittliche Abweichungen in eindimensionalen sowie in zwei- und höherdimensionalen Tabellenfeldern zu berechnen. Auch können kumulierte Angaben gemacht werden, im Stil von „in 90 % aller Tabellenfelder ist die Abweichung kleiner oder gleich 3". Qualitätskennzahlen in dieser Art werden zu den Testrechnungen mit den Einzeldaten der Volkszählung 1987 in Westdeutschland in Höhne (2011) berichtet.

Diese Qualitätsangaben werden nur für Kontrolltabellen berechnet. Nicht-Kontrolltabellen können, wenn die Merkmale im Anonymisierungslauf enthalten waren, aus dem anonymen Datenbestand auch flexibel erzeugt werden und sind ebenfalls sicher. Eine Aussage über deren Qualität ist ex ante aber nicht möglich.

Die Interpretation von mit datenverändernden Verfahren geheim gehaltenen Tabellen unterscheidet sich von anderen Tabellenergebnissen. Der Informationsverlust entsteht, wie oben bereits beschrieben, nicht durch die Unterdrückung von Informationen, sondern durch Unsicherheit in den Informationen aufgrund der Veränderung. Dabei sind Veränderungen um ±2 nötig. Betrachtet man die relativen Veränderungen von Tabellenfeldern, so sind diese bei kleinen Tabellenfeldern besonders groß. Gerade bei den kleinen Tabellenfeldern, hier insbesondere die mit den Häufigkeiten 1 und 2, besteht aber der Schutzbedarf.

Das Verfahren löst aufgrund der Mikroaggregation das Problem, dass keine zu kleinen Fallzahlen mehr auftreten werden. Das Randsummenproblem ist aufgrund der Unsicherheit bei der Interpretation auch nicht mehr vorhanden.

11. Anwendung bei Dateien mit verschiedenen statistischen Einheiten

In einigen Bundesstatistiken sind Angaben zu verschiedenen statistischen Einheiten gemeinsam in einem Mikrodatenbestand enthalten oder die Angaben aus zwei Erhebungen oder Satzarten können verknüpft und kombiniert ausgewertet werden. Solche Datenbestände mit einer hierarchischen Struktur können mit dem SAFE-Verfahren ebenfalls geheim gehalten werden. Dabei sind jedoch einige Aspekte zu beachten. Diese sollen beispielhaft anhand einer fiktiven Pflegestatistik beschrieben werden.

In der Pflegestatistik sind u. a. Angaben zu den Einrichtungen und den Pflegebedürftigen enthalten. Die Einzeldaten beider statistischen Einheiten sollen zu einem Datensatz verknüpft werden, sodass alle Merkmale als Eigenschaften der kleinsten Ebene (pflegebedürftige Personen) gespeichert werden. Bei der Pflegestatistik sollen die Angaben zu den Einrichtungen mit den Angaben der Pflegebedürftigen über die Einrichtungsnummer verknüpft werden. In jeder Einrichtung werden dabei mehrere Pflegebedürftige betreut. Der neue Datenbestand weist in jeder Zeile die Angaben zu einem Pflegebedürftigen auf. Deshalb kann man die Pflegebedürftigen als die führende statistische Einheit in diesem neuen Datenbestand bezeichnen. Die Angaben jeder Pflegeeinrichtung sind nun mehrfach im Datenbestand enthalten, da sie bei jedem Pflegebedürftigen dieser Einrichtung angespielt wurden. Bei Auswertungen der Pflegeeinrichtungen kann dieser

4 | Beispieldatensatz mit verschiedenen statistischen Einheiten

Pflegeeinrichtung					Pflegebedürftige		
ID_Inst	Plätze	An-gestellte	Träger	Zähler_ Einr	ID_ Person	Alter	Pflege-stufe
25	210	120	2	1	233	70	2
25	210	120	2	0	234	85	3
25	210	120	2	0	236	84	2
27	67	35	3	1	238	84	3
27	67	35	3	0	239	81	3

Datenbestand nicht direkt ausgewertet werden, da Mehrfachzählungen der Einrichtungen auftreten würden. Vielmehr muss eine Filtervariable (Zähler) eingefügt werden, die gewährleistet, dass jede Einrichtung nur genau einmal gezählt wird. Bei allen Auswertungen nach Pflegeeinrichtungen wird die Auswertung mit dieser Filtervariable gekreuzt. Sollen die verschiedenen Träger ausgewertet werden, so muss die Häufigkeitstabelle Träger mit der Filterbedingung Zähler_Einr = 1 berechnet werden.

Wendet man das SAFE-Verfahren auf diesen Datenbestand mit verschiedenen statistischen Einheiten an, so entsteht wiederum ein 3-anonymer Datenbestand, aus dem alle Auswertungen geheim sind. Jede Merkmalskombination ist entweder mindestens dreifach oder nicht vorhanden. Dies gilt sowohl für Auswertungen der statistischen Einheit „Pflegebedürftige" als auch der Einheit „Pflegeeinrichtungen", wobei diese auch beim Auswerten des anonymen Bestandes stets mit dem Zähler-Feld kombiniert sein müssen.

Wird bei der Anonymisierung gewährleistet, dass die Kreuzkombinationen mit dem Einrichtungszähl-Feld (Zähler_Einr) als Kontrolltabellen berücksichtigt werden, so gelten die Qualitätsaussagen sowohl für die Auswertungen nach Pflegebedürftigen als auch nach Einrichtungen.

Die Merkmale „Anzahl der Plätze" und „Anzahl der Angestellten" könnten für die Geheimhaltung auch in Größenklassen klassiert werden.

12. Zusammenfassung

Mit SAFE steht ein Verfahren bereit, das einerseits Einzeldaten so anonymisiert, dass man einen anonymen Datenbestand freigeben kann, andererseits ist es ein pre-tabulares Geheimhaltungsverfahren, sodass keine Geheimhaltungsfälle in Auswertungstabellen mehr enthalten sind, die aus dem anonymen Datenbestand erzeugt werden. Das Verfahren optimiert die Lösung für ein vorgegebenes Set an Kontrolltabellen, für die das Verfahren direkt Qualitätsmaße angibt. Gleichzeitig bleibt die flexible Auswertbarkeit der Einzeldaten gewährleistet.

Dr. Jörg Höhne leitet die Abteilung *Gesamtwirtschaft* im Amt für Statistik Berlin-Brandenburg. Er studierte Statistik und Wirtschaftsmathematik in Berlin und Moskau und promovierte 2009 an der Universität Tübingen mit einer Arbeit über „Verfahren zur Anonymisierung von Einzeldaten".

Literatur

- Appel, Günther; Kinzel, Sabine; Nölte, Dieter (1993): SAFE – A Generally Usable Program System for the Anonymization of Individual Data in Official Statistics. In: Proceedings of the International Seminar on Statistical Confidentiality, Dublin, Ireland, 8–10 September 1992, S. 201–228.
- Höhne, Jörg (2003): SAFE – Ein Verfahren zur Anonymisierung und statistischen Geheimhaltung wirtschaftsstatistischer Einzelangaben. In: Berliner Statistik, Statistische Monatszeitschrift, Nr. 3/2003, Berlin 2003, S. 96–107.
- Höhne, Jörg (2008): Anonymisierungsverfahren für Paneldaten. In: Springer, Wirtschafts- und Sozialstatistisches Archiv Band 2/2008, S. 259–275.
- Höhne, Jörg (2010): Verfahren zur Anonymisierung von Einzeldaten. Statistik und Wissenschaft Band 16, Statistisches Bundesamt, Wiesbaden.
- Höhne, Jörg (2011): SAFE – A method for anonymising the German Census. Working Paper 16 at the Joint UNECE/ Eurostat work session on statistical data confidentiality, 26–28 October 2011, Tarragona. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ ge.46/2011/16_Germany.pdf
- Höhne, Jörg (2012): Statistische Geheimhaltung des Zensus 2011. Vortrag im Rahmen der Statistik-Tage Bamberg Fürth 2012, „Die Methoden und Potenziale des Zensus 2011" am 26. und 27. Juli 2012. https://www.statistik.bayern.de/medien/wichtigethemen/st_vortrag_hoehne_27072012.pdf
- Hundepool, Anco; Domingo-Ferrer, Josep; Franconi, Luisa; Giessing, Sarah; Schulte Nordholt, Eric; Spicer, Keith; de Wolf, Peter-Paul (2012): Statistical Disclosure Control. Wiley Series in survey methodology, John Wiley & Sons Ltd, Chichester. Überarbeitete Fassung des ESSNET-SDC Handbook. http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf
- Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (Hrsg.): Wege zu einer besseren informationellen Infrastruktur. Gutachten der vom Bundesministerium für Bildung und Forschung eingesetzten Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik Baden-Baden 2001.
- Lechner, Sandra; Pohlmeier, Wilfried (2003): Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten. In: Gnoss, Roland; Ronning, Gert (Hrsg.): Anonymisierung wirtschaftsstatistischer Einzeldaten, Schriftenreihe „Forum der Bundesstatistik", Band 42, S. 115–137, hrsg. vom Statistischen Bundesamt, Wiesbaden.
- Lenz, Rainer (2010): Methoden der Geheimhaltung wirtschaftsstatistischer Einzeldaten und ihre Schutzwirkung. Statistik und Wissenschaft Band 18, Statistisches Bundesamt, Wiesbaden.
- Ronning, Gerd; Sturm, Ronald; Höhne, Jörg; Lenz, Rainer; Rosemann, Martin; Scheffler, Michael; Vorgrimler, Daniel (2005): Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten. In: Schriftenreihe „Statistik und Wissenschaft", Band 4, hrsg. vom Statistischen Bundesamt, Wiesbaden.
- Sweeney, Latanya (2002): k-anonymity – a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; S. 557–570.
- Zühlke, Sylvia; Christians, Helga; Cramer, Katharina (2007): Das Forschungsdatenzentrum der Statistischen Landesämter – eine Serviceeinrichtung für die Wissenschaft. AStA Wirtschafts- und Sozialstatistisches Archiv 3-4, S. 169–178.

Geheimhaltung

Mindestfallzahlregel versus Randwertregel

– eine Betrachtung der Enthüllungsrisiken

von **Julia Höninger**

Bei der statistischen Geheimhaltung wird üblicherweise anhand der Mindestfallzahlregel und der Dominanzregel geprüft, ob Einzelangaben bei Veröffentlichungen der amtlichen Statistik ausreichend geschützt sind. Vor allem bei Häufigkeitstabellen sollte jedoch das Enthüllungsrisiko durch die Randwertproblematik stärker beachtet werden. Das Enthüllungsrisiko wird anhand von Beispielen aufgezeigt. Es wird begründet, warum auf die Sperrung kleiner Häufigkeiten verzichtet werden könnte. Des Weiteren wird auch zum Randwertproblem bei Wertetabellen Bezug genommen.

Seit dem Hippokratischen Eid ca. 400 Jahre vor Christus ist der Schutz von persönlichen Daten ein bekanntes Konzept und eine besondere Verantwortung für Personen, die aufgrund ihrer sozialen Rolle oder ihres Berufes Kenntnis solcher Daten erlangen. Die Regelung, dass Informationen schützenswert sind, ist für die amtliche Statistik ebenfalls von zentraler Bedeutung: Das Bundesstatistikgesetz[1] verpflichtet die amtliche Statistik ausdrücklich (§ 16) zum Schutz von Einzelangaben. Die wohl bekannteste Geheimhaltungsregel zur Sicherung des Statistikgeheimnisses in diesem Zusammenhang ist die Mindestfallzahlregel. Sie besagt, dass in Häufigkeitstabellen nur Felder veröffentlicht werden dürfen, wenn eine Mindestfallzahl *n* nicht unterschritten wird. Auf Wertetabellen angewandt, dürfen nur solche Tabellenfelder veröffentlicht werden, zu denen mindestens *n* Befragte[2], Betroffene oder Beobachtungseinheiten beitragen. In der amtlichen Statistik wird die Mindestfallzahlregel üblicherweise mit $n \geq 3$ angewandt, zu jedem zu veröffentlichenden Tabellenfeld müssen mindestens drei Beobachtungseinheiten beitragen. Die Mindestfallzahlregel leitet sich aus dem Ausnahmetatbestand in § 16 Abs. 1 Satz 2 Nr. 3 BStatG ab. Dieser besagt, dass Einzelangaben dann veröffentlicht werden dürfen, wenn sie „mit den Einzelangaben anderer Befragter zusammengefasst [wurden] und in statistischen Ergebnissen dargestellt sind".

Die Mindestfallzahlregel ist gut verständlich und bei der Erstellung von Veröffentlichungen leicht anwendbar. Sie wird von der einschlägigen Geheimhaltungsliteratur stets als erste Geheimhaltungsregel benannt.[3] Nach Giessing (1999) verhindert die Mindestfallzahlregel die exakte Offenlegung von Einzelangaben.

Die nächst bekannteren Geheimhaltungsregeln gehören der Gruppe der Dominanz- oder Konzentrationsregeln an, hier gibt es u. a. die (1,k)-, (2,k)-, p%- und (p,q)-Regel. Diese sind nur bei metrischen bzw. quantitativen Variablen anwendbar, sie verhindern die näherungsweise Aufdeckung eines Einzelwertes (Giessing 1999).

Bei Regeln, die die exakte Offenlegung von Einzelwerten verhindern, sollte jedoch neben der Mindestfallzahlregel auch die Randwertregel aufgeführt werden. Ziel der statistischen Geheimhaltung ist es stets, Einzelangaben ausreichend zu schützen. Die Mindestfallzahlregel sperrt jedoch Tabellenfelder, die gar kein Enthüllungsrisiko darstellen. Insofern wird dieses Ziel mit der Mindestfallzahlregel nicht erreicht, sondern es wäre wichtiger, die Randwertregel anzuwenden. Da trotz statistischer Geheimhaltung stets versucht wird, einen möglichst großen Anteil des Informationsgehalts der statistischen Veröffentlichungen zu erhalten, sollte bei bestimmten Konstellationen bei Häufigkeitstabellen auf die Sperrung von gering besetzten Zellen verzichtet werden. Bei Wertetabellen muss jedoch stets auf die Einhaltung der Mindestfallzahl geachtet werden.

1 Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 13 des Gesetzes vom 25. Juli 2013 (BGBl. I S. 2749).

2 Im Weiteren wird der besseren Lesbarkeit wegen die männliche Form verwendet. Gemeint sind stets beide Geschlechter.

3 So beispielsweise in
- Giessing 1999, S. 6 (Fallzahlregeln),
- Poppenhäger 1995, S. 57 (Dreier-Aggregation),
- Dorer, Mainusch und Tobies 1988, S. 96 (Zusammenfassung),
- Hundepool et al. 2010, S. 119 f. (Mindestfallzahlregel = minimum frequency rule),
- Brandt et al. 2009, S. 8 (threshold rule),
- Duncan, Elliot und Salazar 2011, S. 65 (small counts in tables).

Zu diesem Schluss kam im Jahr 2000 auch eine Arbeitsgruppe des Statistischen Bundesamtes: „Die derzeit angewandten Regeln können zu nicht ausreichenden Sperrungen, in anderen Fällen zu überflüssigen Sperrungen führen. Zielsetzung einer Neugestaltung der bestehenden Geheimhaltungsregeln sollte eine Verbesserung der Qualität der Geheimhaltung bei gleichzeitiger Reduzierung der Informationsverluste für die Konsumenten sein". (Gnoss 2000)

Die Randwertproblematik

Zunächst sollen die Randwertproblematik erläutert sowie die Regel genannt werden, mit der Randwertprobleme identifiziert werden können. Danach folgen einige Tabellenbeispiele zu Randwerten und kleinen Fallzahlen.

Wenn die Zellbesetzung eines Tabellenfeldes der Fallzahl der Randsumme entspricht, liegt ein Randwertproblem vor. In Tabellen können Randwerte innerhalb einer Zeile oder einer Spalte auftreten. Manchmal werden statt des Begriffs Randwertregel auch die Begriffe Randsummenregel oder Randsummenkriterium verwendet. Mit der Randwertregel wird geprüft, ob ein Randwertproblem vorliegt, also ob in einem Innenfeld eine Totalbesetzung auftritt. Haben alle Beobachtungseinheiten einer Untergruppe bei einem Merkmal die gleiche Merkmalsausprägung, so ist für jeden Einzelnen enthüllt, welche Ausprägung er in diesem Merkmal hat. Für jeden Merkmalsträger der Untergruppe kann ein Attribut, also eine Ausprägung eines Merkmals, mit Sicherheit enthüllt werden. Sobald man die Merkmalsträger der Gruppe zuordnen lassen, kann für jeden Einzelnen eine Einzelangabe benannt werden, denn „was für alle gilt, gilt auch für den einen". Dabei ist es teilweise nicht notwendig, die genaue Identität zu kennen. Sind beispielsweise alle Einwohner einer Gemeinde evangelisch, so reicht die Information, dass eine Person in jener Gemeinde wohnt, aus, um deren Konfession bestimmen zu können. Eine Unterscheidung in sensible und unsensible Variablen kann es dabei nicht geben, da das Bundesstatistikgesetz keine solche Einteilung vorsieht. Die gültigen Geheimhaltungsregeln müssen grundsätzlich auf alle Merkmale angewendet werden. Alle für eine Bundesstatistik gemachten Einzelangaben sind geheim zu halten.

Randwertprobleme sind bezüglich der Geheimhaltung immer auch inhaltlich zu beurteilen. Einige Randwerte sind inhaltlich aufgrund von logischen Bedingungen nur so möglich, daher müssen sie nicht gesperrt werden (z. B. nicht erwerbstätige Kinder).

Wird als Geheimhaltungsverfahren die Zellsperrung verwendet, so ist zum Schutz neben der mit dem Randwert besetzten Ausprägung im Tabelleninnenfeld auch mindestens eine weitere mögliche Ausprägung zu sperren, um eine Unsicherheit zu erzeugen. Damit ist die Möglichkeit verhindert, aus der Summe auf den Einzelwert zurückzuschließen (Rückrechenbarkeit). Das Verhindern der Rückrechenbarkeit wird als sekundäre Geheimhaltung bezeichnet und ist bei primär gesperrten Feldern immer zu überprüfen. Um die größte Unsicherheit zu erzeugen, kann die ganze Zeile oder Spalte der Tabelle gesperrt werden, in der der Randwert auftritt. Die Randsumme kann dann veröffentlicht werden.

In einer strengeren Version der Randwertregel ist zu sperren, wenn alle Merkmalsträger außer einem die gleiche Ausprägung aufweisen. Anders formuliert, ist zu sperren, wenn ein Tabelleninnenfeld die Häufigkeit des Randfeldes minus eine Beobachtungseinheit enthält. Dann kann die Person oder das Unternehmen, das den Einzelfall darstellt, die Ausprägung für alle anderen Beobachtungseinheiten enthüllen.

Statistik erklärt – **Konzentrations-/Dominanzregeln**

Mindestfallzahl- und Randwertregel finden Anwendung insbesondere in Häufigkeitstabellen. Liegen Tabellenwerten Angaben von mehr als zwei Einheiten (Personen/Unternehmen) zugrunde und haben eine oder zwei der Einheiten einen großen Anteil am Ergebnis, liegt eine Dominanz vor. In diesem Fall kommt in der amtlichen Statistik eine Konzentrationsregel zur Anwendung. Die Dominanzproblematik kann nur bei metrischen Merkmalen (z. B. Einkommen, Umsatz, Investitionen) auftreten, bei denen beispielsweise eine Summe über alle Ausprägungen errechnet werden kann.

(1,k)-Dominanzregel: Ein Tabellenfeld ist primär geheim zu halten, wenn der Anteil des größten Einzelwertes mehr als k % beträgt.

(2,k)-Dominanzregel: Ein Tabellenfeld ist primär geheim zu halten, wenn der Anteil der beiden größten Einzelwerte mehr als k % beträgt.

p%-Regel: Ein Tabellenfeld ist primär geheim zu halten, wenn die Differenz zwischen dem Tabellenwert und dem zweitgrößten Einzelwert den größten Einzelwert um weniger als p% übersteigt. Je stärker der größte Einzelwert geschützt werden soll, umso höhere Werte müssen für p angesetzt werden. Der letztendlich eingesetzte Parameter unterliegt ebenfalls der Geheimhaltung.

In der amtlichen Statistik werden die letzten beiden Konzentrationsregeln zur Wahrung der statistischen Geheimhaltung angewendet, denn (2,k)-Dominanzregel und p%-Regel stellen sicher, dass Brancheninsider mit dem Vorwissen über den Wert des zweitgrößten Einzelbeitrags (d. h. insbesondere das Unternehmen mit dem zweitgrößten Einzelbeitrag selbst) die Angabe des Unternehmens mit dem größten Einzelbeitrag um mindestens (100-k)% des Zellwerts bzw. p% des größten Einzelbeitrags überschätzen.

Quelle: Statistische Ämter des Bundes und der Länder: Handbuch zur statistischen Geheimhaltung, Stand 05. 05. 2014, S. 29 ff., internes Dokument.

Aus methodischer Sicht entsteht das höhere Enthüllungsrisiko bei Häufigkeitstabellen nicht bei der Veröffentlichung kleiner Fallzahlen, sondern bei der Veröffentlichung von Randwertkonstellationen, denn bei einer Veröffentlichung von Randwerten ist eine Information über alle betroffenen Beobachtungseinheiten offen gelegt. Die Einhaltung der Mindestfallzahl bei Häufigkeitstabellen erlaubt meist keinen zusätzlichen Erkenntnisgewinn. Es können höchstens Beobachtungseinheiten zugeordnet werden, für die bereits alle Merkmale bekannt sind.

Beispiele zur Randwertproblematik

Der Bereich der Gesundheitsstatistiken wendet bereits seit vielen Jahren die Randwertregel an und verzichtet gleichzeitig auf die Mindestfallzahlregel. Als Geheimhaltungsverfahren werden in diesem Bereich die Vergröberung und die Zellsperrung angewandt. Bei der Todesursachenstatistik wird das Statistikgeheimnis dadurch gewahrt, dass kleine Fallzahlen sehr wohl veröffentlicht werden, jedoch jeweils überprüft wird, dass keine Randwerte vorliegen. Hier wird beispielsweise kontrolliert, dass die Verstorbenen einer Altersgruppe, eines Geschlechts und in einer gegebenen regionalen Einheit an mehreren unterschiedlichen Todesursachen gestorben sind. Ein „Datenangreifer“ würde eine Einzelangabe dann enthüllen können, wenn alle Personen der Gruppe an der gleichen Ursache gestorben sind. Denn was für die gesamte Gruppe gilt, gilt auch für den Einzelnen. Das nachfolgende fiktive Beispiel ist analog zu Tabellen aus der Todesursachenstatistik (Statistisches Bundesamt 2000, S. 27) aufgebaut.

Tabelle 1a enthält Beispielangaben von verschiedenen Todesursachen A bis D nach Alter in Jahren der Verstorbenen. In dieser Tabelle sind sowohl kleine Fallzahlen (1 und 2) als auch ein Randwertproblem enthalten.

In Tabelle 1b wurde die Mindestfallzahlregel angewendet und das Beispiel aus Tabelle 1a mit Zellsperrung geschützt. Es wurden Primär- und Sekundärsperrungen gesetzt, die mit einem „·“ gekennzeichnet sind. Dennoch kann man aus dieser Tabelle ablesen, dass alle Personen der Altersgruppe 0 bis 19 Jahre an der Todesursache B gestorben sind.

In Tabelle 1c hingegen wurde das Randwertproblem aus Tabelle 1a mit Zellsperrung inklusive Sekundärsperrung geschützt. Die kleinen Fallzahlen wurden hier nicht gesperrt, es wurde keine Mindestfallzahlregel angewandt. Aus dieser Tabelle ist keine Einzelangabe mehr einem Betroffenen zuzuordnen; die Einzelangaben sind hier besser (ausreichend!) geschützt.

Anhand von Beispielen aus anderen Statistikbereichen kann ebenfalls aufgezeigt werden, wie durch Randwerte ein Enthüllungsrisiko entstehen kann. Die fiktiven Tabellenbesetzungen von Tabelle 1a könnten auch bei der Auswertung einer anderen Personenstatistik entstehen, wenn beispielsweise der Familienstand je Altersgruppe tabelliert wird (wobei die Altersgruppen anders gewählt werden).

Aus der fiktiven Tabelle 2 kann herausgelesen werden, dass alle Männer zwischen 30 und 40 Jahren den Familienstand B haben. Wenn der Familienstand B für verheiratet steht, dann wäre kein Mann dieser Altersgruppe ledig, verwitwet oder geschieden (Annahme für dieses Beispiel: es gibt nur diese vier Familienstände). Für alle Männer der Altersgruppe

1a | Originaltabelle: Todesursachen nach Alter in Jahren

Todesursache	Alter in Jahren					
	0 bis 19	20 bis 39	40 bis 59	60 bis 79	80 und älter	Insgesamt
A	–	8	15	24	22	69
B	10	5	5	3	5	28
C	–	1	2	7	1	11
D	–	20	30	25	14	89
Insgesamt	10	34	52	59	42	197

1b | Sperrung nach Mindestfallzahlregel mit Sekundärsperrung

Todesursache	Alter in Jahren					
	0 bis 19	20 bis 39	40 bis 59	60 bis 79	80 und älter	Insgesamt
A	–	8	15	24	22	69
B	10	·	·	3	·	28
C	–	·	·	7	·	11
D	–	20	30	25	14	89
Insgesamt	10	34	52	59	42	197

1c | Sperrung nach Randwertregel mit Sekundärsperrung

Todesursache	Alter in Jahren					
	0 bis 19	20 bis 39	40 bis 59	60 bis 79	80 und älter	Insgesamt
A	–	8	15	24	22	69
B	·	·	5	3	5	28
C	·	·	2	7	1	11
D	–	20	30	25	14	89
Insgesamt	10	34	52	59	42	197

2 | Originaltabelle: Familienstand von Männern nach Altersgruppen in Jahren

Familienstand	Alter in Jahren					
	30 bis 39	40 bis 49	50 bis 59	60 bis 69	70 und älter	Insgesamt
A	–	8	15	24	22	69
B	10	5	5	3	5	28
C	–	1	2	7	1	11
D	–	20	30	25	14	89
Insgesamt	10	34	52	59	42	197

3 | Religionszugehörigkeit der Einwohner einer Gemeinde

	Insgesamt	davon			
		römisch-katholisch	evangelisch	andere Religion	keine Religion
Einwohner.....	500	–	500	–	–

Quelle: Statistisches Bundesamt 2000, S. 27, eigene Bearbeitung

wäre enthüllt, dass sie verheiratet sind. Die Angabe des Familienstandes wurde aber für eine Bundesstatistik gemacht und ist damit geheim zu halten. Problematisch wird es, wenn ein Mann dieser Altersgruppe Freunden, Kollegen oder Bekannten einen anderen Familienstand genannt hat und diese nun die Veröffentlichung des Statistikamtes sehen.

Die Häufigkeit von eins in einem Tabellenfeld (sogenannte Tabelleneins) in der Altersgruppe der über 70-Jährigen hingegen stellt kein Enthüllungsrisiko dar. Über keinen Auskunftgebenden dieser Altersgruppe kann anhand dieser Tabelle etwas dazugelernt werden. Die Merkmalskombination 70 Jahre und älter und Familienstand C ist zwar einzigartig, aber aus der Tabelle selbst kann keinem Beitragenden diese Einzelangabe zugeordnet werden. Ein Familienstand kann keinem der 42 Männer dieser Altersgruppe zugeordnet werden. Als einziges Risiko kann die Einzigartigkeit als „Angriffswissen" bei anderen Publikationen verwendet werden. Diese Logik kann als Grund für Sperrungen benannt werden, aber aus dieser Tabelle selbst entsteht bei der Altersgruppe der über 70-Jährigen kein Enthüllungsrisiko.

Die gleichen Schlussfolgerungen bezüglich des Enthüllungsrisikos würden gelten, wenn in der Tabellenvorspalte statt des Familienstands andere Merkmale tabelliert wären, z. B. Wirtschaftszweig des Betriebs, Stellung im Beruf, Anzahl der Kinder oder Einkommenskategorien.

Ein anderes Beispiel ist eine Auswertung, welcher Religionsgemeinschaft die Einwohner einer Gemeinde angehören. Wenn alle Einwohner wie in Tabelle 3 die Ausprägung evangelisch aufweisen, kein Einwohner römisch-katholisch ist, einer anderen Religion oder keiner Religion angehört, liegt ebenfalls ein Randwert vor. Nach der Mindestfallzahlregel dürfte diese Tabelle veröffentlicht werden, wenn die Gemeinde beispielsweise insgesamt 500 Einwohner umfasst. Allerdings ist für alle Einwohner in diesem Fall die Religionszugehörigkeit enthüllt.

Auch wenn die genannten Beispiele auf den ersten Blick trivial klingen, kann meist ein Szenario konstruiert werden, warum sich ein Betroffener beschweren könnte, dass seine Einzelangabe enthüllt werden kann. Ist beispielsweise ein Mann aus der römisch-katholischen Nachbargemeinde zu seiner evangelischen Ehefrau gezogen und heimlich ebenfalls der evangelischen Kirche beigetreten, so können dies nun alle Verwandten aus der römisch-katholischen Nachbargemeinde aus der Tabelle ablesen. Alle für eine Bundesstatistik gemachten Angaben sind jedoch zu schützen. Wenn der Mann seine Konfession bei einer Bundesstatistik angegeben hat, muss diese Einzelangabe von der amtlichen Statistik geheim gehalten werden.

Bekanntheit der Randwertregel

Die Behandlung der Enthüllungsrisiken durch Randwerte ist in der Geheimhaltungs-[4] und juristischen[5] Literatur sehr unterschiedlich. In der amtlichen Statistik in Deutschland ist die Randwertregel bisher wenig verbreitet. In einem Leitfaden zur Organisation von Arbeitsabläufen und Programmen zur statistischen Geheimhaltung (Statistisches Bundesamt 2008, S. 16) werden beispielsweise alle Geheimhaltungsregeln aufgezählt und Marker für die verschiedenen Geheimhaltungsgründe vorgestellt, Marker für die Randwertregel sind aber nicht vorgesehen.

Mindestfallzahlregel ade: Wann dürfen kleine Fallzahlen veröffentlicht werden?

Poppenhäger (1995, S. 58) kommt zu dem Schluss, dass durch die systematische Stellung des § 16 Abs. 4 Satz 1 BStatG[6] argumentum e contrario eine Weitergabe kleiner Fallzahlen an andere Empfänger nicht erlaubt sein kann. Des Weiteren wird auch die Mindestfallzahlregel mit mindestens drei Befragten von Poppenhäger (1995) aus dem Wortlaut des Gesetzes abgeleitet. Wenn Einzelangaben „mit den Einzelangaben anderer Befragter zusammengefasst und in statistischen Ergebnissen dargestellt sind", so dürfen sie nach § 16 Abs. 1 Satz 3 BStatG veröffentlicht werden. Aufgrund des Plurals bei der Formulierung „anderer Befragter" schlussfolgert Poppenhäger, dass die Angaben von einem Befragten mit den Angaben von mindestens zwei weiteren Befragten zusammengefasst und somit mindestens drei Merkmalsträger zu einem Tabellenfeld beitragen müssen.

Allerdings ist hier ein Paradigmenwechsel notwendig. Die Einzelangaben der Befragten sind nach dem Gesetz geheim zu halten und müssen daher von den Statistischen Ämtern geschützt werden, d. h. Tabellen müssen geprüft werden, ob mit ihrer Hilfe auf Angaben Einzelner zurückgeschlossen werden kann. So lange die gesamte Häufigkeitstabelle bzw. die Tabellenspalte oder -zeile die Angaben von mehreren Befragten zusammenfasst, ist kein Rückschluss mehr auf die Einzelangaben möglich. Geringe Häufigkeiten von 1 oder 2 stellen kein Enthüllungsrisiko dar und müssen nicht gesperrt

4 In Hundepool et al. (2010), einem umfassenden europäischen Handbuch zur Geheimhaltung von statistischen Tabellenergebnissen und zur Anonymisierung von Mikrodaten, wird die Randwertregel in Kapitel 4.2 Wertetabellen (S. 117 ff.) und in der dort abgebildeten Übersicht der „sensitivity rules" (Geheimhaltungsregeln) nicht erwähnt. In Kapitel 5 Häufigkeitstabellen wird unter den verschiedenen Enthüllungsrisiken jedoch das Problem der „group disclosure" (S. 168) erläutert.
In Brandt et al. (2009, S. 9), einem Leitfaden zur Prüfung der statistischen Geheimhaltung für Forschungsdatenzentren, der in einem europäischen Forschungsprojekt entstanden ist, wird die Randwertregel (group disclosure) neben der Fallzahlregel und den Dominanzregeln gleichberechtigt benannt und es wird für ihre Anwendung plädiert.
Duncan, Elliot und Salazar (2011) umreißen in Kapitel 4 kurz alle modernen Möglichkeiten, wie Tabellen pre- oder posttabular geschützt werden können. Die Randwertregel wird nicht erwähnt. Mindestfallzahlregel und die Dominanzregeln werden inklusive Beispiel kurz dargestellt.

5 Dorer, Mainusch und Tobies (1988), erläutern in ihrem Kommentar zum Bundesstatistikgesetz, dass Zusammenfassungen von Einzelangaben nötig sind, damit diese Aggregate als statistische Ergebnisse veröffentlicht werden dürfen; auch Dominanzen erläutern sie inhaltlich ohne Formeln. Das Randwertproblem wird nicht erwähnt.
Der aktuellere juristische Text von Poppenhäger (1995) behandelt die Ausnahmetatbestände von § 16 Abs. 1 BStatG detaillierter. Er nennt explizit eine Mindestfallzahl von drei („Drei-er-Aggregation", S. 57). Randwertprobleme werden nur in der Fußnote 169 (S. 58) als „Totalbesetzung eines Merkmals" inklusive Beispiel als problematisch erwähnt.

6 Die Weitergabe von Tabellen mit kleinen Fallzahlen, „auch soweit Tabellenfelder nur einen einzigen Fall ausweisen", an oberste Bundes- und Landesbehörden ist nach § 16 Abs. 4 Satz 1 BStatG explizit erlaubt.

werden. Auch in den beiden Beiträgen von Smith und Elliot (2008) und Hochgürtel (2013) vertreten die Autoren die Meinung, dass mit kleinen Fallzahlen in Tabellenfeldern, zu denen nur ein oder zwei Merkmalsträger beigetragen haben, kein unmittelbares Enthüllungsrisiko verbunden ist. Aus Tabellen mit kleinen Fallzahlen kann kein Rückschluss auf den einzelnen Befragten gezogen werden, aus Tabellen mit Randwertproblematik jedoch schon. Die Einzelangaben sind bei Randwertproblemen nicht geschützt. Man sollte daher die Tabelle als die „Zusammenfassung von Einzelangaben" ansehen und darin enthaltene kleine Häufigkeiten veröffentlichen, jedoch beim Vorliegen von Randwertproblemen Sperrungen oder andere Geheimhaltungsmaßnahmen vornehmen.

4a | Anzahl der Betriebe nach Wirtschaftszweigen

Wirtschaftszweig	Anzahl der Betriebe
Insgesamt	11
davon	
45.22.1	10
45.22.2	1

4b | Beschäftigtengrößenklassen nach Wirtschaftszweigen

Wirtschaftszweig	Anzahl	davon Betriebe mit … Beschäftigten		
		1	2	3 und mehr
Insgesamt	11	5	4	3
davon				
45.22.1	10	5	3	3
45.22.2	–	–	1	–

Quelle: Statistisches Bundesamt 2000, S. 9, eigene Bearbeitung

Aus der Tabelle 4a, in der eine nach der Mindestfallzahlregel geheim zu haltende geringe Häufigkeit auftritt, ist nur eine Existenzfeststellung möglich. Es gibt einen Betrieb, der im Wirtschaftszweig (WZ) 45.22.2 tätig ist. Wer den Betrieb kennt, lernt nichts Neues, denn er wusste vor dem Lesen der Tabelle, dass es einen Betrieb gibt. Durch die Tabelle wird bekannt, dass die Merkmalskombination einzigartig ist. Dieses Wissen kann in anderen Tabellen als „Angriffswissen" verwendet werden. Aufgrund der Argumentation, dass die Befragten eine Enthüllung befürchten könnten, sollten kleine Fallzahlen gesperrt werden. Eine Enthüllung würde aber aufgrund der „anderen" Tabelle, in der man die Information der Einzigartigkeit nutzt, entstehen, nicht durch die hier dargestellte Tabelle 4a.

In Tabelle 4b kann man über den einzigen Betrieb im WZ 45.22.2 nun eine Eigenschaft ablesen, nämlich die Anzahl der Beschäftigten. Wenn der Betrieb aufgrund der Zugehörigkeit zum Wirtschaftszweig identifiziert werden kann, so kann ihm die für die Bundesstatistik gemachte Einzelangabe zugeordnet werden. Dieses Enthüllungsrisiko entsteht nicht aufgrund einer zu kleinen Häufigkeit, sondern aufgrund der Randwertproblematik. Einzigartige Merkmalskombinationen stellen für sich noch kein Enthüllungsrisiko dar – wenn sie in einer anderen höherdimensionalen Tabelle jedoch nach einem weiteren Merkmal aufgegliedert werden, entsteht in allen Fällen ein Randwertproblem.

Randwerte und kleine Fallzahlen – auf die Streuung kommt es an

Ein sehr eindrucksvolles Beispiel, dass ein Randwertproblem enthüllend ist, präsentierte Wolfgang Walla in einem Artikel aus dem Jahr 1994. In der Tabelle 5a kann jeder ablesen, welchen Berufen die Erwerbstätigen einer Gemeinde nachgehen und wie alt sie sind. In dieser fiktiven Gemeinde herrscht eine absolute berufliche Spezialisierung und alle Erwerbstätigen sind von Beruf Maurer. Da in dieser Tabelle sowohl in den Spalten (nur ein Beruf besetzt) als auch in den Zeilen (nur eine Altersgruppe besetzt) jeweils Randwerte vorliegen, kann für alle erwerbstätigen Einwohner auch das Alter exakt abgelesen und somit enthüllt werden. In diesem Extremfall einer Tabelle benötigt ein Leser kein weiteres Vorwissen, um den Beruf ablesen zu können.

Wäre die Verteilung der Erwerbstätigen eine andere als in Tabelle 5b dargestellt, so wären Enthüllungen von Einzelangaben schon erschwert, aber dennoch weiterhin möglich. Bei Kenntnis des genauen Alters kann der Beruf abgelesen werden. Und bei Vorwissen über den Beruf kann eindeutig auf das Alter des Erwerbstätigen geschlossen werden.

Damit eine exakte Zuordnung einer Merkmalsausprägung verhindert werden kann, müssen sich die Angaben über die verschiedenen Kategorien verteilen. Es muss ein Mindestmaß an Streuung vorhanden sein, also mehrere Kategorien in jeder Zeile und Spalte besetzt sein, damit das Statistikgeheimnis gewahrt wird. Die Verteilung der Erwerbstätigen in Tabelle 5b weist eine größere Streuung als in Tabelle 5a auf, ist jedoch nicht ausreichend, um die Einzelangaben zu schützen.

5a | Erwerbstätige nach Altersgruppen und ausgewählten Berufen

Ausgewählte Berufe	Erwerbstätige im Alter von … bis unter … Jahren										
	15 bis 20	20 bis 25	25 bis 30	30 bis 35	35 bis 40	40 bis 45	45 bis 50	50 bis 55	55 bis 60	60 bis 65	Insgesamt
Ingesamt	–	–	–	10	–	–	–	–	–	–	10
davon											
Bäcker	–	–	–	–	–	–	–	–	–	–	–
Flaschner	–	–	–	–	–	–	–	–	–	–	–
Förster	–	–	–	–	–	–	–	–	–	–	–
Friseur	–	–	–	–	–	–	–	–	–	–	–
Lehrer	–	–	–	–	–	–	–	–	–	–	–
Maler	–	–	–	–	–	–	–	–	–	–	–
Maurer	–	–	–	10	–	–	–	–	–	–	10
Mechaniker	–	–	–	–	–	–	–	–	–	–	–
Metzger	–	–	–	–	–	–	–	–	–	–	–
Schlosser	–	–	–	–	–	–	–	–	–	–	–

Quelle: Walla (1994, S. 104)

Randwertproblem bei Voll- und Stichprobenerhebungen

Das Randwertproblem ist vor allem bei Vollerhebungen kritisch. Bei Vollerhebungen weiß der Leser einer Veröffentlichung, in der ein Randwertproblem publiziert wird, etwas über alle Beobachtungseinheiten, die der Gruppe angehören. Bei Stichprobenerhebungen wird nur etwas über die Personen oder Unternehmen enthüllt, die an der Befragung teilgenommen haben. Wenn die Teilnahme nicht bekannt ist, herrscht Unsicherheit. Ist jedoch die Teilnahme bekannt, z. B. weil nach dem Erhebungsdesign stets alle Bewohner einer Anschrift befragt werden, weiß ein Teilnehmer auch, dass seine Nachbarn ebenfalls befragt wurden. Die Unsicherheit durch die Stichprobe entfällt.

Mindestfallzahlregel bei Wertetabellen unverzichtbar

Bei Wertetabellen muss die Mindestfallzahlregel der dahinter liegenden Beobachtungseinheiten jedoch stets beachtet werden. Beispielsweise muss vor der Veröffentlichung von Summen geprüft werden, von wie vielen Merkmalsträgern Einzelwerte addiert wurden. Wenn eine Summe auf nur einem Merkmalsträger beruht, dann würde genau die Einzelangabe des einen Merkmalsträgers veröffentlicht. Das Plädoyer, dass die Randwertregel bei der statistischen Geheimhaltung wichtiger ist als die Fallzahlregel und daher auf die Prüfung von Mindestfallzahlen verzichtet werden kann, bezieht sich ausdrücklich nur auf Fallzahltabellen.

Randwertprobleme können auch in Wertetabellen enthüllend wirken, wenn die Werte in mehrdimensionalen Tabellen dargestellt werden. In der nachfolgenden Tabelle 6b sind in jeder Altersgruppe in jedem Kreis eine ausreichend hohe Zahl an Abiturienten vorhanden, wie die Häufigkeiten in Tabelle 6a anzeigen. Dennoch kann der Tabelle 6b auch ohne eine Veröffentlichung der Tabelle 6a entnommen werden, dass alle Abiturienten in Kreis A 17 Jahre alt sind. Bisher wurde in diesem Beitrag für Fallzahltabellen im Stil der Tabelle 6a erläutert, dass auf Randwerte zu prüfen ist. Allerdings ist auch bei Wertetabellen darauf zu achten, dass keine Randwerte veröffentlicht werden.

In einem anderen Szenario wird zunächst Tabelle 6a berechnet, der Randwert identifiziert, gesperrt und durch sekundäre Geheimhaltungsmaßnahmen gesichert. Wenn nun als nächstes Tabelle 6b erstellt wird, müssen die gleichen Tabelleninnenfelder gesperrt werden, damit nicht aus der Wertetabelle die Information herausgelesen werden kann. Falls eine Wertetabelle zu einer Fallzahltabelle veröffentlicht werden soll, in der bereits ein Randwert identifiziert wurde, sind Sperrmuster immer zu übertragen. Zusätzlich müssen in der Wertetabelle sowohl alle Felder auf Dominanzen als auch die Einhaltung der Mindestfallzahlregel geprüft werden. Eine separate Überprüfung, ob die Mindestfallzahl in jedem Tabellenfeld erfüllt ist, ist bei Wertetabellen allerdings nicht nötig, wenn alle Felder mit der p%-Regel geprüft werden. Die p%-Regel umfasst

5b | Erwerbstätige nach Altersgruppen und ausgewählten Berufen

Ausgewählte Berufe	Erwerbstätige im Alter von … bis unter … Jahren										
	15 bis 20	20 bis 25	25 bis 30	30 bis 35	35 bis 40	40 bis 45	45 bis 50	50 bis 55	55 bis 60	60 bis 65	Insgesamt
Ingesamt	1	1	1	1	1	1	1	1	1	1	10
davon											
Bäcker	1	–	–	–	–	–	–	–	–	–	1
Flaschner	–	–	1	–	–	–	–	–	–	–	1
Förster	–	–	–	–	–	–	1	–	–	–	1
Friseur	–	–	–	–	1	–	–	–	–	–	1
Lehrer	–	–	–	–	–	1	–	–	–	–	1
Maler	–	1	–	–	–	–	–	–	–	–	1
Maurer	–	–	–	–	–	–	–	–	1	–	1
Mechaniker	–	–	–	1	–	–	–	–	–	–	1
Metzger	–	–	–	–	–	–	–	–	–	1	1
Schlosser	–	–	–	–	–	–	–	1	–	–	1

Quelle: Walla (1994, S. 104)

6a | Anzahl Abiturientinnen und Abiturienten nach Altersjahren und Kreisen

	Insgesamt	Davon … Jahre alt			
		17	18	19	20
Kreis A	25	25	–	–	–
Kreis B	27	10	6	6	5
Kreis C	30	5	8	9	8
Insgesamt	82	30	14	15	13

6b | Durchschnittsnoten von Abiturientinnen und Abiturienten nach Altersjahren und Kreisen

	Insgesamt	Davon … Jahre alt			
		17	18	19	20
Kreis A	*2,1*	*2,1*	–	–	–
Kreis B	*2,4*	*2,3*	*2,5*	*2,5*	*2,2*
Kreis C	*2,8*	*2,9*	*3*	*2,7*	*2,5*
Insgesamt	*2,4*	*2,3*	*2,8*	*2,6*	*2,4*

die Mindestfallzahlregel mit n≥3 gleich mit: Bei allen Tabellenfeldern, zu denen nur ein oder zwei Merkmalsträger beitragen, wird das Sicherheitsniveau p nicht erreicht und die p%-Regel weist das Feld als sensibel aus.

Julia Höninger, Diplom-Volkswirtin, leitet das Referat *Volkswirtschaftliche Gesamtrechnungen, Erwerbstätigkeit* des Amtes für Statistik Berlin-Brandenburg. Zuvor arbeitete sie in mehreren Projekten zu den Themen statistische Geheimhaltung und Mikrodatenzugang.

Zusammenfassung

In diesem Beitrag wird für einen Paradigmenwechsel plädiert. Es wurde gezeigt, dass ein Enthüllungsrisiko in Häufigkeitstabellen typischerweise durch die Randwertproblematik entsteht. Eine Geheimhaltung von kleinen Fallzahlen, also die Prüfung der Mindestfallzahl, ist in vielen Fällen jedoch unnötig und führt dazu, dass die veröffentlichten Tabellen Informationspotenzial verlieren. Daher sollte der auch von anderen Gremien bereits hervorgebrachte Vorschlag aufgegriffen und der Randwertregel eine deutlich stärkere Aufmerksamkeit zugewendet werden. Dieses Plädoyer gilt nur für Häufigkeitstabellen. Bei Wertetabellen muss stets geprüft werden, dass ausreichend viele Merkmalsträger zu einer Summe beitragen. Diese Prüfung der Mindestfallzahl ist in der Anwendung der p%-Regel aber bereits enthalten.

Literatur

- Brandt, Maurice; Franconi, Luisa; Guerke, Christopher; Hundepool, Anco; Lucarelli, Maurizio; Mol, Jan; Ritchie, Felix; Seri, Giovanni; Welpton, Richard (2009): Guidelines for the checking of output based on microdata research. ESSnet SDC. Verfügbar unter [zuletzt besucht am 16.03.2012]: http://neon.vb.cbs.nl/casc/..%5Ccasc%5CESSnet%5CGuidelinesForOutputChecking_Dec2009.pdf
- Dorer, Peter; Mainusch, Helmut; Tubies, Helga (1988): Bundesstatistikgesetz. Verlag C.H. Beck.
- Duncan, George T.; Elliot, Mark; Salazar-González, Juan-José (2011): Statistical Confidentiality - Principles and Practice. Statistics for Social and Behavioral Sciences (Series Editors: Stephen E. Fienberg, Wim J. van der Linden). Springer Science+Business Media.
- Giessing, Sarah (1999): Statistische Geheimhaltung in Tabellen. In: Statistisches Bundesamt (Hrsg.): Methoden zur Sicherung der statistischen Geheimhaltung. Forum der Bundesstatistik, Band 31, Statistisches Bundesamt, Wiesbaden, S. 6–26.
- Gnoss, Roland (2000): Vorschläge für eine Neugestaltung der Regelungen zur primären Geheimhaltung. Vortrag auf der Statistischen Woche 2000 in Nürnberg. Abstract verfügbar unter: http://www.archiv.statistik.nuernberg.de/stawo/abstracts/Gnoss01.pdf
- Hochgürtel, Tim (2013): Die Messung der Enthüllungsrisiken von Ergebnissen statistischer Analysen, HTW Saar. Institut für Diskrete Mathematik und Angewandte Statistik, Arbeitspapier Nr. 3. http://www.htw-saarland.de/forschung/struktur/forschungseinrichtungen/dmas/arbeitspapiere/die-messung-der-enthullungsrisiken-von-ergebnissen-statistischer-analysen
- Hundepool, Anco; Domingo-Ferrer, Josep; Franconi, Luisa; Giessing, Sarah; Lenz, Rainer; Naylor, Jane; Schulte Nordholt, Eric; Seri, Giovanni; de Wolf, Peter-Paul (2010): Handbook on Statistical Disclosure Control. Version 1.2. http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf
- Poppenhäger, Holger (1995): Die Übermittlung und Veröffentlichung statistischer Daten im Lichte des Rechts auf informationelle Selbstbestimmung. Schriften zum Recht des Informationsverkehrs und der Informationstechnik Band 12, Duncker & Humblot Berlin.
- Smith, Duncan; Elliot, Mark (2008): A Measure of Disclosure Risk for Tables of Count, Transactions on Data Privacy 1, S. 34–52. http://www.tdp.cat/issues/tdp.a003a08.pdf
- Statistisches Bundesamt (2000): Internes Arbeitspapier.
- Statistisches Bundesamt (2007): Entwurf – Leitfaden zur Festlegung eines p-Wertes für die p%-Regel zur Tabellengeheimhaltung. Anlage 6 zum Sachstandsbericht an den AOU vom September 2007, internes Dokument der amtlichen Statistik, Statistisches Bundesamt, Wiesbaden.
- Statistisches Bundesamt (2008): Geheimhaltung – Leitfaden zur Organisation von Arbeitsabläufen und Programmen zur Erstellung von Verbundaufbereitungen unter Berücksichtigung der statistischen Geheimhaltung. Internes Dokument, Version 2.0.
- Walla, Wolfgang (1994): Das Kreuz mit der »1«, in: Baden-Württemberg in Wort und Zahl, Heft 3/1994, S. 103–106.

Fachgespräch mit Oberregierungsrätin Sarah Giessing

„Das Ziel sind einheitliche Geheimhaltungsprozesse in den einzelnen Statistiken."

Sarah Giessing leitet das Referat *C 104 Statistische Geheimhaltung; Mathematisch-statistische Methoden für Plausibilisierung und Imputation* im Statistischen Bundesamt. Bereits seit 1996 beschäftigt sie sich schwerpunktmäßig mit Methoden zur statistischen Geheimhaltung in Tabellen. Sie leitet die Bund-Länder-Arbeitsgruppe zur statistischen Geheimhaltung.

Womit beschäftigt sich die Bund-Länder-Arbeitsgruppe Geheimhaltung?

Die Arbeitsgruppe wurde vor zehn Jahren als Unterarbeitsgruppe Geheimhaltung der AG Standardisierung der Prozesse in der amtlichen Statistik (SteP) eingerichtet. Sie sollte ein Konzept für eine verbundweit einheitliche Geheimhaltung am Beispiel der Umsatzsteuerstatistik erarbeiten. Das ist gelungen: Seit 2009 wird in der Umsatzsteuerstatistik die Geheimhaltung mit einer im Verbund abgestimmten, maschinell mit der Software τ- ARGUS durchgeführten Zellsperrung durchgeführt. Inzwischen sind auch noch einige andere Wirtschaftsstatistiken auf diesem Weg.
Seit 2011 arbeitet die Arbeitsgruppe zusammen mit einer Arbeitsgruppe der Forschungsdatenzentren an modularen Geheimhaltungsleitfäden. Die bestehen aus zunächst zwei Modulen: Zum einen erklärt ein Methodenhandbuch die Rahmenbedingen und die im Verbund praktizierten Methoden und Verfahren. Aufgenommen haben wir aber auch Ansätze aus der internationalen Fachliteratur, die in anderen Ländern eingesetzt werden oder besonders vielversprechend erschienen. Damit meine ich vor allem die datenverändernden Ansätze.
Um hier ein Beispiel zu nennen: Bevor letztlich beim Zensus 2011 die Entscheidung für den Einsatz des Verfahrens SAFE aus Ihrem Haus fiel, wurde als denkbare Alternative ein Konzept des australischen Statistikamts geprüft. Darauf aufbauend habe ich angefangen, an dem, was inzwischen als „stochastische Überlagerung mit Rundung" im Verbund diskutiert wird, zu arbeiten. Das Konzept wurde von der Arbeitsgruppe intensiv begleitet und in den Pilotprojekten Umsatzsteuer- und Beherbergungsstatistik getestet.
Aber zurück zu den Geheimhaltungsleitfäden: Die haben nämlich noch ein zweites wichtiges Modul: die statistik-spezifischen Leitfäden. Mit denen soll die Geheimhaltungspraxis der einzelnen Statistiken dargestellt und dokumentiert werden. Ein Stück weit sind sie auch als Diskussionsgrundlage gedacht. Denn, ganz klar, Ziel ist es, in den einzelnen Statistiken zu einheitlichen Prozessen zwischen den Statistischen Ämtern im Verbund zu kommen, einschließlich der Forschungsdatenzentren. Dazu muss die bisherige Praxis in einigen Statistiken auf den Prüfstand gestellt werden.

Warum wird aktuell überhaupt über andere Verfahren als die bisher praktizierte Zellsperrung diskutiert?

Ich denke, vor allem aus zwei Gründen, und beide hängen mit der Zielsetzung des einheitlichen Vorgehens im Verbund zusammen. Dabei muss man sich zuallererst einigen, was überhaupt geheim zu halten ist. Gerade wenn es um Häufigkeitstabellen geht, zeigt die Praxis, dass es gelegentlich Interpretationsunterschiede des einschlägigen Paragraphen aus dem Bundesstatistikgesetz[1] gibt. Zellsperrung als Geheimhaltungstechnik erfordert aber klare Entscheidungsregeln: Was zu sperren ist und was nicht. Geheimhaltungsverfahren, die grundsätzlich bei sämtlichen Ergebnissen eine gewisse Unsicherheit über den exakten Wert aus der Erhebung bewirken, bieten hier einen Vorteil. Denn wenn in den ausgewiesenen statistischen Ergebnissen eine Unsicherheit über den exakt erhobenen Wert besteht, kann der Nutzer daraus keine Rückschlüsse auf Einzelangaben ziehen bzw. sind solche Rückschlusse mit nicht unerheblichen Irrtumswahrscheinlichkeiten behaftet.
Ein weiteres Problem der Zellsperrung ist, dass sich die Sekundärsperrung nur dann gut maschinell umsetzen lässt, wenn die Tabellen bzw. das Tabellenprogramm feststehen. Auf Verdacht ein sehr umfangreiches Programm festzulegen, damit jede denkbare Sonderauswertung bedient werden kann, die dann vielleicht gar nicht gefragt wird, führt zu vielen Zusatzsperrungen – auch bei stark von Nutzern nachgefragten Ergebnissen. Beschränkt man sich umgekehrt auf ein Minimalprogramm, muss bei jeder Sonderauswertung oder Nutzeranfrage aufwändig

nachgearbeitet werden und auch das führt oft zu keiner befriedigenden Lösung. Die Zielsetzung, hier zu einem einheitlichen Verfahren im Verbund zu kommen, bedeutet, dass Tabellenprogramme, die die Grundlage für die Sekundärsperrung bilden, gemeinsam erarbeitet werden müssen und flexible Sonderauswertungen darüber hinaus nicht oder nur eingeschränkt möglich sind. Es ist zu befürchten, dass sich bei manchen Statistiken die Ausarbeitung gemeinsamer Tabellenprogramme sehr mühsam und konfliktträchtig gestalten wird und die gemeinsame Geheimhaltung auch in der späteren Umsetzung mit organisatorischen Herausforderungen verbunden sein wird.
Bei einer Statistik wie dem Zensus, wo neben einem sehr umfangreichen statischen Auswertungsrahmen explizit auch dynamische Auswertungsmöglichkeiten gefordert sind, ist eine vollständige und konsistente Geheimhaltung mit Zellsperrung gar nicht realisierbar.

Handelt es sich bei der datenverändernden Geheimhaltung um einen allgemeinen Trend in der amtlichen Statistik oder ist diese Form der Geheimhaltung lediglich für spezielle Statistiken geeignet?

| Da zumindest in Deutschland erst bei einer Statistik – dem Zensus 2011 – ein datenveränderndes Verfahren zur Geheimhaltung eingesetzt wird, kann man noch nicht von einem Trend sprechen. Auch Tests wurden bislang bei nur drei Statistiken durchgeführt. Rein methodisch wäre es möglich, für jede Statistik ein geeignetes Verfahren zu finden, aber auch die Zellsperrung stellt bei manchen Statistiken eine gute Alternative dar. Bei Wirtschaftsstatistiken auf Basis von Stichprobenerhebungen sind den Auswertungsmöglichkeiten ohnehin durch Stichprobendesign und Zufallsfehler gewisse Grenzen gesetzt. Ich stelle mir vor, dass es bei solchen Statistiken vielleicht auch im Verbund relativ einfach ist, sinnvolle Tabellenprogramme für einen bundesweit abgestimmten Zellsperrungsprozess festzulegen.

Inwiefern beeinflusst die Anwendung von Geheimhaltungsverfahren die Qualität bzw. den Informationsgehalt der Daten?

| Wenn durch Zellsperrung geheim gehalten wird und bestimmte Auswertungen nicht möglich sind, weil sich die Risiken, dass geheim gehaltene Tabellenfelder eventuell aufgedeckt werden könnten, anders nicht sinnvoll kontrollieren lassen, ist damit ein Informationsverlust verbunden. Ein Informationsverlust tritt natürlich auch bei den von Sekundärsperrung betroffenen Feldern auf. Tabellenfelder mit geheim zu haltenden Einzelangaben darf man dabei nicht mitzählen, denn hier hat die amtliche Statistik keine Wahl: Diese Angaben müssen geheim gehalten werden.
Bei Datenveränderung entspricht der Informationsverlust abstrakt gesehen der Unsicherheit über die Daten, die das Verfahren erzeugt. Der Datennutzer weiß, dass die Daten nicht völlig identisch mit den beobachteten Werten sind. Über die Größenordnung des Unsicherheitsintervalls werden Informationen bereitgestellt. Dieser Informationsverlust muss aber im Verhältnis zur Qualität der Erhebungsdaten beurteilt werden. Denn als Statistiker wissen wir: Auch fehlende oder mit Fehlern erhobene Daten, wie sie in der Realität nun einmal leider vorkommen, verursachen statistische Fehler und haben somit eine gewisse Unsicherheit in den Ergebnissen zur Folge. Werden beispielsweise fehlende Angaben imputiert, führt auch dies zu einer Unsicherheit in den Ergebnissen.
Von einem datenverändernden Verfahren sollte verlangt werden, dass die erzeugten Veränderungen entweder irrelevant für dargestellte Ergebnisse sind (sprich: geringe relative Abweichung bei stark aggregierten Daten) oder bei schwächer aggregierten Daten, dass statistische Strukturen auch nach der Veränderung noch klar hervortreten. Ergebnisse, auf die das nicht zutrifft, müssen für die Datennutzer erkennbar sein.

Gibt es „das eine" perfekte Geheimhaltungsverfahren?

| Sicherlich nicht. Alle Verfahren haben ihre Vor- und Nachteile. Das gilt natürlich auch für datenverändernde Verfahren. Hier werden zwei grundsätzlich verschiedene Ansätze unterschieden: die pre-tabularen Verfahren wie z. B. SAFE, die die Mikrodaten verändern, und die post-tabularen Verfahren, die die Veränderung jeweils für ein konkretes Tabellenfeld festlegen.
Wenn ein pre-tabulares Verfahren erst einmal über die Daten gelaufen ist, braucht man für die spätere Tabellenproduktion normalerweise keine speziellen Auswertungsinstrumente. Dies ist bei post-tabular arbeitenden Verfahren anders: Hier muss der Auswertungsprozess in irgendeiner Weise modifiziert sein, um das Verfahren zu integrieren. Dafür muss ein entsprechendes Werkzeug geschaffen werden und man muss dafür sorgen, dass nur Ergebnisse publiziert oder Nutzern anderweitig zugänglich gemacht werden, die mit diesem Werkzeug berechnet wurden.

Wenn andererseits bei einem pre-tabularen Verfahren eine bestimmte Datenqualität erreicht werden soll, muss das Verfahren sinnvoll konfiguriert werden, was recht aufwändig sein kann. Bei SAFE z. B. empfiehlt es sich, dem Programm sozusagen als Steuerinformation die Strukturen aller vorgesehenen Auswertungen mitzugeben. Es kann passieren, dass man da an bestimmte Grenzen stößt. Das muss im Vorfeld gut untersucht und getestet werden. Bei einem Datenvolumen wie dem des Zensus 2011 können solche Tests mit sehr langen Rechenzeiten verbunden sein. Und natürlich ist SAFE nicht für den Einsatz in der Wirtschaftsstatistik entwickelt worden – hier werden auf jeden Fall andere Verfahren benötigt.

Wo steht Deutschland mit der hier praktizierten Geheimhaltung im internationalen Vergleich?

Das ist nicht einfach zu sagen. Denn auf internationalen Konferenzen wird ja eher die Speerspitze der Forschung vorgetragen. Ob ein in diesem Rahmen diskutiertes Verfahren nur bei einer Statistik oder in großer Bandbreite eingesetzt wird, erfährt man so nicht.
Bei Zellsperrung, denke ich, sind die nationalen Verbundanwendungen mit τ-ARGUS, bei denen sehr konsequent auf tabellenübergreifende Konsistenz geachtet wird, sicher „best practice". Diese Technik wird allerdings bislang nur bei wenigen Statistiken eingesetzt.
Beim Zensus 2011 haben einige Länder, wie Deutschland auch, nicht auf Zellsperrung gesetzt. Am häufigsten wurde mit einer Form des Record Swapping[2] gearbeitet. Wie bei SAFE werden dabei veränderte Mikrodaten erzeugt. Wie stark die Daten durch so ein Verfahren verändert werden, hängt von den Details ab. Diese werden nur sehr zurückhaltend publiziert, da dadurch Zusatzwissen entstehen könnte, das womöglich in bestimmten Konstellationen die Schutzwirkung des Verfahrens vermindert.
In der Wirtschaftsstatistik weiß ich, dass das United States Census Bureau bei mindestens einer Statistik auch eine pre-tabulare stochastische Überlagerung eingesetzt hat.
Und dann gibt es eben den Ansatz der Australier für post-tabulare stochastische Überlagerung, der dort bei Zensusdaten und auch bei einer anderen großen Haushaltsstatistik eingesetzt wird. Aus den Publikationen erfährt man aber, dass intensiv daran gearbeitet wird, den Ansatz noch breiter nutzen zu können, z. B. innerhalb des Forschungsdatenzugangs und nicht nur bei Häufigkeitsauszählungen, sondern auch für Auswertungen quantitativer Merkmale.

Was bedeutet ein Wechsel beim Geheimhaltungsverfahren für die Kundinnen und Kunden sowie die Kolleginnen und Kollegen in den Fachbereichen der amtlichen Statistik?

Erste Erfahrungen mit dem Wechsel haben wir im Zensus 2011 gemacht. Insgesamt ist es sehr wichtig, einen Wechsel wirklich gut vorzubereiten. Die Nutzer jedenfalls werden nur dann die Vorteile erkennen können, wenn dadurch unser Datenangebot zumindest im Vergleich zu dem, was mit im Verbund konsequent durchgeführter Zellsperrung möglich ist, spürbar größer und besser zugänglich wird. Für die Kolleginnen und Kollegen in den Fachbereichen wird das vermutlich bedeuten, dass sie sich an neue Auswertungssysteme gewöhnen müssen.
Ganz wichtig ist, dass es uns gelingt, den Nutzerinnen und Nutzern zu vermitteln, dass die Zuverlässigkeit der Auswertungen nicht wesentlich beeinträchtigt ist, soweit es Ergebnisse betrifft, die im bisherigen Zellsperrungsverfahren nicht der primären Geheimhaltung unterliegen. Bevor ein Wechsel zu datenverändernden Verfahren vollzogen wird, müssen Strategien für die Kommunikation der Datenveränderung den Nutzern gegenüber sowie Fachkonzepte für geeignete Auswertungssysteme entwickelt werden. Um einen Wechsel vorzubereiten, müssten hier in den nächsten Jahren die Arbeitsschwerpunkte der amtlichen Statistik in der Geheimhaltung liegen.

1 § 16 des Bundesstatistikgesetzes (BStatG) vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 13 des Gesetzes vom 25. Juli 2013 (BGBl. I S. 2749), schreibt für Bundesstatistiken Verfahrensregeln zum Schutz der Vertraulichkeit von Einzeldaten vor.

2 Beim Record Swapping werden die Ausprägungen einzelner Merkmale (typischerweise: Gebietsgliederung in der feinsten Darstellungsebene) zwischen Erhebungseinheiten, die in Bezug auf die Ausprägung bestimmter Kontrollmerkmalen (z. B. zur Haushalts-/Familien-/Altersstruktur) identisch sind, getauscht.

Geheimhaltung

FiRe – Ein Schritt zur Teilautomatisierung der Geheimhaltungsprüfung

von **Jakob Pohlisch, Julia Höninger, Ramona Voshage**

Das Programm FiRe kann bei der Geheimhaltungsprüfung von statistischen Ergebnissen in Forschungsdatenzentren eingesetzt werden, wenn das Statistiksoftwareprogramm Stata verwendet wird. Es übernimmt automatisch einzelne Schritte der Inputkontrolle, indem einzelne Befehle unterdrückt und nicht ausgeführt werden. Im Rahmen einer automatisierten Prozesskontrolle wird von Befehlen in Stata nur der Teil des Outputs angezeigt und im Log-File ausgegeben, der unter Geheimhaltungsgesichtspunkten unkritisch ist und veröffentlicht werden darf. Bei dem vorliegenden Beitrag handelt es sich um eine geringfügig modifizierte Fassung des FDZ-Arbeitspapiers Nr. 47, erschienen im Januar 2015, Düsseldorf.

1 Einleitung

Seit 2002 wurden bei den großen amtlichen Datenproduzenten Forschungsdatenzentren (FDZ) eingerichtet, um der Wissenschaft den Zugang zu Mikrodaten zu ermöglichen. So gibt es inzwischen unter anderem die FDZ der Statistischen Ämter des Bundes und der Länder (Zühlke et al. 2007), ein FDZ der Deutschen Rentenversicherung (Stegmann 2009) und ein FDZ der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung (Allmendinger und Kohlmann 2005). Alle FDZ ermöglichen verschiedene Wege des Mikrodatenzugangs, die meisten FDZ bieten Wissenschaftlern[1] die Möglichkeit, entweder an speziell ausgestatteten Gastwissenschaftlerarbeitsplätzen in den Räumen der Datenproduzenten zu arbeiten oder die kontrollierte Datenfernverarbeitung zu nutzen. Bei letztgenannter Art des Datenzugangs senden Wissenschaftler Auswertungsprogramme an das FDZ, die FDZ-Mitarbeiter führen diese auf den Originaldaten aus und senden die Ergebnisse an die Wissenschaftler zurück. Alle Ergebnisse werden dabei vor der Freigabe an die Wissenschaftler von FDZ-Mitarbeitern auf die Einhaltung des Statistikgeheimnisses geprüft. Diese statistische Geheimhaltungsprüfung ist teilweise sehr aufwändig.

Im Projekt „infinitE – Eine informationelle Infrastruktur für das E-Science Age“ (Brandt und Zwick 2009), das vom Bundesministerium für Bildung und Forschung von Mai 2009 bis Dezember 2012 finanziert worden ist, wurden die Grundlagen für Ansätze zu einer (teil-)automatisierten Geheimhaltungsprüfung untersucht. Der Zugang zu Mikrodaten könnte durch ein echtes Fernrechnen, den sogenannten Remote Access, für die Wissenschaft komfortabler werden und in den FDZ weniger Ressourcen binden. Im Projekt wurden zwei Ansätze entwickelt: Alternative II wurde „Morpheus“ genannt (Höhne und Höninger 2012). Die Idee des ganzheitlichen Ansatzes zur Automatisierung der Geheimhaltungsprüfung bei Morpheus ist, dass die Wissenschaftler auf anonymen Daten in Echtzeit rechnen und zu jedem Ergebnis ein Gütemaß erhalten. Bei der Alternative I wurde die Idee einer kombinierten Input-, Prozess- und manuellen Stichprobenkontrolle konzeptionell entwickelt (Hochgürtel und Brandt 2011). Inputkontrolle bedeutet hier, dass Befehle vor der Ausführung geprüft werden und solche Befehle, die in allen Fällen ein Enthüllungsrisiko darstellen, deaktiviert und nicht ausgeführt werden. Viele Befehle erzeugen jedoch Ergebnisse, die nur in manchen Konstellationen ein Enthüllungsrisiko darstellen, es müssen also jeweils die Rahmenbedingungen geprüft werden. Diese Prüfung übernimmt die Prozesskontrolle, die nur den Teil der Ergebnisse anzeigt, von denen kein Enthüllungsrisiko ausgeht. Der vorliegende Beitrag erläutert das Programm FiRe, das als ein Baustein in solch einer kombinierten Alternative I eingesetzt werden könnte. Mit FiRe kann die Input- und Prozesskontrolle bei Verwendung des Statistiksoftwareprogrammes Stata teilautomatisiert werden.

2 Grundlegende Idee von FiRe

Die Kontrollierte Datenfernverarbeitung erlaubt als einziger Zugangsweg der Forschungsdatenzentren die Analyse formal anonymisierter Einzeldaten. Für Datennutzer besteht jedoch kein direkter Zugang zu diesen Mikrodaten. Die Datennutzer erhalten Strukturdatensätze (Dummy-Dateien), die in Aufbau und Merkmalsausprägungen den Originaldaten weitestgehend gleichen. Mittels dieser Dummy-Dateien

1 Zur sprachlichen Vereinfachung wird in diesem Text nur die männliche Form des Wortes „Wissenschaftler“ und seine Synonyme verwendet, wobei die Forschungsdatenzentren natürlich weibliche und männliche Wissenschaftler betreuen. Gleiches gilt für die im Folgenden verwendete Bezeichnung „FDZ-Mitarbeiter“.

können Auswertungsprogramme in den Analyseprogrammen SPSS, SAS, R oder Stata erstellt werden, mit denen die FDZ-Mitarbeiter in den statistischen Ämtern anschließend die Originaldaten auswerten. Die Datennutzer erhalten nach einer notwendigen Geheimhaltungsprüfung schließlich die Ergebnisse dieser Auswertung.

Die Geheimhaltungsprüfung in den FDZ wird momentan ausschließlich manuell durchgeführt und ist außerordentlich zeit- und ressourcenintensiv. Um den Aufwand zu reduzieren, wurde im Rahmen des Projektes InfinitE nach Möglichkeiten einer automatisierten Geheimhaltung gesucht. Zwar existieren Verfahren zur automatischen Geheimhaltung von Tabellen, diese sind jedoch für Zwecke der FDZ nicht flexibel genug einsetzbar, um den Arbeitsaufwand in den FDZ zu reduzieren. Als eine gute Möglichkeit hat sich dagegen die Manipulation[2] von Befehlen herausgestellt. Befehle werden dabei so manipuliert, dass der zu generierende Output automatisch einer primären Geheimhaltungsprüfung unterzogen wird. Die ebenfalls notwendige sekundäre Geheimhaltung obliegt weiterhin den Mitarbeitern der FDZ.

Das Programm FiRe wurde im FDZ-Standort Berlin-Brandenburg entworfen und programmiert. Hier wurden einige erste Stata-Befehle umgeschrieben, um die grundsätzliche Machbarkeit einer Input- und Prozesskontrolle bei dem Statistiksoftwareprogramm Stata zu zeigen. Damit die umprogrammierten Befehle im Tagesgeschäft der FDZ eingesetzt werden können, wurde ein Programm namens FiRe, als Abkürzung von „Find & Replace", entwickelt. Es ist innerhalb der Alternative I zur automatisierten Geheimhaltungsprüfung ein Baustein, um die Möglichkeit der Umsetzung der Prozesskontrolle stellvertretend am Statistiksoftwareprogramm Stata zu demonstrieren. Die von den Nutzern verfassten Do-Files werden mit Hilfe von FiRe so verändert, dass statt der Standardbefehle die neuen umprogrammierten Befehle verwendet werden. Durch diese werden die Bedingungen für eine Freigabe von statistischen Ergebnissen bis zu einem gewissen Grad automatisiert geprüft. Das Programm ist so konzipiert, dass neu umprogrammierte Befehle sehr leicht eingebunden werden können. Eine Textdatei enthält die Namen aller manipulierten Befehle. Kommen neue Befehle hinzu, kann die Textdatei entsprechend erweitert werden.

Insgesamt wurden vom FDZ-Standort Berlin-Brandenburg in einem ersten Schritt acht Stata Ado-Files manipuliert. Die Eingriffe sind im Abschnitt 4 dokumentiert. Die Prozesskontrolle durch das Programm FiRe kann bereits alleinstehend verwendet werden, ohne dass die Elemente Input- oder Stichprobenkontrolle die im infinitE-Projekt entwickelte Alternative I zu einem vollautomatisierten Datenfernzugang ergänzen würden.

2 Der Begriff Manipulation wird hier in dem Sinne verwendet, dass die Originalbefehle, wie sie in dem Statistiksoftwareprogramm Stata implementiert sind oder wie sie von anderen externen Programmierern bereitgestellt werden, geändert und modifiziert werden. Sie entsprechen danach nicht mehr exakt dem Original; dieses wird verändert.

3 FiRe

Anstelle der von den Nutzern angegebenen Originalbefehle sollen die manipulierten Befehle genutzt werden. An dieser Stelle greift FiRe ein, indem es den Aufruf der Originalbefehle durch ein „Suchen und Ersetzen" („Find & Replace"), durch einen Aufruf der manipulierten Befehle, ersetzt. FiRe ist ein VBA-Makro, welches als add-on installiert und anschließend aus Microsoft Word gestartet werden kann. Die aktuelle Version des Programms FiRe ist Programmversion 3.5. Das Programm besteht aus den Teilen Pre- und Post-FiRe.

3.1 Pre-FiRe

Pre-FiRe durchsucht die Do-Files der Datennutzer vor deren Ausführung nach vorher definierten Befehlen und benennt diese um. So werden für die originäre Kontrollierte Datenfernverarbeitung manipulierte Ado-Files verwendet. Alle Befehle, für die eine manipulierte Version vorliegt, können in der Datei „preFire.txt" aufgelistet werden. Es wurde vereinbart, dass manipulierte Ado-Files nach folgendem Muster zu benennen sind: Dem ursprünglichen Befehlsnamen wird der Präfix „fdz" vorangestellt. Die manipulierte Version von „summarize" heißt dann beispielsweise „fdzsummarize".

Das bearbeitete Do-File wird am gleichen Speicherort mit einer Erweiterung des Dateinamens um _FiRe.do gespeichert. Das originale Do-File des Wissenschaftlers wird somit nicht verändert. Wenn der FDZ-Mitarbeiter nun das geänderte Do-File startet, werden jeweils die manipulierten Befehle statt der Originalbefehle verwendet. Je mehr manipulierte Ado-Files bereitgestellt werden, umso mehr kann die manuelle Outputprüfung automatisiert werden.

Darüber hinaus fügt Pre-FiRe zu Beginn des Do-Files automatisch zusätzliche Informationen ein. Für den Wissenschaftler wird durch Kommentarzeilen deutlich und transparent gemacht, dass das Do-File mit FiRe bearbeitet wurde. Es wird angegeben, bei welchen Befehlen Namenserweiterungen durchgeführt wurden. Wichtig ist auch der Verweis auf den Ordner, in dem alle umgeschriebenen Befehle gesammelt werden, damit Stata diese neuen Befehle anwenden kann.

Da Befehle in Stata oft abgekürzt werden, werden auch alle zulässigen Abkürzungsvarianten geprüft. Dabei wurde darauf geachtet, dass nur Befehle den Namenszusatz „fdz" erhalten, nicht jedoch andere Wörter oder ähnlich lautende Befehle, in denen die Buchstabenfolgen, beispielsweise „reg" oder „sum", vorkommen. Die neuen manipulierten Befehle müssen demnach auch in der abgekürzten Version bereitgestellt werden.

3.2 Post-FiRe

Um den Aufwand bei der Geheimhaltungsprüfung weiter zu reduzieren, können mit Post-FiRe Outputs zu einem gewissen Grad vorgeprüft werden. Es werden beispielsweise Angaben wie „1 missing value generated" oder „1 obs not used" automatisch gesperrt.

Sollten weitere Ersetzungen gewünscht sein, können diese in der Datei „postFire.txt" ergänzt werden. In einer Zeile steht der zu ersetzende, in der darauffolgenden Zeile der Text, der stattdessen eingefügt werden soll. In jeder Zeile müssen zu Beginn zwei Rautezeichen (##) stehen, da die zu ersetzenden Textstellen auch mit einem Leerzeichen beginnen können.

Das mit Post-FiRe bearbeitete Log-File wird mit der erweiterten Dateiendung _FiRe.log im gleichen Ordner wie die Ausgangsdatei gespeichert. Somit wird auch hier sichergestellt, dass die originale Outputdatei nicht verändert wird.

4 Manipulierte Ado-Files im Detail

Die Idee besteht darin, die Befehle bezüglich ihres Outputs zu verändern. Leider liegen jedoch nicht von allen Befehlen die entsprechenden Quellcodes vor. Darüber hinaus würde eine Manipulation dieser die Nutzung der originalen Befehle unmöglich machen.

Anstatt die Befehle direkt zu manipulieren, werden die neuen Befehle auf Grundlage der existierenden programmiert. Dabei wurden bisher u. a. die Befehle regress, summarize und tabstat entsprechend verändert. Dies liegt zum einen an dem Aufwand der Manipulation der Befehle und zum anderen an der Häufigkeit, mit der diese im täglichen Betrieb der FDZ vorkommen.

Die Manipulation der Befehle unterscheidet sich grundlegend, je nach Vorhandensein eines manipulierbaren Ado-Files. Ist kein solches Programm vorhanden, muss eines geschrieben werden, welches den ursprünglichen Befehl unter Berücksichtigung der Geheimhaltung perfekt imitiert. Ist ein Ado-File vorhanden, kann dies meist deutlich einfacher angepasst werden.

Durch die Manipulation der Ado-files ergeben sich naturgemäß Fehlerquellen durch eine inkorrekte Programmierung. Die Vermeidung von Fehlern in den manipulierten Ado-Files muss als mindestens so wichtig eingeschätzt werden wie die Geheimhaltung selbst. Nichts wäre ein herberer Rückschlag auf dem Weg zur automatisierten Geheimhaltung als falsch generierter und anschließend durch Wissenschaftler veröffentlichter Output. Die bisher veränderten Befehle wurden daher vor der Einführung in den Alltagsbetrieb der FDZ ausgiebigen Tests unterzogen.[3] Im Folgenden werden die veränderten Befehle genannt und näher erläutert.

4.1 regress

Durch den regress-Befehl kann in Stata eine lineare Regression durchgeführt werden. Die Frage, welche Enthüllungsrisiken unter anderem durch lineare Regressionen entstehen können, wurde bereits von Reznek (2003), Reznek und Riggs (2005) und Vogel (2011) untersucht. Es muss an dieser Stelle festgehalten werden, dass die hier implementierte Geheimhaltung nicht vor allen denkbaren Angriffsszenarien schützt. Eine kontinuierliche, gewissenhafte Prüfung des generierten Schätzoutputs muss somit auch weiterhin gewährleistet sein.

In Vogel (2011) finden sich folgende Enthüllungsszenarien, denen durch die Umprogrammierung des Befehls regress begegnet wird:

- Variablen mit nur ein oder zwei von Null verschiedenen Beobachtungen,
- Variablen, die für alle, außer für ein oder zwei Beobachtungen, nur sehr kleine Werte annehmen.

Um diese Risiken zu eliminieren, muss jede einzelne im regress-Befehl aufgeführte unabhängige Variable auf diese beiden Risiken hin untersucht werden. Sowohl Szenario (a) als auch (b) stellen Dominanzfälle dar und können durch eine einfache Dominanzprüfung der beteiligten Variablen ausgeschlossen werden.

Der manipulierte Ado-File fdzregress enthält eine solche Prüfung. Dabei wurde sowohl eine Prüfung nach der (2,k)-Regel als auch eine Prüfung nach der p%-Regel implementiert. Die (2,k)-Regel überprüft, ob 2 Beobachtungseinheiten mehr als k % der Variablensumme auf sich vereinigen. Die p%-Regel hingegen prüft, ob sich der Wert der betragsmäßig größten Beobachtungseinheit mit Hilfe des zweitgrößten Einzelwertes schätzen lässt. Ist der Schätzfehler kleiner als p%, liegt ein Dominanzfall vor. Da sich die für k und p zu nutzenden Werte je nach Statistik unterscheiden können und darüber hinaus geheim zu halten sind, wurde der Befehl fdzregress um die Option critical() erweitert. Hier kann innerhalb der Klammern der Wert für die p%-Regel in Prozent spezifiziert werden. Soll beispielsweise p=15 sein, so lautet der Befehl:

```
fdzregress [varlist], critical(15)
```

Ist bei einer Statistik der Parameter k der (2,k)-Regel vorgegeben, so kann der Parameter p der p%-Regel, der das gleiche Schutzniveau wie die die (2,k)-Regel bietet, anhand der folgenden Formel berechnet werden (Statistisches Bundesamt 2007):

$$p = 100\,\frac{100-k}{100}$$

Soll die (2,k)-Regel mit k=85% angewendet werden, so bietet eine p%-Regel mit

$$p = 100\,\frac{100-85}{100} = 17{,}6$$

den gleichen Schutz vor annäherungsweiser Enthüllung einer Einzelangabe.

Sollte ein Dominanzfall vorliegen, wird die entsprechende Variable, die Art der Dominanzprüfung und der errechnete Wert für p bzw. k ausgegeben. Handelt es sich bei einer der überprüften Variablen um eine Variable mit einer Fallzahl von weniger als drei Beobachtungen für eine Ausprägung, so wird zusätzlich eine Häufigkeitstabelle der Variablen ausgegeben. Sowohl die Fallzahlprüfung als auch die Dominanzprüfung wird für jede Variable einzeln und ggf. innerhalb der Gruppierungen durchgeführt, welche beispielsweise über den Befehl „bysort" spezifiziert werden können.

3 Die manipulierten Ado-Files können auf Anfrage zur Verfügung gestellt werden.

Der so produzierte Output enthält nun zunächst die geheim zu haltenden Informationen über p und k. Diese können jedoch leicht mit Hilfe von Post-FiRe gelöscht werden.

Die Entscheidung, ob und in welcher Weise der Output gesperrt werden muss, obliegt nach wie vor den Mitarbeitern der FDZ. Vorteil dieses Verfahrens ist jedoch, dass jede Regression auf eventuelle Risiken untersucht wird und eine manuelle Prüfung auf Dominanz entfällt.

An dieser Stelle muss erwähnt werden, dass sich die Prüfung auf Variablen beschränkt, die nur positive reelle Zahlen enthalten.[4] Ist eine Variable auf ganz $\mathbb{R}$ definiert, führt die Prüfung zu falschen Ergebnissen, da sich bei der Summation der Einzelwerte positive und negative Werte gegenseitig aufheben können. Somit wäre die berechnete Gesamtsumme und damit auch der berechnete Anteil der betragsmäßig größten Beobachtungseinheiten nicht korrekt. Dieses Problem muss den Prüfenden unbedingt bewusst sein, um unnötige Sperrungen zu verhindern.

4.2 tabstat

Der tabstat-Befehl generiert allgemeine Statistiken von zu definierenden Variablen, bezogen auf den gesamten Datensatz oder beliebig definierbare Untergruppen. Die Geheimhaltung dieser Statistiken ist in den FDZ des Bundes und der Länder einheitlich festgeschrieben. Diese Regeln wurden in die frei verfügbare Syntax des Befehls integriert. Es werden keine Statistiken auf der Grundlage von weniger als drei Beobachtungseinheiten veröffentlicht. Zusätzlich werden die Optionen min (Minimum) und max (Maximum) im Output gänzlich unterdrückt, da diese Variablenwerte einzelne Beobachtungen darstellen. Die Fallzahlregel bezieht sich hier insbesondere auch auf die Ausgabe von Perzentilen. So wird das 1%-Perzentil erst ab einer zugrundeliegenden Fallzahl von 300, das 5%-Perzentil erst ab einer Fallzahl von 60 etc. veröffentlicht.

Sollte der Befehl tabstat auf eine Variable angewendet werden, die für alle, außer für ein oder zwei Beobachtungen, nur sehr kleine Werte annimmt, so geben die Statistiken Werte aus, die näherungsweise Einzelwerte darstellen können. Dieses Risiko kann mit einer Dominanzprüfung eliminiert werden. Wie schon im regress-Befehl wurde auch im tabstat-Befehl eine solche Prüfung implementiert, die über die Option critical() steuerbar ist. Der entstandene manipulierte Befehl wurde unter dem Namen fdztabstat.ado gespeichert und steht zur Implementierung bereit.

4.3 summarize

Der Output des summarize-Befehls ist dem des tabstat-Befehls sehr ähnlich. Dies ist der Tatsache geschuldet, dass der tabstat-Befehl den summarize-Befehl zur Berechnung seiner Statistiken nutzt. Im Unterschied zum tabstat-Befehl ist die Syntax des summarize-Befehls jedoch nicht zugänglich und von daher nicht manipulierbar. Der Befehl fdzsummarize beruht somit nicht auf einer Manipulation des ursprünglichen Befehls. Vielmehr handelt es sich um einen Klon des Befehls, dessen Rechenroutinen weiterhin auf summarize beruhen, jedoch mit einer integrierten Geheimhaltungsprüfung. Der Output ist dem des Originalbefehls nachempfunden. Unterliegen Teile des Outputs der Geheimhaltung, werden sie nicht ausgegeben. Die Geheimhaltung erfolgt nach den gleichen Regeln wie bei fdztabstat. Auch im Befehl fdzsummarize wurde eine Dominanzprüfung nach dem gleichen Prinzip wie schon bei fdzregress und fdztabstat integriert.

4.4 codebook

Der Befehl „codebook" gibt einen Überblick über die Eigenschaften und Ausprägungen einer Variable und gibt entweder die Wertelabel und eine einfache Häufigkeitsverteilung bei einem kategorialen Merkmal oder deskriptive Kennzahlen bei einem metrischen Merkmal aus. Durch die Umprogrammierung wird das Intervall, in welchem sich die Ausprägungen der aufgezählten Variablen befinden („range:"), nicht mehr angezeigt, da das Minimum und das Maximum zu schützende Einzelangaben sind. Die Anzahl der unterschiedlichen Ausprägungen („unique values:") wird nicht mehr angezeigt, da sie in einigen Fällen einen Rückschluss auf den größten Wert zulassen. Weiterhin muss manuell überprüft werden, ob die Fallzahl zur Angabe von Perzentilen ausreicht und ob die Mindestfallzahl bei den verschiedenen Kategorien in Häufigkeitstabellen gewahrt ist.

4.5 utest

Mit dem Befehl „utest" kann getestet werden, ob zwischen zwei Merkmalen eine U-förmige Beziehung besteht. Es handelt sich um einen Befehl, der als Ado-File installiert werden kann, er ist nicht standardmäßig in Stata enthalten. Im Output des Befehls wird durch die Manipulation sowohl der Extremwert des abhängigen Merkmals als auch das Intervall der Werte auf der x-Achse nicht angezeigt. Intern werden diese Werte berechnet, nur die Anzeige wird unterdrückt.

4.6 xtsum

Der Befehl gibt deskriptive Kennzahlen für eine Panelvariable aus. Dabei werden der Mittelwert und die Streuung sowie Minimum und Maximum sowohl für die ganze Variable ausgegeben als auch die Streuung in zwei Komponenten zerlegt: die Streuung zwischen den Beobachtungseinheiten im Querschnitt und die Streuung der Beobachtungseinheiten im Zeitverlauf. Die Extremwerte sollen auch bei diesem Befehl nicht herausgegeben werden. Der Befehl „xtsum" wurde so umprogrammiert, dass die Extremwerte zwar berechnet werden, statt der Zahlangaben aber die Textangabe „gesperrt" ausgegeben wird.

4.7 inspect

Dieser Befehl erzeugt ein rudimentäres Histogramm und gibt an der x-Achse den kleinsten und größten Wert an. Diese Einzelangaben müssen jedoch geheim gehalten werden. Mit FiRe kann eingestellt werden, dass der Befehl „fdzinspect" anstatt des Be-

4 Die Null ist in diesem Fall Element der positiven reellen Zahlen.

fehls „inspect“ ausgeführt wird, wobei anstatt einer Ausführung des Befehls der Text „*FDZ: Befehl beim Fernrechnen nicht zulässig“ erscheint. FiRe bietet insofern auch eine Teilautomatisierung der Inputkontrolle.

4.8 xtdescribe

Der Befehl „xtdescribe“ gibt Kennzahlen zur Panelstruktur aus, z. B. die vorhandenen Teilnahmestrukturen der Beobachtungseinheiten und deren Häufigkeiten. Bei der Angabe der Anzahl der Teilnehmer im Querschnitt werden einige Beispiel-IDs ausgegeben, die in der manipulierten Variante des Befehls nun durch den Text „*FDZ: gesperrt“ ersetzt werden.

5 Ausblick

Die Prozesskontrolle durch FiRe und die manipulierten Ado-Files können zwar bereits eigenständig genutzt werden, stellen aber nur einen ersten Schritt in Richtung vollautomatisierter Ergebniskontrolle dar.

Durch den Einsatz von FiRe kann von einer enormen Zeitersparnis bei der manuellen Prüfung auf Geheimhaltung ausgegangen werden. Die größte Reduktion des Prüfaufwandes ist sicherlich durch die Umprogrammierung des Befehls „tabstat“ zu erzielen. Dieser Befehl wird von den Wissenschaftlern im FDZ besonders häufig verwendet. Um den Nutzen weiter zu steigern, müssen in Zukunft weitere Befehle manipuliert und eingebunden werden.

Durch die zusätzlich in die manipulierten Befehle integrierten Prüfungen auf Fallzahl- und Dominanzregelungen steigt selbstverständlich auch die für deren Ausführung benötigte Rechenzeit. Im Falle von „regress“ und „tabstat“ benötigt der manipulierte Befehl im arithmetischen Mittel ca. 3-mal länger als das Original. Der „summarize“-Befehl hingegen benötigt im arithmetischen Mittel ca. 4-mal so lang wie der Originalbefehl. Die ermittelten Zeiten beruhen auf mehrmaligen Berechnungen mit dem gleichen Datensatz auf unterschiedlichen Computern. Dabei waren keine deutlichen Unterschiede zwischen den relativen Rechenzeiten bei unterschiedlichen Hardwarespezifikationen und Betriebssystemen feststellbar. Da es sich bei den Rechenzeiten um Sekunden handelt, ist die entstehende Verzögerung einer manuellen Prüfung jedoch in jedem Fall vorzuziehen.

Sowohl die Inputkontrolle als auch die Prozesskontrolle kommt durch FiRe teilautomatisiert schon heute in den FDZ zum Einsatz. Eine vollständige Implementierung der im Projekt infinitE entwickelten Alternative I zur (voll-) automatisierten Geheimhaltungsprüfung sind diese ersten Beispielprogrammierungen aber noch nicht. Auch können diese bisher nur eingesetzt werden, wenn die Wissenschaftler die Daten des FDZ mit der Software Stata auswerten. FiRe ist allerdings auf jedes statistische Softwareprogramm erweiterbar, dessen Syntax und Befehle in einem Texteditor bearbeitet werden können.

Jakob Pohlisch entwickelte im Rahmen seiner Tätigkeit im Referat *Mikrodaten, Analysen, Forschungsdatenzentrum* im Amt für Statistik Berlin-Brandenburg das hier vorgestellte Programm FiRe.

Julia Höninger leitet das Referat *Volkswirtschaftliche Gesamtrechnungen, Erwerbstätigkeit.* Zuvor arbeitete sie als wissenschaftliche Mitarbeiterin im Referat *Mikrodaten, Analysen, Forschungsdatenzentrum* des Amtes für Statistik Berlin-Brandenburg.

Ramona Voshage leitet das Referat *Mikrodaten, Analysen, Forschungsdatenzentrum* des Amtes für Statistik Berlin-Brandenburg.

Literatur

- Allmendinger, Jutta und Kohlmann, Annette (2005): Datenverfügbarkeit und Datenzugang am Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung. Allgemeines Statistisches Archiv 88, S. 159–182.
- Brandt, Maurice und Zwick, Markus (2009): infinitE – Eine informationelle Infrastruktur für das E-Science Age; Verbesserung des Mikrodatenzugangs durch „Remote-Access“. Wirtschaft und Statistik 7/2009, Statistisches Bundesamt, Wiesbaden.
- Hochgürtel, Tim und Brandt, Maurice (2011): „InfinitE – Eine informationelle Infrastruktur für das E-Science Age: Verbesserung des Mikrodatenzugangs durch „Remote-Access“, Vortrag auf der Konferenz für Sozial- und Wirtschaftsdaten 13.–14.01.2011 in Wiesbaden.
- Höhne, Jörg und Höninger, Julia (2012): Das Verfahren Morpheus – Auf dem Weg zu Remote Access. Working Paper Series des Rates für Sozial- und Wirtschaftsdaten (RatSWD) 205/2012, Berlin. Verfügbar unter: http://www.ratswd.de/download/RatSWD_WP_2012/RatSWD_WP_205.pdf
- Reznek, Arnold P. (2003): Disclosure Risks in Cross-Section Regression Models. American Statistical Association 2003, Proceedings of the Section on Government Statistics and Section on Social Statistics: 3444–3451.
- Reznek, Arnold P. und Riggs, T. Lynn (2005): Disclosure Risks in Releasing Output Based on Regression Residuals. American Statistical Association 2005, Proceedings of the Section on Goverment Statistics and Section on Social Statistics: 1397–1404.
- Statistisches Bundesamt (2007): Entwurf – Leitfaden zur Festlegung eines p-Wertes für die p%-Regel zur Tabellengeheimhaltung. Anlage 6 zum Sachstandsbericht an den AOU vom September 2007, Internes Dokument, Statistisches Bundesamt, Wiesbaden.
- Stegmann, Michael (2009): Das aktuelle Datenangebot und Neuentwicklungen im FDZ-RV, DRV-Schriften Band 55/2009, S. 27–36.
- Vogel, Alexander (2011): Enthüllungsrisiko beim Remote Access: Die Schwerpunkteigenschaft der Regressionsgerade. FDZ-Arbeitspapier Nr. 36. Statistische Ämter des Bundes und der Länder, Düsseldorf
- Zühlke, Sylvia; Christians, Helga und Cramer, Katharina (2007): Das Forschungsdatenzentrum der Statistischen Landesämter – eine Serviceeinrichtung für die Wissenschaft. AStA Wirtschafts- und Sozialstatistisches Archiv 3-4, S. 169–178.

Neuerscheinung

Interaktive Zensusergebnisse für Berlin jetzt auch kleinräumig

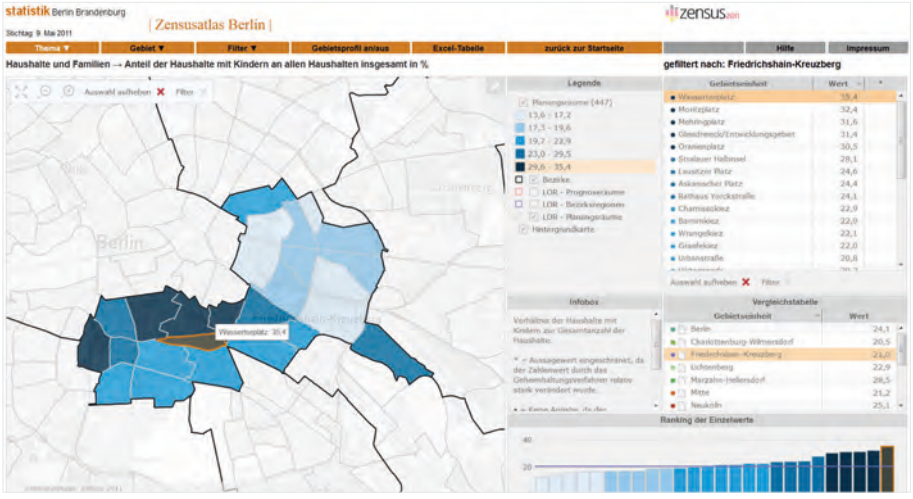
In welchem Kiez ist der Anteil der Haushalte mit Kindern am höchsten? Wo wohnen besonders viele ledige Männer? Welcher Stadtteil besitzt den jüngsten Wohngebäudebestand? Fragen wie diese lassen sich für Berlin nun mit wenigen Klicks beantworten. Die Ergebnisse werden in Form von dynamischen Karten, Tabellen und Grafiken im Internetangebot des Amtes für Statistik Berlin-Brandenburg (AfS) visualisiert.

Nach der Veröffentlichung der beiden Zensusatlanten für die Länder Berlin und Brandenburg im vergangenen Jahr[1] erfolgte im April 2015 eine räumliche und thematische Erweiterung der Online-Anwendung. Für das Land Berlin wurden ausgewählte Ergebnisse des Zensus 2011 nun auch in tiefer regionaler Gliederung zur interaktiven Nutzung aufbereitet. Unterhalb der Bezirksebene können sich Nutzerinnen und Nutzer eine Vielzahl von Merkmalen aus den Themenblöcken Bevölkerung, Gebäude und Wohnungen sowie Haushalte und Familien auf allen drei Ebenen der Lebensweltlich orientierten Räume (LOR) – Prognoseräume, Bezirksregionen und Planungsräume – anzeigen lassen. Eine nun integrierte Hintergrundkarte sorgt für eine bessere räumliche Orientierung beim Hineinzoomen in die Karte. Auf kleinräumiger Ebene besteht zudem die Möglichkeit, Ergebnisse nach Bezirken, der Stadtaufteilung in Innen-/Außenstadtbereich und der historischen Ost-/West-Teilung zu filtern. Mit Hilfe einer Vergleichstabelle lassen sich die Werte einzelner Stadtgebiete mit denen der Bezirke oder der Stadt vergleichen.

Im Zensusatlas für das Land Brandenburg werden die Erhebungsergebnisse in gleicher Weise auf Ebene der Kreise und Gemeinden präsentiert. Nutzerinnen und Nutzer können die kleinräumigen Gemeindedaten nach Landkreisen, Planungsregionen, Mittelbereichen zentraler Orte sowie der Unterteilung Berliner Umland und weiterer Metropolenraum filtern. Eine noch kleinteiligere Darstellung der Zensusergebnisse für das Land Brandenburg auf der Ebene einzelner Ortsteile wird angestrebt.

In beiden Atlanten kann für jede beliebige Raumeinheit ein Gebietsprofil dargestellt werden. Es enthält alle Indikatoren des Zensusatlas und zeigt deren Abweichungen zum Landesdurchschnitt an. Durch die Auswahl mehrerer Gebiete ist auch ein Vergleich statistischer Profile möglich. Das komplette, kleinräumige Datenangebot lässt sich aus der Anwendung heraus als Excel-Tabelle herunterladen. Als besondere Form der Interaktion ermöglichen die Atlanten weiterhin eine individuell veränderbare Datenklassifizierung (Methode und Klassenanzahl) und Farbauswahl.

Bei der Veröffentlichung kleinräumiger statistischer Daten ist auch die Frage der Geheimhaltung ein wichtiges Thema. Die Zensusatlanten tragen dem Rechnung, indem die veröffentlichten Ergebnisse zuvor einem Geheimhaltungsverfahren unterzogen wurden. Hierbei werden die Originaldaten teilweise leicht verändert. Nur in wenigen Einzelfällen ist die Abweichung vom Originalwert so groß, dass eine Ausweisung des Wertes nicht erfolgen kann. Die Gebietseinheit ist dann entsprechend gekennzeichnet. Etwas häufiger tritt der Fall auf, dass der Wert relativ stark vom Originalwert abweicht, aber aus Geheimhaltungssicht dennoch veröffentlicht werden kann. Diese Werte sind speziell markiert, um den Nutzerinnen und Nutzern die eingeschränkte Verwertbarkeit der betreffenden Zahl zu verdeutlichen.



Die Zensusatlanten für Berlin und Brandenburg sind abrufbar unter:
www.statistik-berlin-brandenburg.de/instantatlas/interaktive-karten.asp

1 Siehe Zeitschrift für amtliche Statistik Berlin Brandenburg, Heft 2/2014, S. 8 f.

Zensus

Erstbezugseigentümer in Berlin und Brandenburg

– eine generationenbezogene Analyse von Personen-, Haushalts-, Gebäude- und Wohnungsmerkmalen auf Basis der Ergebnisse des Zensus 2011

von **Verena Kutzki** und **Marco Schwarz**

Den Schwerpunkt dieses Beitrags bildet eine generationenbezogene Analyse von Bevölkerung und Haushalten der Länder Berlin und Brandenburg auf Basis der Ergebnisse des Zensus 2011. Ziel hierbei ist, das Leben und Wohnen der Generationen in einem allgemeinen Überblick darzustellen. Gleichzeitig wird diese Analyse genutzt, um Verbesserungsansätze für den Zensus 2021 aufzuzeigen.

Einleitung

Die Ergebnisse des Zensus 2011 und der Gebäude- und Wohnungszählung (GWZ) stellen einen in der Bundesrepublik Deutschland einmaligen Datenbestand dar, der es ermöglicht, die bundesweit flächendeckend vorliegenden Vollerhebungsergebnisse der Gebäude- und Wohnungszählung 2011 (GWZ) mit registerbereinigten demografischen Ergebnissen und sozioökonomischen Stichprobenergebnissen kombiniert auszuwerten. Diese Kombinierbarkeit der vier statistischen Teilbereiche Personen, Haushalte, Gebäude und Wohnungen ist so nur mit den Daten des Zensus 2011 möglich. Aber auch die Daten der einzelnen Teilbereiche liefern aktualisierte Basisinformationen und werden für Statistiken, wie beispielsweise die Bevölkerungs- und Bautätigkeitsfortschreibungen, als neue Datengrundlage genutzt und spiegeln die vorherrschende Situation zum 9. Mai 2011 in Deutschland wider.

In diesem Beitrag werden zum einen Merkmale der einzelnen Zensus-Teilbereiche auf Basis einer Generationenabgrenzung vorgestellt, um einen Überblick der vorherrschenden Gesamtsituation zum Zensusstichtag zu zeigen. Zum anderen wird eine komplexe Analyse – eine Kombination verschiedener Merkmale des Zensus 2011 – vorgestellt. Hierzu werden Erstbezugseigentümer – Personen bzw. Haushalte, deren Einzugsdatum mit dem Baujahr nahezu übereinstimmt – betrachtet.

Das Fazit am Ende dieses Beitrags widmet sich, unbeschadet der erstmals erfolgreich angewandten registergestützten Methodik des Zensus 2011, einigen Verbesserungsansätzen im Hinblick auf den bevorstehenden Zensus 2021. Angesprochen werden fachliche Einzelaspekte, welche aufgrund der bereits bestehenden bundesgesetzlichen Möglichkeiten keine langwierigen parlamentarischen oder juristischen Abstimmungsverfahren erforderlich machen sollten.

Zur Methodik der personen- und haushaltebezogenen Analysen

Die nachfolgenden Auswertungen mit Berichtszeitpunkt 9. Mai 2011 (Zensusstichtag) erfolgen auf Basis der endgültigen Ergebnisse des Zensus 2011. Aus Qualitätsgründen sowie aus Gründen der Nachvollziehbarkeit werden die nachfolgenden Auswertungen aus dem Datenmaterial erstellt, welches gemäß § 22 Abs. 2 Zensusgesetz 2011 (Gesetz über den registergestützten Zensus im Jahre 2011 (Zensusgesetz 2011 – ZensG 2011) vom 8. Juli 2009 (BGBl. I S. 1781)) auf Nachfrage auch den abgeschotteten Statistikstellen der Gemeinden und Gemeindeverbände für ausschließlich kommunalstatistische Zwecke zur Verfügung gestellt werden kann. Mit Rücksicht auf die statistische Geheimhaltung werden absolute Werte und Verhältniszahlen stets gerundet dargestellt.

Es werden zwei methodisch voneinander getrennte Analysen vorgenommen. Hierzu zählen eine personenbezogene Untersuchung, die alle 3 292 365 Einwohner Berlins umfasst, sowie eine haushaltebezogene Betrachtung.

Die haushaltebezogene Betrachtung, die in erster Linie eine Strukturanalyse der Zusammenhänge von Haushalts-, Gebäude- sowie Wohnungsmerkmalen zum Ziel hat, führt zu einer Reduzierung der betrachteten Personenanzahl. Der Grund hierfür sind Mehr-Personen-Haushalte. Um diese in Bezug auf das Thema dieses Beitrags analysieren und in einen Generationskontext setzen zu können, muss zunächst eine Bezugsperson je Haushalt bestimmt werden, die diesen Haushalt repräsentiert.

Da die Bestimmung der Generationenzugehörigkeit von Personen ausschließlich auf Basis des Geburtsjahres und anschließender Klassenzuordnung erfolgt, die verschiedenen Personen innerhalb eines Haushalts jedoch im Regelfall unterschiedliche Geburtsjahre aufweisen, ist eine gesamthaushalts- oder familienbasierte Generationenauswertung nur

über die diesen Haushalt repräsentierende Bezugsperson möglich.

Um den Generationenbezug zu den fünf im Zensus 2011 ermittelten Haushaltstypen herzustellen, muss in Mehr-Personen-Haushalten zunächst eine zentrale Person je Haushalt ausgewählt werden. Als zentrale, generationengebende Personen werden hier die im Rahmen der Haushaltegenerierung zur Haushaltetypisierung nach Geschlecht, Alter und Familienstand bestimmten Bezugspersonen der jeweiligen Haushalte herangezogen, sodass für die Haushaltstypen der Mehr-Personen-Haushalte mit einer Kernfamilie (*Paare ohne Kinder, Paare mit Kindern und Alleinerziehende*) eine eindeutige für den Haushalt stehende Person festgelegt werden konnte.

Eine Bezugspersonenfestlegung beim Haushaltstyp *Mehrpersonenhaushalt ohne Kernfamilie*, beispielsweise Wohngemeinschaften, ist mit diesem Ansatz nicht möglich. Daher werden diese Haushalte aus der Analysemasse ausgeschlossen.

Zusammen mit den Ein-Personen-Haushalten ergeben sich für das Land Berlin somit ca. 1,69 Mill. und für das Land Brandenburg 1,14 Mill. Bezugspersonen. Dies entspricht 51,3 % der Berliner Bevölkerung bzw. 94,1 % der 1 794 936 Haushalte. In Brandenburg stellen die Bezugspersonen einen Anteil von 46,4 % der Einwohner bzw. 96,3 % der Haushalte. Eine genaue Übersicht zeigt Tabelle 1.

Für die folgenden Analysen werden die Berliner und Brandenburger Bevölkerung zum Stichtag 9. Mai 2011 in je sechs Generationen untergliedert (vgl. Frisoli, Schmitz-Veltin 2015). Aussagen und Fallzahlen zu Mehrpersonenhaushalten mit Kernfamilien beziehen sich innerhalb dieses Beitrags stets auf die jeweilige Bezugsperson. Zur Vereinfachung der Lesbarkeit wird im weiteren Verlauf nicht zwischen *Haushalten* und *Bezugspersonen* differenziert. Es findet stets der Begriff *Haushalte* Verwendung. Des Weiteren wird bei den merkmalskombinierten Analysen die jüngste Generation Z, 1995 bis 9. Mai 2011 nicht betrachtet, da, bezogen auf die gestellte Fragestellung, zu wenig Fälle im Datensatz vorhanden sind.

Auswertung von personen-, haushalte-, gebäude- und wohnungsbezogenen Einzelmerkmalen

Tabelle 2 stellt die Einteilung der Gesamtbevölkerung beider Länder inklusive der Bezugspersonenanzahl für die jeweiligen Generationen als Basis der folgenden Einzelmerkmals- und Kombinationsanalysen dar.

Den mit 21,9 % höchsten Anteil an der Berliner Gesamtbevölkerung weist die Generation X (1966 bis 1979) auf. Die Nachkriegsgeneration ist mit 11,4 % am geringsten vertreten. Nach der Eingrenzung auf die Bezugspersonen pro Haushalt (Haushaltsbetrachtung) zeigen sich nicht nur prozentuale Änderungen, sondern auch Änderungen in der Reihenfolge der Häufigkeiten. Generation X nimmt in beiden Analysen den höchsten Anteil ein. Während bei den Bezugspersonen als nächst höhere die Generation der Babyboomer folgt, lebten zum 9. Mai 2011 anteilsmäßig mehr Personen aus der Generation Y in Berlin als aus der Generation der Babyboomer.

In Brandenburg werden die höchsten Anteile an der Gesamtbevölkerung von der Vorkriegsgeneration (vor 1945) mit 21,9 % und der Generation X (1966 bis 1979) mit 19,1 % gestellt. Wie in Berlin nimmt die in der Gesamtbevölkerung am stärksten vertretene Generation auch den höchsten Anteil (29,2 %) der Bezugspersonen ein. Des Weiteren folgen wie in Berlin die Babyboomer mit 22,6 %.

1 | Bestimmung der Bezugspersonen für haushaltebezogene Auswertungen in den Ländern Berlin und Brandenburg am 9. Mai 2011

	Berlin
3 292 365	Einwohnerzahl
– 46 000	Personen in Sonderbereichen
– 1 000	Im Ausland tätige Angehörige der Bundeswehr, der Polizeibehörden und des Auswärtigen Dienstes sowie ihre dort ansässigen Familien
– 156 000	Personen in Mehrpersonenhaushalten ohne Kernfamilie
– 676 000	Partner der Bezugspersonen
– 651 000	Nachkommen der Bezugspersonen
– 72 000	sonstig lebende Personen im Haushalt der Bezugspersonen
1 690 000	Bezugspersonen bzw. Haushalte
	Brandenburg
2 455 780	Einwohnerzahl
– 41 000	Personen in Sonderbereichen
– 5 000	Im Ausland tätige Angehörige der Bundeswehr, der Polizeibehörden und des Auswärtigen Dienstes sowie ihre dort ansässigen Familien
– 46 000	Personen in Mehrpersonenhaushalten ohne Kernfamilie
– 642 000	Partner der Bezugspersonen
– 503 000	Nachkommen der Bezugspersonen
– 80 000	sonstig lebende Personen im Haushalt der Bezugspersonen
1 139 000	Bezugspersonen bzw. Haushalte

2 | Personenfallzahlen der Generationenabgrenzung in den Ländern Berlin und Brandenburg am 9. Mai 2011

Generationenbezeichnung	Geburtsjahrgänge	Anzahl der Personen		Anzahl der Bezugspersonen	
		absolut	%	absolut	%
		Berlin			
Vorkriegsgeneration.....	vor 1945	609 000	*18,5*	407 000	*24,1*
Nachkriegsgeneration.....	1945 bis 1954	376 000	*11,4*	239 000	*14,1*
Babyboomer.....	1955 bis 1965	502 000	*15,2*	326 000	*19,3*
Generation X.....	1966 bis 1979	722 000	*21,9*	452 000	*26,8*
Generation Y.....	1980 bis 1994	638 000	*19,4*	264 000	*15,6*
Generation Z.....	1995 bis 9. Mai 2011	446 000	*13,5*	1 000	*0,1*
		Brandenburg			
Vorkriegsgeneration.....	vor 1945	538 000	*21,9*	332 000	*29,2*
Nachkriegsgeneration.....	1945 bis 1954	319 000	*13,0*	184 000	*16,2*
Babyboomer.....	1955 bis 1965	455 000	*18,5*	257 000	*22,6*
Generation X.....	1966 bis 1979	470 000	*19,1*	250 000	*22,0*
Generation Y.....	1980 bis 1994	368 000	*15,0*	115 000	*10,0*
Generation Z.....	1995 bis 9. Mai 2011	306 000	*12,5*	1 000	*0,0*

Personenbezogene Einzelmerkmalsanalyse

Die Untersuchung der 3,3 Mill. Einwohner Berlins nach den festgelegten Generationen und deren demografischen Merkmalen zeigt erwartungsgemäß ähnliche Strukturen wie die Untersuchung der Gesamtbevölkerung. So sind die Geschlechterproportionen innerhalb der Generationen, mit Ausnahme der Vorkriegsgeneration, in welcher der Anteil der Frauen 58 % beträgt, nahezu ausgeglichen. Der höhere Frauenanteil in der Vorkriegsgeneration begründet sich durch die höhere Lebenserwartung von Frauen. Ebenso spiegelt sich das erwartete Muster, dass der Anteil von verheirateten, geschiedenen und verwitweten Personen je Generation mit steigendem Alter bis zur Nachkriegsgeneration zunimmt, in den Daten wider. Der Anteil der Verheirateten in der Vorkriegsgeneration nimmt jedoch zugunsten der verwitweten Personen mortalitätsbedingt ab. Die Geschlechter- und Familienstandsverhältnisse weisen im Land Brandenburg die gleichen Strukturen auf. Bei den Ausländeranteilen in Berlin weisen die Generationen X und Y Werte von 18 % bzw. 14 % auf, wohingegen diese in den übrigen Generationen zwischen 5 % und 11 % liegen. Mit 3,3 % ist der Ausländeranteil in der Generation X im Land Brandenburg am höchsten, gefolgt von der Generation Y mit 2,8 %.

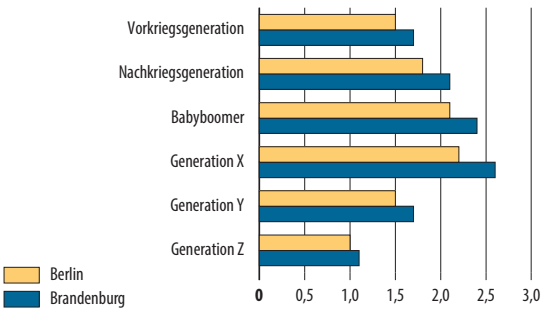
Haushaltebezogene Einzelmerkmalsanalyse

Während der Anteil der Ein-Personen-Haushalte in den Generationen der Geburtsjahrgänge 1945 bis einschließlich 1979 (Nachkriegsgeneration, Babyboomer, Generation X) circa 45 % beträgt, nimmt der Anteil der Haushalte mit drei und mehr Personen von Generation zu Generation zu (15 %, 30 %, 36 %). Im Land Brandenburg liegen die Anteile der Ein-Personen-Haushalte für die drei genannten Generationen bei jeweils rund 27 %. Hier leben die Einwohner vermehrt in Mehr-Personen-Haushalten.

Im Gegensatz hierzu leben bereits zwei Drittel der Generation Y in Berlin und 59 % in Brandenburg in Ein-Personen-Haushalten und 12 % bzw. 17 % in Haushalten mit drei und mehr Personen.

Die durchschnittliche Haushaltsgröße in den Generationen beträgt im Land Berlin 1,0 bis 2,2 Personen. Im Land Brandenburg fällt diese mit 1,1 bis 2,6 Personen etwas höher aus (vgl. Abbildung a).

a | Durchschnittliche Personenanzahl je Haushalt nach Generationen in den Ländern Berlin und Brandenburg am 9. Mai 2011



Die Ein-Personen-Haushalte weisen in Berlin in allen Generationen den höchsten Anteil bei den Haushaltstypen auf. Paare ohne Kinder sind in der Vorkriegs-, Nachkriegsgeneration und Generation Y am zweithäufigsten vertreten (36,9 %, 35,5 % und 16,0 %), während es bei den Babyboomern und der Generation X die Paare mit Kindern sind. Die Daten im Land Brandenburg zeigen ein entgegengesetztes Bild. Hier sind bei den Generationen der Babyboomer (35,2 %) und der Generation X (46,1%) die Paare mit Kindern am häufigsten vertreten. Mit knapp der Hälfte stellen Paare ohne Kinder in der Vor- und Nachkriegsgeneration den anteilsmäßig häufigsten Haushaltstyp.

Da die Existenz eigener Kinder im Rahmen des Zensus 2011 nicht umfassend erhoben wurde, können nur Aussagen zu am Stichtag 9. Mai 2011 im Haushalt lebenden Kindern getroffen werden. Haushalte mit der Merkmalsausprägung *Paare ohne Kind(er)* haben demnach nicht zwangsläufig keine Kinder, es leben lediglich keine Kinder in den entsprechenden Haushalten, da sie beispielsweise ausgezogen sein können. Eine Analyse des zugrundeliegenden Merkmals *Typ des privaten Haushalts (nach Familien)* in Bezug auf die gesamtfamiliären Zusammenhänge, wie z. B. die Anzahl der geborenen Kinder oder das Wanderungsverhalten derselben, ist aufgrund dessen nicht möglich.

Der Anteil der Berliner und Brandenburger Haushalte, in denen alle Personen einen Migrationshintergrund aufweisen, nimmt mit Verjüngung der Generationen zu (vgl. Tabelle 3). Den höchsten Anteil stellen dennoch die Haushalte, in denen keine Person einen Migrationshintergrund besitzt.

Gebäude- und wohnungsbezogene Einzelmerkmalsanalyse

Für Wohngebäude (ohne Wohnheime) und sonstige Gebäude mit Wohnraum in Berlin ergibt sich das Durchschnittsbaujahr 1955 mit durchschnittlich sechs Wohnungen je Gebäude. In Brandenburg sind die Gebäude durchschnittlich ein Jahr jünger und verfügen im Mittel über zwei Wohnungen.

Während die Haushalte, deren Bezugspersonen zum Zensusstichtag 17 bis 31 Jahre alt waren (Generation Y), in Berlin auf durchschnittlich 57,5 m^2 Wohnfläche lebten, lag dieser Wert für Brandenburg bei 61,4 m^2. Die Haushalte der übrigen Generationen verfügten im Mittel über 75,8 m^2 (Berlin) bzw. 88,7 m^2 (Brandenburg) Wohnfläche. Die Babyboomer lebten in beiden Ländern mit 78,0 m^2 bzw. 94,9 m^2 Durchschnittswohnfläche in den größten Wohnungen. Dieser Unterschied zeigt sich auch in der durchschnittlichen Raumanzahl der Haushalte, die in Berlin zwischen 2,9 Räumen (Generation Y) und 3,6 Räumen (ältere Generationen) sowie in Brandenburg zwischen 2,9 und 4,5 Räumen lag. In beiden Ländern lebten die Babyboomer mit durchschnittlich 3,7 bzw. 4,5 Räumen am großzügigsten.

Die Unterscheidung der vom Eigentümer bewohnten und zu Wohnzwecken vermieteten Wohnungen zeigt, dass der Eigentümeranteil in den Berliner Haushalten der ältesten Generationen (Vor- und

Nachkriegsgeneration) bei rd. 21 % liegt und mit jeder jüngeren Generation kontinuierlich abnimmt (18,7 %, 11,8 %, 4,9 %). Dieses bestätigt die Charakterisierung Berlins als *Mieterstadt*.

Anders verhält es sich in Brandenburg. Die höchsten Eigentümeranteile mit 55,9 % und 55,7 % finden sich innerhalb der Nachkriegsgeneration sowie der Babyboomer; die Vorkriegsgeneration und die Generation X weisen mit je rd. 43 % nahezu gleiche Eigentümeranteile auf.

Die Analyse von Einzelmerkmalen lieferte eine Übersicht über erste Ergebnisse. Diese Ergebnisse bestätigen durchaus erwartete Zusammenhänge.

Auswertung kombinierter Einzelmerkmale am Beispiel der Erstbezugseigentümer

Anhand einer spezifischen Fragestellung soll die nur im Zensus 2011 mögliche Kombinationsanalyse der verschiedenen Teilerhebungen dargestellt werden. Hierbei muss aber beachtet werden, dass diese nur unter gewissen Einschränkungen möglich ist. In Bezug auf das Thema dieses Beitrags wurde folgende Fragestellung für eine tiefergehende Analyse ausgewählt:

In welchem Umfang haben die verschiedenen Generationen in Berlin und Brandenburg Wohneigentum (Eigentumswohnungen oder Einfamilienhäuser) bei gleichzeitigem Erstbezug erworben und leben bis heute (Zensusstichtag) darin? Im Folgenden werden diese *Erstbezugseigentümer* genannt.

Um diese Frage zu beantworten, müssen folgende Einschränkungen vorgenommen werden:

- Bestimmung der Bezugspersonen zwecks Haushaltegenerationenfestlegung zur Schaffung einer neuen Grundgesamtheit für eine haushalte- und generationenbezogene Analyse. Reduktion:
  → der Berliner Bevölkerung um 48,7 % (vgl. Tab. 1),
  → der Brandenburger Bevölkerung um 53,6 %.
- Selektion der Bezugspersonen nach vom Eigentümer bewohnten Wohnungen. Es verbleiben:
  → in Berlin ca. 264 000 Haushalte,
  → in Brandenburg ca. 509 000Haushalte.
- Abgleich von Wohnungsbezugsdatum und Gebäudebaujahr. Analysemasse:
  → in Berlin ca. 55 000 Haushalte,
  → in Brandenburg ca. 138 000 Haushalte.

Bei der Auswertung treten unter Umständen Unplausibilitäten zwischen Bezugsdaten und Baujahren auf, wie beispielsweise fehlende Melderegisterangaben zum Wohnungsbezugsdatum, von Auskunftspflichtigen geschätzte Gebäudebaujahre oder fehlende Gemeinde- und Wohnungsbezugsdaten bei Fehlbestandspersonen. Dies führt zu einer Untererfassung der Analysemasse, welche auch mittels weiterführender methodischer Schritte – hier u. a. einer festgelegten Toleranz von Bezugsdatum und Baujahr von bis zu einem Jahr – nicht vermieden werden kann. Dennoch handelt es sich bei diesen Ergebnissen um einzigartige Informationen, die ausschließlich aus den Ergebnissen des Zensus 2011 ermittelt werden können.

Bei Erstbezugseigentümern im Sinne dieses Beitrags muss es sich trotz annähernder Gleichheit von Gebäudebaujahr und Wohnungsbezugsdatum bei gleichzeitigem Vorliegen der Wohnungsnutzungsausprägung *vom Eigentümer bewohnt* nicht notwendigerweise um die direkt nach Bezugsfertigstellung des Gebäudes im Grundbuch eingetragenen Personen handeln. In die Gruppe der Erstbezugseigentümer fallen neben Erben beispielsweise auch ehemalige Mieter, die das Objekt, in dem sie seit Bezugsfertigstellung lebten, im Laufe ihrer Mietdauer vom ursprünglichen Eigentümer erworben haben.

Abbildung b stellt die Anteile der Erstbezugseigentümer an allen Eigentümern dar. In Berlin sind der durchschnittlich höhere Anteil von Erstbezugseigentümern im ehemaligen Berlin-Ost (Bezirke Marzahn-Hellersdorf, Lichtenberg, Treptow-Köpenick und Pankow) sowie in den westlichen und nord-westlichen Außenbereichen (Bezirke Spandau und Reinickendorf) auffällig. Die innenstadtnahen Prognoseräume in Berlin-West weisen insbesondere im Bezirk Charlottenburg-Wilmersdorf die geringsten Werte auf.

3 | Migrationsstatus der privaten Haushalte nach Generationen in den Ländern Berlin und Brandenburg am 9. Mai 2011

Haushalt	Anteil je Generation in %					
	Vorkriegs-generation	Nachkriegs-generation	Baby-boomer	Generation X	Generation Y	Generation Z
			Berlin			
alle Personen mit Migrationshintergrund.....	*7,6*	*14,6*	*16,8*	*20,8*	*18*	*41,6*
Personen zum Teil mit Migrationshintergrund.....	*3,7*	*5,9*	*6*	*8,2*	*5,2*	*0,5*
alle Personen ohne Migrationshintergrund....	*88,7*	*79,5*	*77,3*	*71*	*76,8*	*57,8*
			Brandenburg			
alle Personen mit Migrationshintergrund.....	*2,9*	*2,2*	*2,3*	*3,1*	*3,7*	*8,0*
Personen zum Teil mit Migrationshintergrund.....	*4,2*	*4,1*	*3,4*	*3,3*	*1,5*	–
alle Personen ohne Migrationshintergrund....	*92,9*	*93,8*	*94,3*	*93,6*	*94,7*	*92,0*

Die höchsten Anteilswerte der Erstbezugseigentümer im Land Brandenburg weisen die kreisfreien Städte Frankfurt (Oder) und Cottbus mit 43,7 % und 42,6 % auf. Die Landeshauptstadt Potsdam bewegt sich mit 27,6 % im Landesdurchschnitt. Anteilswerte von 35 und mehr Prozent konzentrieren sich im Berliner Umland. Im Weiteren Metropolenraum treten diese höheren Anteilswerte in der Gemeinde Schwedt/Oder sowie im Umland von Cottbus und Frankfurt (Oder) auf. Bezogen auf die einzelnen Generationen besitzt die Generation Babyboomer in Berlin mit ca. 16 000 (29,9 %) Haushalten den höchsten absoluten Anteil an Eigentumswohnungen bzw. Einfamilienhäusern. 15 000, 12 000 bzw. 11 000 Haushalte der Vorkriegs-, Nachkriegsgeneration bzw. Generation X besaßen am Zensusstichtag Wohneigentum. Etwa 1 % der Erstbezugseigentümer gehörte bereits zur Generation Y. Auch in Brandenburg stellen die Babyboomer mit ca. 46 000 Haushalten (32,9 %) den höchsten Generationenanteil innerhalb der Erstbezugseigentümer, gefolgt von der Generation X (36 000 Haushalte bzw. 26,0 %) und der Nachkriegsgeneration (28 000 Haushalte bzw. 20,5 %).

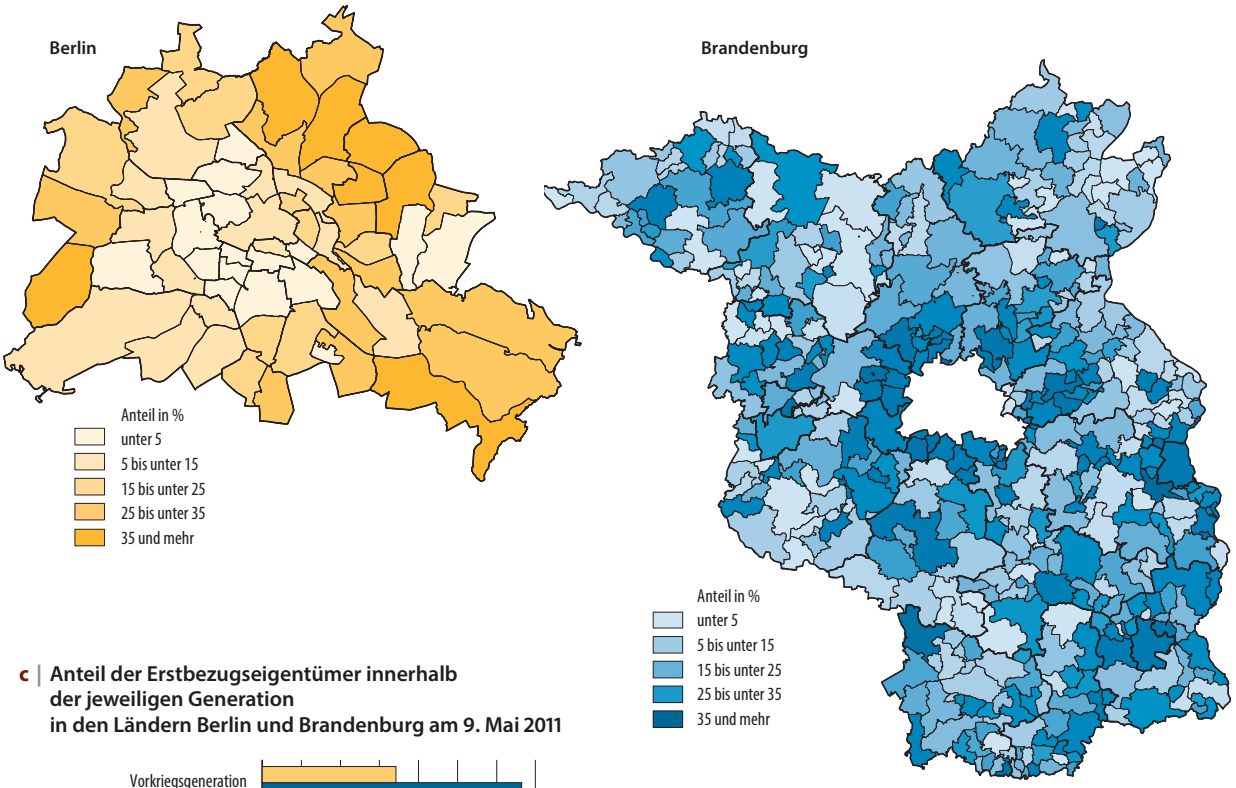
Es zeigt sich, dass der Anteil der Erstbezugseigentümer innerhalb der Generation der Babyboomer, die am 9. Mai 2011 in vom Eigentümer bewohnten Wohnungen lebten, mit 26,7 % am höchsten ausfällt, gefolgt von der Nachkriegsgeneration mit 23,9 % (vgl. Abbildung c). In Brandenburg ist dieser Anteil in der Vorkriegsgeneration mit 33,2 % am höchsten, gefolgt von den Babyboomern mit 26,2 %.

Abbildung d skizziert den Anteil der Erstbezugseigentümer je Generation nach Baujahr des Wohneigentums ab 1920 für die Vorkriegsgeneration bis zur Generation Y. Zwecks besserer Veranschaulichung sind die Jahreseinzelwerte jeder Generation durch Linien miteinander verbunden. Dabei zeigt sich ein Anstieg des Wohneigentumsumfangs von Erstbezugseigentümern ab ungefähr den 1970er Jahren. Eine detaillierte Darstellung dieser Baujahre bzw. Erstbezugsdaten ist aus Abbildung e ersichtlich.

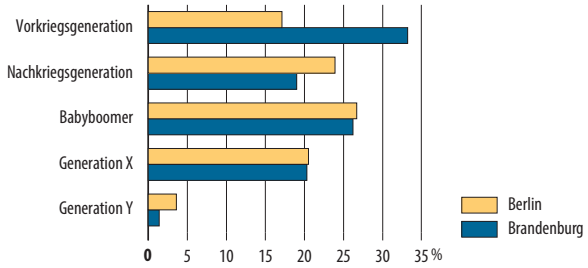
Die nachfolgenden strukturellen Auffälligkeiten sind in den Ländern Berlin und Brandenburg gleichermaßen zu beobachten:

Sowohl bei der Vorkriegs- als auch bei der Nachkriegsgeneration sind zwei Maximalwerte erkennbar. Diese erstrecken sich von Mitte bis Ende der 1970er sowie Mitte bis Ende der 1990er Jahre (Vorkriegsgeneration) bzw. von Mitte bis Ende der 1980er und erneut Mitte bis Ende der 1990er Jahre (Nachkriegsgeneration). Diese Höchstwerte der

b | Anteil der Erstbezugseigentümer an allen Eigentümern in den Ländern Berlin und Brandenburg am 9. Mai 2011



c | Anteil der Erstbezugseigentümer innerhalb der jeweiligen Generation in den Ländern Berlin und Brandenburg am 9. Mai 2011



Erstbezugseigentümeranteile von Mitte bis Ende der 1990er Jahre sind ebenfalls innerhalb der Babyboomer-Generation vorhanden. Ein ggf. eintretendes zweites Maximum innerhalb der Babyboomer-Generation ist mit den aktuell vorliegenden Daten (noch) nicht ablesbar und kann, sofern sich hier ein Muster zeigt, frühestens mit Vorliegen der Ergebnisse des Zensus 2021 nachgewiesen werden.

Innerhalb der Generation Y beginnt der Anstieg zum ersten Höchstwert der Erstbezugseigentümeranteile bereits 30 Jahre nach der unteren Altersgruppengrenze, wohingegen innerhalb der älteren Generationen die ersten Anstiege erst nach jeweils ca. 40 Jahren feststellbar sind.

Der überdurchschnittlich hohe Anteil der Erstbezugseigentümer innerhalb der Generation Y ab Mitte der 2000er Jahre ist auf den gegenüber den älteren Generationen deutlich kürzeren Existenzzeitraum dieser Generation (seit 1980) zurückzuführen. Während sich die prozentualen Jahreshäufigkeiten der Erstbezugseigentümer, ausgehend vom Zensusstichtag, von Generation X bis zur Vorkriegsgeneration auf bis zu 45, 56, 66 und mehr als 66 Jahre verteilen, führt der eingeschränkte Darstellungszeitraum der Generation Y (31 Jahre) grundsätzlich zu höheren Jahresanteilswerten der Erstbezugseigentümer.

Da derartige Jahreshöchstwerte der Erstbezugseigentümeranteile ausgehend von der Vorkriegsgeneration erst seit Mitte der 1970er Jahre zu verzeichnen sind, tritt dieser stets sichtbare Effekt bei gleichzeitiger und relativer Darstellung aller Generationen seit einschließlich 1970 auch dann auf, wenn die Summe der Erstbezugseigentümer jeder Generation ab einschließlich 1970 auf 100 % festgesetzt wird.

Eine von diesem Phänomen bereinigte Darstellung kann beispielsweise mit Vorliegen von Ergebnissen des Zensus 2031 erfolgen.

Die kombinierte Analyse der Zensusmerkmale zeigt, dass die Erstbezugseigentümer in zwei Phasen des Lebens vermehrt Wohneigentum erwerben. Dieses könnte auch auf Wohnbauförderungen innerhalb dieser Jahre zurückzuführen sein.

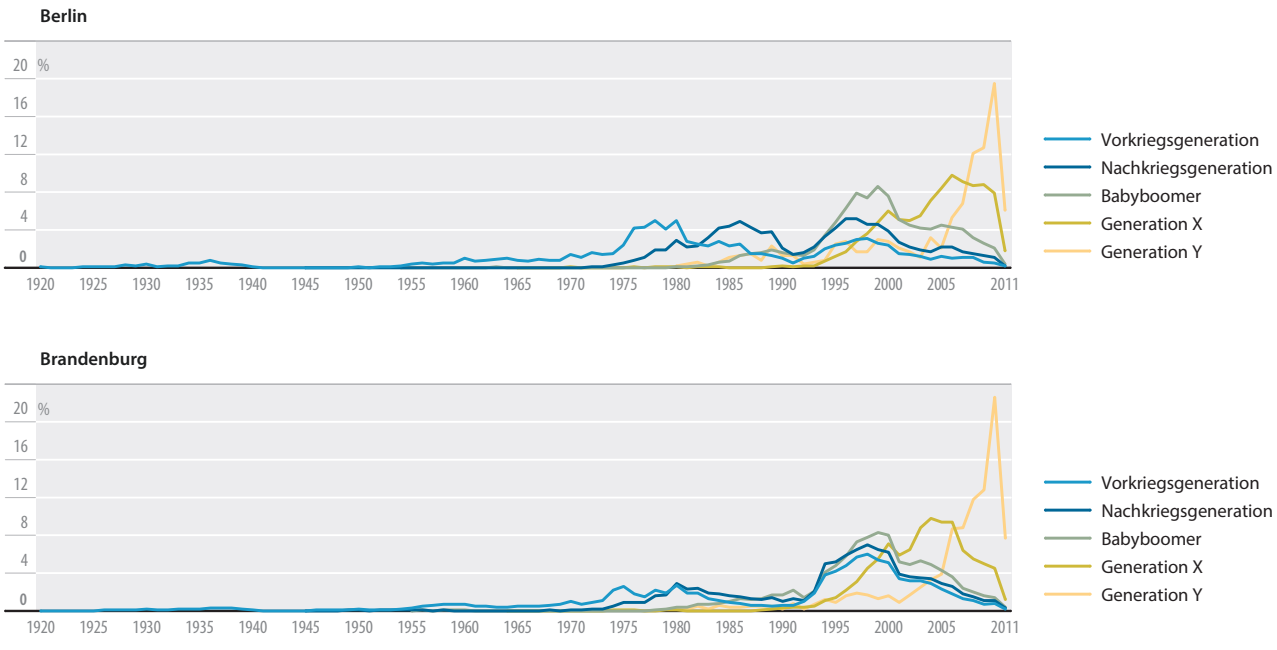
Das Durchschnittsalter der Erstbezugseigentümer in Berlin nimmt ausgehend von der Vorkriegsgeneration bis zu den Babyboomern mit 45, 37 bzw. 31 Jahren stetig ab. Diese Entwicklung ist ebenso in Brandenburg festzustellen, wobei die Altersdurchschnitte der genannten Generationen um jeweils vier Jahre niedriger ausfallen. Es kann somit festgehalten werden, dass der Eigentumserstbezug in Brandenburg in durchschnittlich jüngeren Lebensjahren erfolgt.

Abbildung f stellt die Anteile der Erstbezugseigentümer nach Alter und Generation dar. Wie schon in Abbildung e führt der zunehmend eingeschränkte Darstellungszeitraum der jüngeren Generationen grundsätzlich zu höheren Jahresanteilswerten. Dennoch sind ausgehend von der Vorkriegsgeneration bis hin zur Generation X klare Höchstwerte in den Altersjahren zwischen 35 und 40 Jahren erkennbar. Etwa zehn Jahre nach den ersten Anteilshöchstwerten zeigen sich bei der Vorkriegs- sowie der Nachkriegsgeneration weitere Anteilsspitzen.

In Brandenburg hat ein überdurchschnittlicher Anteil der Vorkriegsgeneration ebenfalls in einem Alter von 35 bis 40 Jahren neu errichtetes Wohneigentum erworben und bezogen. Die zweite Anteilsspitze zeigt sich hier jedoch erst rund 20 Jahre später. Auch die Nachkriegsgenerationen zeigen zwei Maximalwerte in unterschiedlichen Altersjahren, wobei die jüngeren Erstbezugseigentümer bereits mit Anfang 30 Immobilien besitzen.

Die Babyboomer und die Generation X weisen bis einschließlich Zensusstichtag jeweils nur ein Alters-

d | Anteil der Erstbezugseigentümer je Generation nach Baujahr des Wohneigentums in Berlin und im Land Brandenburg am 9. Mai 2011



maximum (38 Jahre bzw. 32 Jahre) bei Erstbezug innerhalb der jeweiligen Generation auf. Generation Y war zum Zensusstichtag maximal 31 Jahre alt und bleibt daher in dieser Altersauswertung unberücksichtigt.

Mithilfe von richtig kombinierten Merkmalen können somit – eventuell neue – Erkenntnisse von Zusammenhängen oder Strukturen auf Basis des Zensus 2011 gewonnen werden. Eine Ursachenanalyse ist aber weiterhin nur durch externe Analysen möglich.

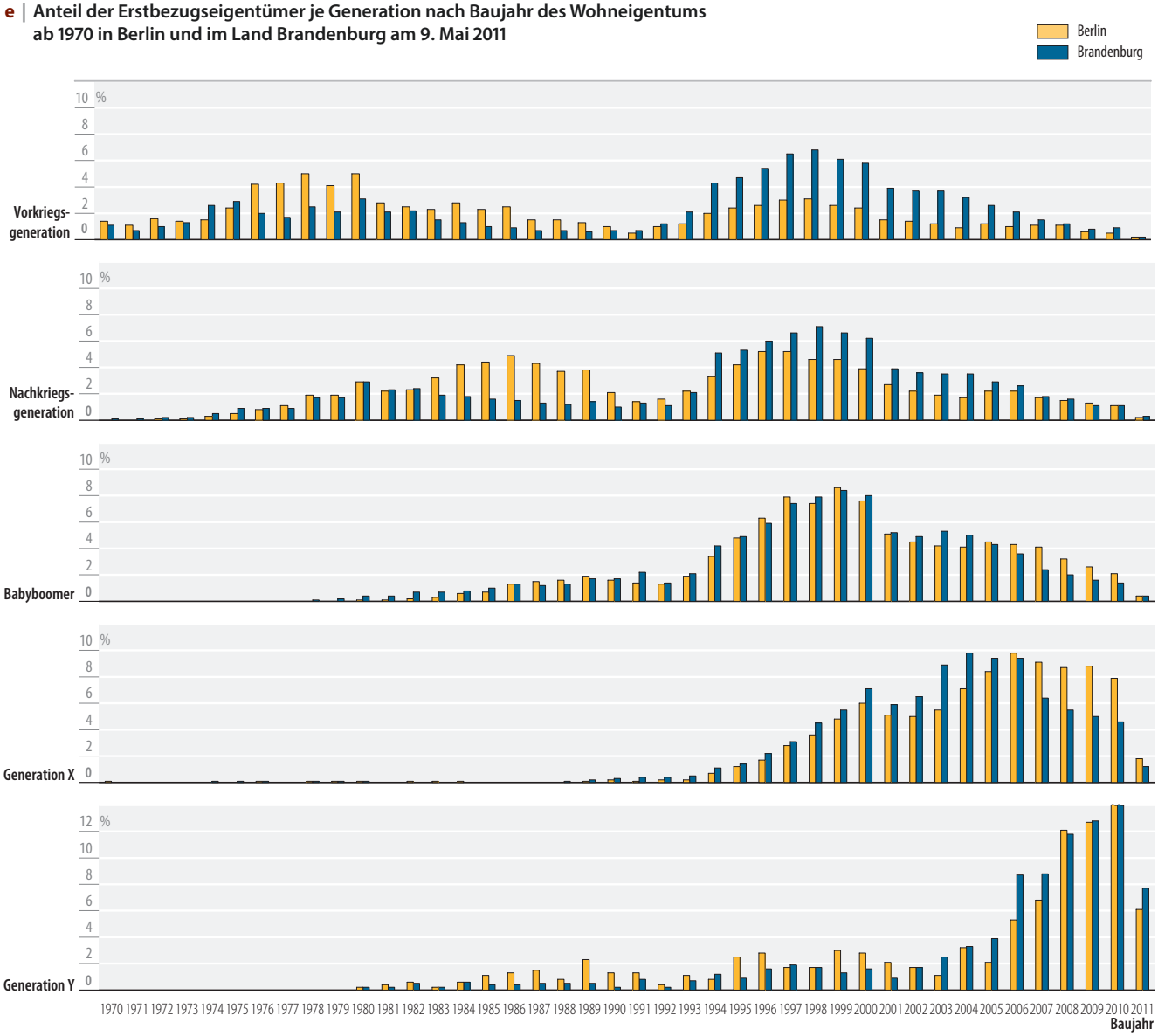
Fazit

Mit den im Zensus 2011 erhobenen Merkmalen werden bereits bekannte Zusammenhänge, Strukturen und Erwartungen in den Themenfeldern Bevölkerung, Haushalte und Wohnen bestätigt. Neue Erkenntnisse aus Einzelmerkmalen können aus dem jetzigen Merkmalskranz kaum abgeleitet werden. Zusätzlich stellt die vorliegende Ergebnisdatenstruktur des Zensus 2011 eine Erschwernis dar, sodass kombinierte Auswertungen meist sehr komplex werden und mit Einschränkungen einhergehen.

Für eine Aufdeckung neuer Erkenntnisse oder bisher unbekannter Strukturen ist eine nachhaltige Auswahl und Festlegung von Erhebungsmerkmalen sowie Sicherstellung der Kombinierbarkeit derselben im Vorfeld des anstehenden Zensus 2021 erforderlich. Auch die Anfragen von Datennutzerinnen und -nutzern aus Wissenschaft und Verwaltung zeigen, dass das bisherige Merkmalsspektrum zu eingeschränkt ist und einer Erweiterung bedarf, um die jeweilig formulierten Fragestellungen beantworten zu können. Hierzu werden Merkmale wie beispielsweise energetische Gebäudeinformationen, Leerstandsgrund und -dauer, Geschossanzahl oder Barrierefreiheit (GWZ) bzw. Einkommen, Kinderanzahl oder Pendlerbewegungen innerhalb Berlins (Haushaltebefragung) benötigt. Des Weiteren weisen die Anfragen darauf hin, dass Definitionen, Merkmalsabgrenzungen und -beschreibungen missverstanden werden sowie Ergebnisdaten bedarfsgerechter anpassbar und flexibler auswertbar werden müssen.

Beispiele hierfür sind u. a. das Merkmal *Typ des privaten Haushalts (nach Familien)*, aus dessen Ausprägungen sich keine Rückschlüsse auf die Anzahl

e | Anteil der Erstbezugseigentümer je Generation nach Baujahr des Wohneigentums ab 1970 in Berlin und im Land Brandenburg am 9. Mai 2011



der geborenen Kinder oder das Vorhandensein mehrerer Familien innerhalb eines Haushalts ziehen lassen, oder das Merkmal *Zahl der Räume*, das aufgrund seiner Definition nicht dem allgemeinen Sprachverständnis entspricht, da eine Küche seitens der Auskunftspflichtigen sowie Datennutzerinnen und -nutzer im Regelfall nicht als separater Raum aufgefasst wird.

Die immer noch anhaltende Nachfrage nach Sonderauswertungen der Ergebnisse des Zensus 2011 seitens Politik, Wissenschaft, Senatsverwaltung, Wohnungswirtschaft und Forschung belegt die immense Bedeutung kleinräumig vorliegender Personen-, Gebäude- Wohnungs- und Haushaltsdaten für Stadtentwicklungs-, Umwelt- und Bedarfsplanungen, Wohn- und Infrastrukturuntersuchungen, Klima- und Machbarkeitsstudien, Entwicklung energetischer Quartierskonzepte, die Erstellung von Emissionskatastern und Umweltatlanten, Gutachten zur Mietwohnungssituation nebst wohnungspolitischen Diskussionen, die Erstellung von Mietspiegeln, Wahlanalysen sowie nicht zuletzt für Studien- und Diplomarbeiten. Neben der Ermittlung der Einwohnerzahlen, die für eine Vielzahl von Rechtsvorschriften in Deutschland und beispielsweise für Ausgleichszahlungen zwischen den Ländern (Länderfinanzausgleich), für kommunale Planungen oder bei der Einteilung der Wahlkreise benötigt werden, bilden die Ergebnisse der GWZ 2011 die wesentliche Basis für die Neuberechnung der Bruttowertschöpfung der Wohnungsvermietung in den Volkswirtschaftlichen Gesamtrechnungen (VGR).

Die amtliche Statistik sieht sich als Dienstleister, der, durch bundes- und landesgesetzliche Aufgaben legitimiert, qualitativ hochwertige Daten bereitstellt. Die Wertung und Bewertung dieser Daten ist jedoch Aufgabe von Politik und Wissenschaft. Anfragen belegen allerdings, dass eben dieses für die Ergebnisse des Zensus 2011 nur eingeschränkt bzw. nur durch komplexe, mit Einschränkungen behaftete Analysen möglich ist. Der Ausgestaltung zukünftiger Zensen kommt somit im Hinblick auf Datenqualität und Merkmalsspektrum eine besondere Bedeutung zu. Hierbei ist es nicht Aufgabe der amtlichen Statistik, über die gesetzliche Zulässigkeit notwendiger Nachjustierungen zu befinden, sondern konkrete und abgestimmte fachliche Anforderungen zu definieren, die dann ggf. nach Prüfung durch Parlamentarier und Datenschützer ihren Weg in die Zensusgesetze 2021 finden.

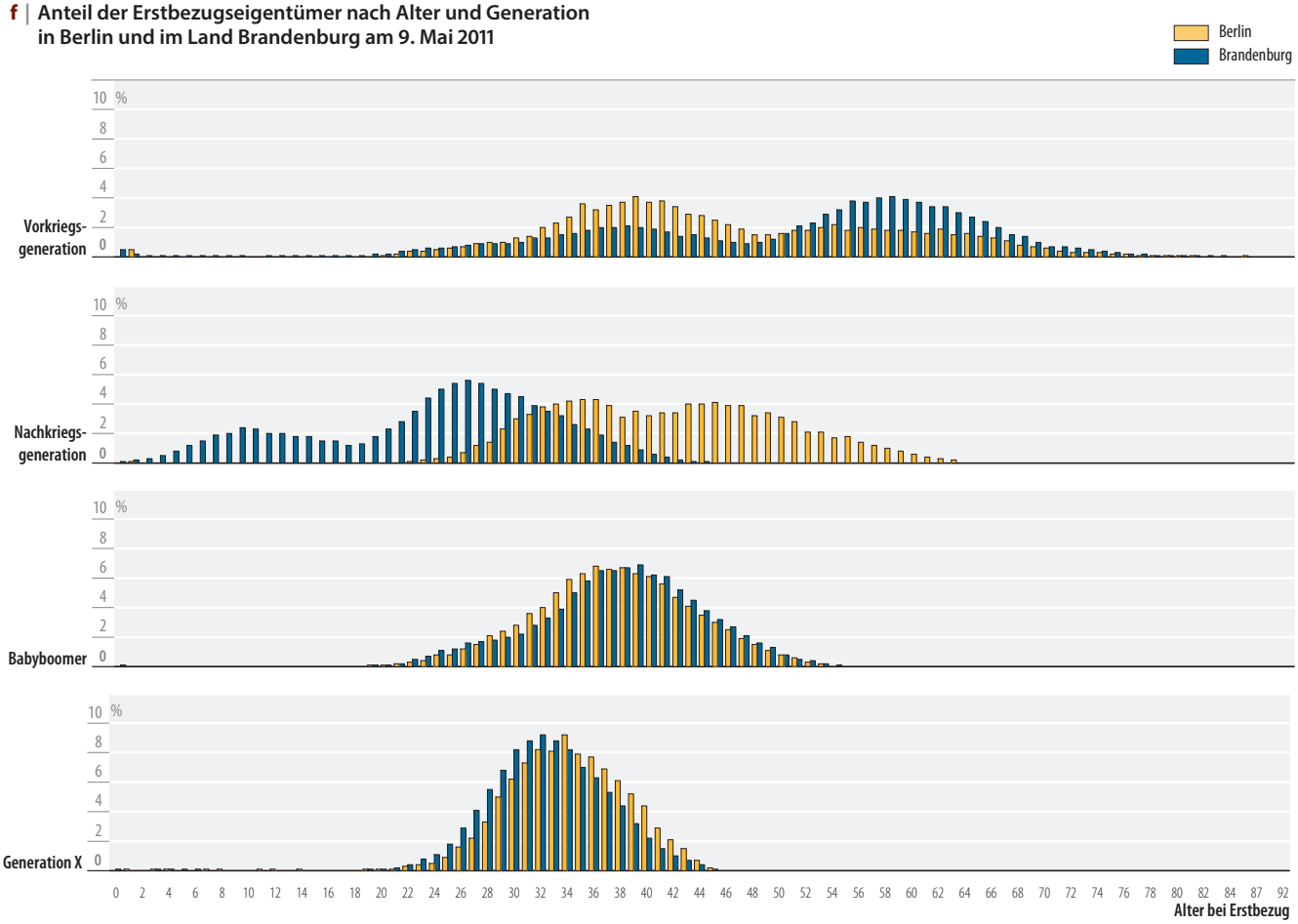
Verena Kutzki, Diplom-Volkswirtin und Master of Science, ist wissenschaftliche Mitarbeiterin im Referat *Zensus* des Amtes für Statistik Berlin-Brandenburg. Zuvor war sie Mitarbeiterin des Instituts für Wirtschaftsforschung Halle (IWH).

Marco Schwarz, Diplom-Geograph, ist seit 2008 im Referat *Zensus* des Amtes für Statistik Berlin-Brandenburg als Referent, GWZ-Teilprojektleiter und wissenschaftlicher Mitarbeiter tätig.

Literatur

- Frisoli, Pasquale; Schmitz-Veltin, Ansgar (2015): Abgrenzung und demografische Analyse von Generationen: Herausforderungen für das Informationsmanagement. In: Stadtforschung und Statistik 1/2015, S. 36–42.

f | Anteil der Erstbezugseigentümer nach Alter und Generation in Berlin und im Land Brandenburg am 9. Mai 2011



Wirtschaft

Unternehmen und Betriebe

– Entwicklung in Berlin und Brandenburg

von **Thomas Heymann**

In diesem Beitrag wird die Entwicklung von Unternehmen und Betrieben in Berlin und Brandenburg vorgestellt, wie sie sich durch die Meldungen Gewerbetreibender in der Gewerbeanzeigenstatistik und Meldungen der Insolvenzgerichte in der Insolvenzstatistik abzeichnet. Die Zeitreihen beginnen im Januar 2004.

Die Gewerbeanzeigenstatistik

liefert monatlich Informationen über die Zahl der Gewerbean- und -abmeldungen nach Wirtschaftsbereichen, Rechtsformen und Veränderungen im Lebenszyklus der Gewerbebetriebe. Außerdem werden Geschlecht und Staatsangehörigkeit der Gewerbetreibenden ermittelt. Die An- und Abmeldungen werden danach unterschieden, welche Gründe maßgeblich waren. Abbildung a zeigt den Aufbau und Zusammenhang der wesentlichen Formen der An- und Abmeldung eines Gewerbes.

Ein Gewerbe ist jede erlaubte, selbstständige, nach außen erkennbare Tätigkeit, die planmäßig, für eine gewisse Dauer und zum Zwecke der Gewinnerzielung ausgeübt wird und kein freier Beruf ist. Die *Gewerbeanmeldung* ist erforderlich bei der Neugründung eines Betriebes, der Wiedereröffnung eines Betriebes nach Verlegung (Zuzug), der Gründung eines Betriebes nach dem Umwandlungsgesetz, der Änderung der Rechtsform, dem Eintritt von Gesellschaftern und der Übernahme durch Erbfolge, Kauf oder Pacht eines Betriebes. Dem steht die *Gewerbeabmeldung* gegenüber, die bei einer vollständigen Aufgabe eines Betriebes, der Verlagerung eines Betriebes in ein anderes Bundesland (Fortzug), der Abmeldung eines Betriebes nach dem Umwandlungsgesetz, der Änderung der Rechtsform, dem Austritt von Gesellschaftern und der Übergabe durch Erbfolge, Kauf oder Pacht eines Betriebes vorgeschrieben ist.

Die Zeitreihe der Abbildung b stellt die Entwicklung des Gründungsgeschehens seit 2004 dar. Unter einer Betriebsgründung wird die Gründung einer Haupt- bzw. Zweigniederlassung oder unselbstständigen Zweigstelle durch eine natürliche oder juristische Person verstanden, die entweder im Handels-, Vereins- oder Genossenschaftsregister eingetragen ist oder die Handwerkseigenschaft besitzt oder mindestens einen Arbeitnehmer beschäftigt. Entsprechende Bedingungen müssen ebenfalls bei einer vollständigen Aufgabe eines Betriebes erfüllt sein.

a | Gewerbemeldung mit den wichtigsten Formen der meldepflichtigen Veränderungen des Gewerbes

Region	Gewerbeanmeldungen									
	Insgesamt	Neuerrichtungen						Umwandlungen	Übernahmen	Zuzüge
		insgesamt	Neugründungen				sonstige Neugründungen			
			Zusammen	Betriebsgründungen						
				zusammen	Hauptniederlassungen	Zweigniederlassungen				

Region	Gewerbeabmeldungen									
	Insgesamt	Gewerbeaufgaben						Umwandlungen	Übernahmen	Fortzüge
		insgesamt	Vollständige Aufgaben				sonstige Stilllegungen			
			Zusammen	Betriebsaufgaben						
				zusammen	Hauptniederlassungen	Zweigniederlassungen				

Die absoluten Werte des Gründungsgeschehens in Berlin und Brandenburg verdeutlicht die Dominanz des Standortes Berlin für die Metropolregion. Ein Trend sowohl für Brandenburg mit seinen peripheren ruralen Räumen als auch für die Metropolregion Berlin einschließlich des urbanen Gürtels, der sich im Zentrum Brandenburgs herausbildet, lässt sich nur schwer herauslesen (Abbildung b).

Hier kann der Saldo von Betriebsgründungen und -aufgaben je Bundesland und Monat helfen (grau hinterlegte Kurven der Abbildung c). Die Fluktuationen im Verlauf eines Jahres haben erstens saisonale Gründe und zweitens sind sie zum Jahresende auf Löschungen „von Amts wegen" zurückzuführen. Für eine anschauliche Ansicht möglicher Trends wurde für jeden Monat ein Mittelwert der jeweils zwölf vorherigen Monate errechnet (gleitender Mittelwert mit einer Periode von 12 Monaten).

Für Berlin ist über den gesamten Zeitraum ein wachsender und für Brandenburg ein abnehmender Trend zu konstatieren, wobei für beide Länder der durchschnittliche Saldo immer noch positiv ist, das heißt die Betriebsgründungen überwiegen die Betriebsaufgaben. 2013 war für Brandenburg eine zunehmende Tendenz feststellbar, die sich 2014 nicht fortsetzte, sondern wie im Jahr 2012 wieder abnahm. Für Berlin scheint sich der Trend um das Niveau des Saldos von durchschnittlich 150 Gründungen herauszubilden. Möglicherweise beeinflussen die Löschungen „von Amts wegen" in Brandenburg den Trend stärker als in Berlin.

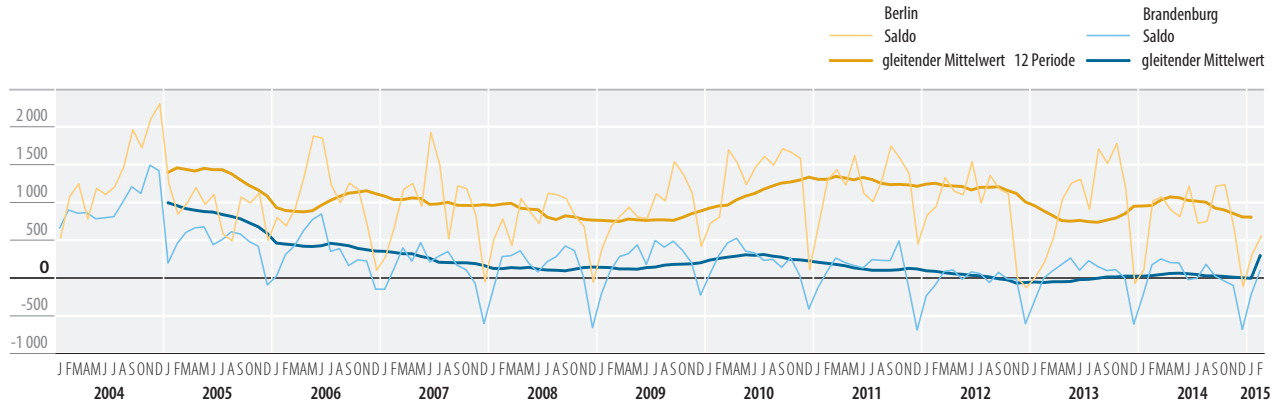
Die Insolvenzstatistik der eröffneten Verfahren

ist eine monatliche Statistik. Aufgabe dieser Statistik ist es, über die Situation von überschuldeten oder zahlungsunfähigen juristischen oder natürlichen Personen, deren Fälle vor Gericht verhandelt werden, zu berichten und den möglichen volkswirtschaftlichen Schaden zu beschreiben. Weiterhin wird die Insolvenzstatistik dazu herangezogen, die Effizienz des Insolvenzrechts zu bewerten.

Die Insolvenzordnung definiert mehrere Schuldnergruppen, wie z. B. Unternehmen, ehemals selbstständig Tätige und Verbraucher, die einem von zwei Verfahrensabläufen zugeordnet werden: Regel- oder einfaches Insolvenzverfahren. Vor Eröffnung eines Verfahrens wird geprüft, wie groß die Chancen sind, den Forderungen der Gläubiger nachzukommen. Können nicht einmal die Verfahrenskosten aufgebracht werden, wird der Antrag auf Eröffnung eines Insolvenzverfahrens „mangels Masse" abgewiesen, also nicht eröffnet. In Abbildung d wird die Häufigkeit der eröffneten Verfahren in Monatsschritten für jede der oben genannten Schuldnergruppen aufgezeigt.

b | **Betriebsgründungen und -aufgaben in den Ländern Berlin und Brandenburg seit 2004**

c | **Saldo der Gewerbean- und -abmeldungen in den Ländern Berlin und Brandenburg seit 2004**



Die dünnen Linien in der Grafik der eröffneten Verfahren von *Verbrauchern* zeigt den absoluten Kurvenverlauf. Mit der Methode des gleitendes Mittels (Periode: 6 Monate) werden die Verläufe auch bei den beiden anderen Grafiken ein wenig geglättet. Die von der Verfahrensanzahl her größte Schuldnergruppe der überschuldeten privaten Haushalte (Verbraucher) wurde mit dem Inkrafttreten der Insolvenzordnung vor 15 Jahren eingeführt. Daher nahm die Anzahl der eröffneten Verfahren in Berlin und Brandenburg bis 2007 kontinuierlich zu. Seit 2008 ist hier ein langsamer Rückgang festzustellen.

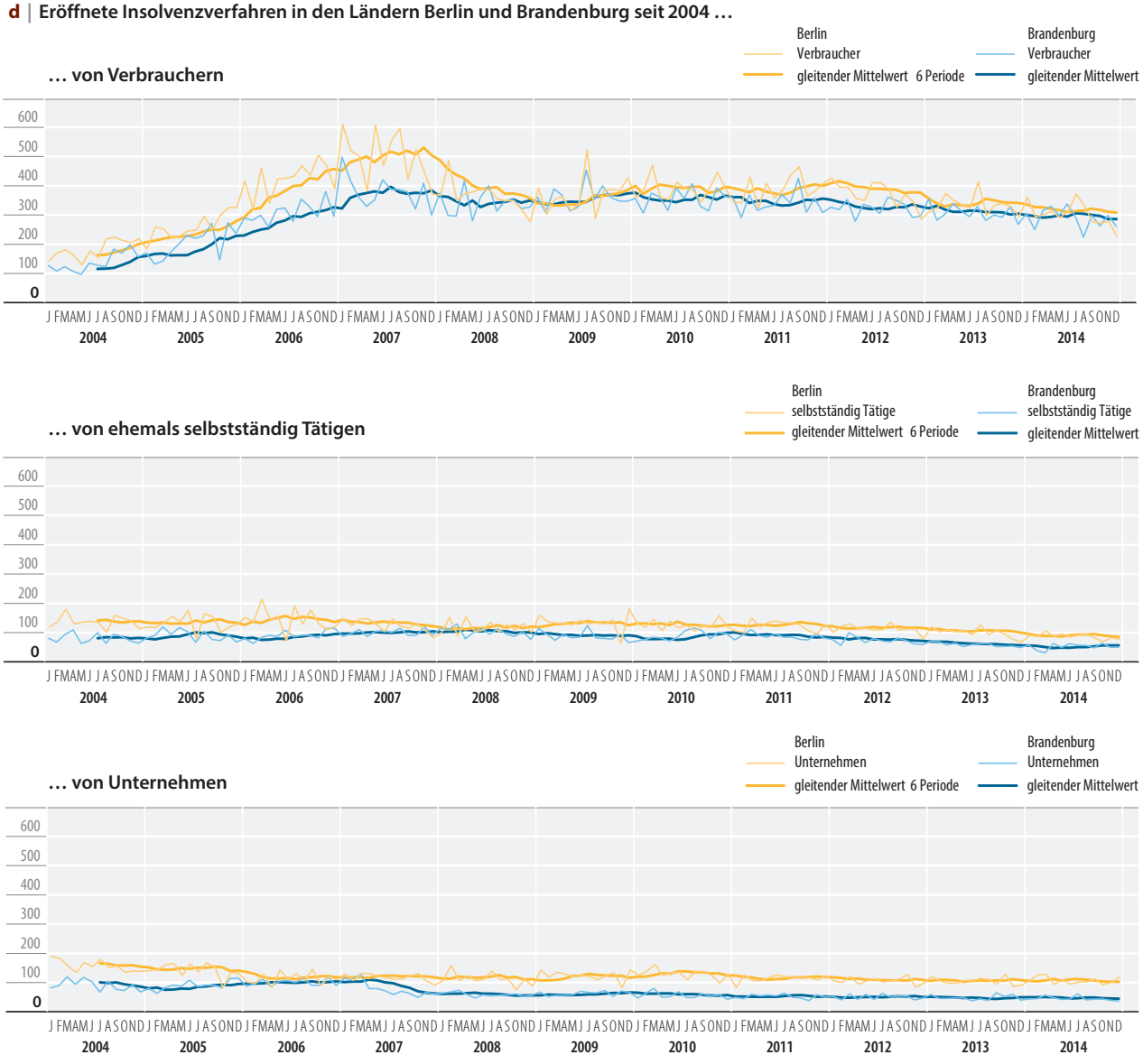
Die Insolvenzverfahren der *ehemals selbstständig Tätigen* nehmen seit 2004 kontinuierlich in beiden Bundesländern ab. Bei dieser Schuldnergruppe handelt es sich um Privatpersonen, die aus ihrer früheren Selbstständigkeit unbefriedigte Forderungen in ihre „Privatheit" mitnehmen mussten. Hierbei wird der Grad der Verschuldung an der Anzahl der Gläubiger gemessen: Handelt es sich um Forderungen von weniger als 20 Gläubigern, wird ein einfaches Insolvenzverfahren für Verbraucher eröffnet; warten 20 und mehr Gläubiger auf die Befriedigung ihrer Forderungen, dann wird ein Regelinsolvenzverfahren für Unternehmen eröffnet.

Über die Entwicklung der Anzahl eröffneter Insolvenzverfahren von *Unternehmen* (Regelinsolvenzen) ist festzustellen, dass sie anfänglich leicht abnehmend sich im Mittel auf ein Niveau von 100 monatlichen Verfahrenseröffnungen in Berlin und 10 in Brandenburg einzupendeln scheinen.

Die Anzahl der Insolvenzverfahren sagt noch nichts über ihre Größenordnung an finanziellem Schaden und Verlust an Arbeitsplätzen aus. Die hier diskutierte monatliche Statistik der eröffneten Insolvenzverfahren wird zukünftig ergänzt durch die Statistik der beendeten Insolvenzverfahren, welche dann neben anderen Informationen über den tatsächlichen volkswirtschaftlichen Schaden Auskunft geben wird.

Dr. Thomas Heymann leitet das Referat *Unternehmensregister, Gewerbeanzeigen, Insolvenzen* des Amtes für Statistik Berlin-Brandenburg.

d | Eröffnete Insolvenzverfahren in den Ländern Berlin und Brandenburg seit 2004 …



Historisches

Über Inhalt und Methode einer Berliner Schulstatistik

Schulstatistik um 1870 – Teil 2

von **Jürgen Hübner** und **Holger Leerhoff**

Vorbemerkung

Im vierten Jahrgang der Schrift *Berlin und seine Entwicklung. Städtisches Jahrbuch für Volkswirthschaft und Statistik* von 1870 findet sich die Abhandlung „Über Inhalt und Methode einer Berliner Schulstatistik" von Dr. H. Schwabe und Dr. F. Bartholomäi, der in dieser Zeitschrift als Fortsetzungsartikel besprochen wird. Der erste Teil, in dem der Hintergrund und Auftrag zur Abhandlung dargelegt wurden, findet sich in Heft 1/2015; in dieser Ausgabe sollen die Teile A.2: „Die allgemeinen Requisiten einer Schulstatistik" und A.3: „Die Methode der Erhebung" vorgestellt werden.

A. Die theoretischen Zielpunkte der Schulstatistik

2. Die allgemeinen Requisiten einer Schulstatistik

Gleich zu Beginn der Ausführungen findet sich eine heute etwas ungewöhnlich erscheinende Überlegung: „Um keines der vielen, eine Schulstatistik bestimmenden Elemente in ihrer mannigfaltigen Verwebung zu übersehen, fingieren wir eine bereits fertige Schulstatistik und stellen sie einem ebenfalls fingierten, pädagogisch gebildeten Vater zur Disposition, der nach ihr für seine Kinder eine oder mehrere Schulen auszuwählen gedenkt." Zwar gehört auch heute die omnipräsente ‚interessierte Öffentlichkeit' zur Zielgruppe der Arbeit der Statistiker, aber sicherlich nicht an erster Stelle. Auch ließe der im ersten Teil dieser Serie beschriebene Auftrag des Herrn Seydel eher Verwaltung und Politik als Zielgruppe der Statistik vermuten.

Drei Gesichtspunkte werden von den Autoren als besonders wichtig für die Bedürfnisse des so angenommenen Vaters erachtet: „das körperliche Wohl des Kindes, die Beschaffenheit der Erziehung und des Unterrichtes in der Schule."

Hinsichtlich der Schule selbst sind diverse Faktoren in Betracht zu ziehen, etwa die Entfernung der Schule zum Elternhaus: „[…] denn unter übrigens gleichen Verhältnissen ist die nächste Schule die beste." Schließlich musste der Schulweg seinerzeit mehrfach täglich zurückgelegt werden, da zwischen Vor- und Nachmittagsunterricht zu Hause das Mittagessen eingenommen wurde. Relevant war die Entfernung auch aus anderen Gründen, denn „endlich mehrt sich mit der Länge des Wegs die Gelegenheit, Neues, Auffälliges, die Aufmerksamkeit Fesselndes und darum dem beginnenden Unterrichte Widerstrebendes wahr= und aufzunehmen" und es könnte gar vorkommen, dass das Kind „durch die auf dem längeren Wege gebotene Vervielfältigung der Gelegenheit sich verführen lässt, um die Schule herumzugehen." Bei dem Schulweg sollte zudem berücksichtigt werden, „ob die Kinder Straßen zu passieren haben, welche die Schlupfwinkel der gewerbsmäßigen Unzucht in sich bergen." Entsprechend sollten in der Statistik die Anschriften der Schule selbst und die aller Schüler erfasst werden, um den Einzugsbereich der Schule bestimmen zu können; ein kleinerer Einzugsbereich wurde dabei als wünschenswert erachtet.

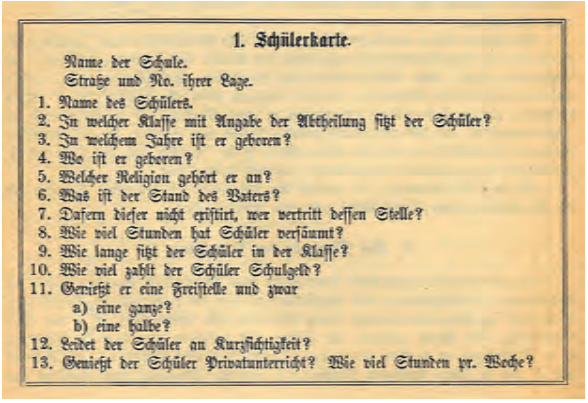
Doch auch die eigentliche Lage der Schule war wichtig: Der Boden, auf dem sie gebaut war, sollte trocken sein und die in der Nähe befindlichen „gewerblichen Anlagen, welche die Luft mit übelriechenden und gesundheitsgefährlichen Gasen erfüllen", bekannt und be-

schrieben sein. Eine Erörterung der von „der Feuchtigkeit der Wände“ der Schule ausgehenden möglichen Gefahren gemahnt auf unangenehme Weise wieder an die Gegenwart in Berlin.

Den Lichtverhältnissen in der Schule wurde ebenfalls große Aufmerksamkeit geschenkt. Gerade weil die damaligen Möglichkeiten insbesondere der künstlichen Beleuchtung noch eingeschränkt waren, spielte das Tageslicht eine weit wichtigere Rolle als heute. Die Ausstattung des Schulgebäudes mit Fenstern und die Größe und Ausrichtung der Klassenräume sollten erfasst werden. Eine ausreichende Menge an Sonnenlicht war wichtig, damit es „die Feuchtigkeit sowie die Ausdünstungen der Kinder aufzehren könne“. Auch die durch die Sonne verursachte Erhitzung der Räume wurde in Betracht gezogen.

Im Zusammenhang mit den Lichtverhältnissen gehen die Autoren zudem auf die Ursachen der verbreiteten Kurzsichtigkeit ein: „Gewiß wirkt bald die eine, bald die andere Ursache, bald allein, bald im Bunde mit anderen: Disposition, kleine Schrift, Onanie, Lichtmangel, vorgebeugtes Sitzen.“ An dieser Stelle merkt nun der heutige aufgeklärte Bürger, dass es in den letzten fast 150 Jahren einen nicht unbedeutenden wissenschaftlichen Erkenntniszuwachs – zumindest auf einem Spezialgebiet – gegeben hat. Für die Schulstatistik fordern die Autoren, dass die Anzahl der kurzsichtigen Schüler und der jeweilige Grad der Kurzsichtigkeit zu erfassen sei. Eine entsprechende Untersuchung könne durch den Lehrer, besser aber durch junge Augenärzte vorgenommen werden, die „sich gewiß gerne der dankbaren Aufgabe“ unterzögen.

Ein großes Problem zu Zeiten der Ofenheizung in den Schulen war die Luft in den Klassenzimmern: „Das Kind muß den Staub, die Ausdünstungen der Kleider, die Produkte des Athmungsprozesses und der Hautthätigkeit seiner Mitschüler einathmen.“ Potenzielle gesundheitliche Gefahren wurden auch schon diskutiert; entsprechend sollten in die Schulstatistik ebenfalls Angaben zu Anzahl und Alter der Schüler, das Volumen der Klassenräume, den Aufbewahrungsort der Kleider der Kinder sowie zu Heizung und Ventilation aufgenommen werden.



1. Schülerkarte.

Name der Schule.
Straße und No. ihrer Lage.

1. Name des Schülers.
2. In welcher Klasse mit Angabe der Abtheilung sitzt der Schüler?
3. In welchem Jahre ist er geboren?
4. Wo ist er geboren?
5. Welcher Religion gehört er an?
6. Was ist der Stand des Vaters?
7. Dafern dieser nicht existirt, wer vertritt dessen Stelle?
8. Wie viel Stunden hat Schüler versäumt?
9. Wie lange sitzt der Schüler in der Klasse?
10. Wie viel zahlt der Schüler Schulgeld?
11. Genießt er eine Freistelle und zwar
    a) eine ganze?
    b) eine halbe?
12. Leidet der Schüler an Kurzsichtigkeit?
13. Genießt der Schüler Privatunterricht? Wie viel Stunden pr. Woche?

Die Versorgung der Kinder mit Trinkwasser wird auch thematisiert. Während besorgte Eltern heute zwischen hunderten Modellen an Trinkflaschen aus Aluminium, verschiedenen Kunststoffen und in den unterschiedlichsten Formen und Farben wählen können, damit ihrem Nachwuchs der Gang zum Wasserhahn im Schulgebäude erspart bleibt, stellte sich die Lage im 19. Jahrhundert noch anders dar. Die Autoren zitieren als Experten einen Herrn Falk mit den Worten: „Daß eine ausreichende Versorgung der Schule mit Trinkwasser nothwendig ist […], leuchtet ein, wenn wir uns aus unserer eigenen Schulzeit des Anblicks erinnern, welche in den Sommernachmittagszwischenstunden die dicht um den Brunnen gescharte durstige Jugend gewährte, welche öfters die dem Brunnen Näheren zwang, den Becher kreisen zu lassen. Jede Klasse muß wenigstens ein Glas in ihrem Inventarium haben.“ Durstige Kinder könnten aber dem Unterricht nicht folgen oder gar gesundheitliche Schäden davontragen – die Schulstatistik sollte entsprechend auch Angaben zum Brunnenwasser enthalten, zu dessen Temperatur, chemischer Zusammensetzung und mechanischen Beimengungen.

Für die Lehr- und Lernsituation wird festgestellt: „Aber der Unterricht erfordert, daß die Sitze so bequem als möglich sind, damit das Sitzen so leicht als möglich sei und von ihm aus keine Störung der Gedankenbewegung eintrete.“ Ein weiterer – heute wohl in Vergessenheit geratener – Gedanke: „Die Ruhe muß durch Bewegung unterbrochen werden, und je kleiner die Kinder sind, desto öfters. So wie die Ruhe muß auch die Bewegung eine ungestörte sein. Die Schule bedarf also eines Spiel= oder Tummelplatzes.“ Heute scheint allerdings die bundesdeutsche Schule eher ein Tummelplatz sehr verschiedener Unterrichtsmethoden und -inhalte zu sein. – Statistisch zu erfassen wären hier die Körpergröße der Kinder (halbjährlich) sowie bei den Sitzbänken deren Länge, Tiefe, Höhe und Angaben zum Vorhandensein einer Lehne. Auch die Größe des Spielplatzes sowie Ausweichmöglichkeiten bei Regenwetter oder strenger Kälte sollten registriert werden.

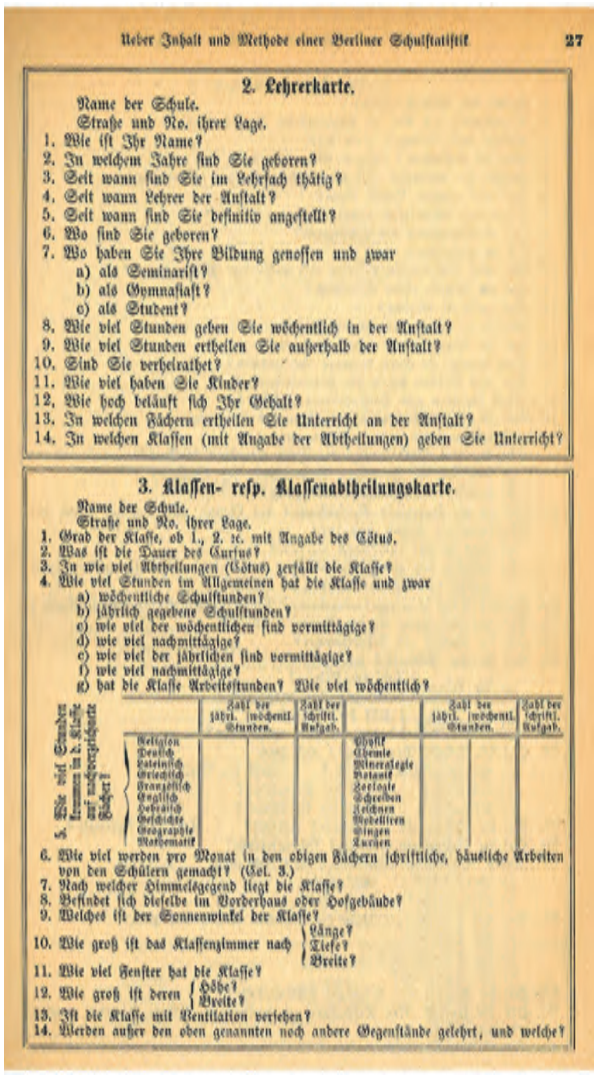
In den weiteren Ausführungen wird sich nun der „Morbidität und Mortalität" – also dem Krankenstand und der Sterblichkeit – unter den Kindern zugewandt. Das heute vielleicht am kontroversesten diskutierte Thema in der Bildungslandschaft ist die Inklusion, wobei die Statistik in einigen Bundesländern damit zu kämpfen hat, dass schon die Feststellung und statistische Erfassung einer vorliegenden Behinderung als Diskriminierung der betreffenden Person verstanden wird und entsprechend nicht mehr erfolgt(!), was zu beinahe absurden Konsequenzen etwa bei der Mittelvergabe an Institutionen führt. In dieser Hinsicht war man vor rund 150 Jahren deutlich unbekümmerter: „[…] so sind außerdem insbesondere diejenigen Kranken hervorzuheben, die zwar nicht die Abwesenheit vom Unterricht bedingen, aber störend auf denselben einwirken, wie die Blinden, Tauben, Stotternden, Cretinen, sodann die, welche den Unterricht nicht stören, aber an augenfälligen Gebrechen leiden, wie Lahmheit, Verkrüppelung etc." Eine detaillierte Erfassung wäre heute, wie gesagt, völlig undenkbar.

Hinsichtlich der Sterblichkeit wird festgestellt: „Denn […], so ist bekannt, daß die größere Sterblichkeit der ärmeren Volksklassen zum Theil auf Rechnung der ungeeigneten Wohnungen zu setzten ist. Nahrung, Kleidung, Wohnung, Arbeit und Genuss bewirken, daß den Kindern dieser Volksklassen in Durchschnitt 18 Jahre Lebensdauer weniger mit auf den Weg gegeben wird, als den Kindern der Wohlhabenden und Reichen."

Bereits im ersten Teil dieser Serie wurde angesprochen, dass der – wie man es heute ausdrücken würde – Zusammenhang von häuslichem Umfeld der Kinder und deren Bildungschancen schon vor 150 Jahren wahrgenommen und durchaus kritisch diskutiert wurde. Soziologisch analysiert wurden die unterschiedlichen Dimensionen dieses Zusammenhangs in den 80er Jahren des letzten Jahrhunderts von Bourdieu, worauf aufbauend in der heutigen Bildungsberichterstattung die Konzeption der Risikolagen erarbeitet wurde (siehe dazu auch Rockmann, Ulrike; Rehkämper, Klaus; Leerhoff, Holger: „Bildungskapital verringert Bildungsrisiken", DJI Impulse 3/2014, S. 26–29; nachgedruckt in Ausgabe 1/2015 dieser Zeitschrift).

Die Dres. Schwabe und Bartholomäi führen dazu etwa aus, dass es „eine bekannte psychologische und leicht begreifliche Erscheinung [ist], daß des Kindes Bildungs= und Wissenstrieb im Durchschnitt der Bildung seiner Umgebung proportional ist. Das Kind lesender und sich geistig beschäftigender Eltern kommt mit der Erwartung des Lernens in die Schule, mit der Richtung auf Lesen, Schreiben und so weiter. Wohingegen [sic!] das Beispiel der geistigen Arbeit fehlt, da verhalten sich auch die Kinder zu ihr gleichgültig." Die Aufgabe der Lehrer sei, diese Unterschiede zu vermindern, indem sie die Kinder „für das Lernenwollen gewinnen, sie für die Schule zu fanatisiren, ihnen dieselbe lieb und werth zu machen".

Nicht nur dazu bedurfte es der geforderten ordentlichen Schulräume, die in Berlin damals aber nicht ausreichend vorhanden waren: „Die Schulen Berlin's leiden an Ueberfüllung. […] Ueberfüllte Klassenzimmer sind ein Unglück für jede Schule." Zu den heute auch noch bekannten Ursachen wurde festgestellt: „Die Ueberfüllung der Klassen ist eine Folge der beschränkten Finanzverhältnisse, erspart aber durchaus nicht so



Ueber Inhalt und Methode einer Berliner Schulstatistik 27

2. Lehrerkarte.

Name der Schule.
Straße und No. ihrer Lage.

1. Wie ist Ihr Name?
2. In welchem Jahre sind Sie geboren?
3. Seit wann sind Sie im Lehrfach thätig?
4. Seit wann Lehrer der Anstalt?
5. Seit wann sind Sie definitiv angestellt?
6. Wo sind Sie geboren?
7. Wo haben Sie Ihre Bildung genossen und zwar
   a) als Seminarist?
   b) als Gymnasiast?
   c) als Student?
8. Wie viel Stunden geben Sie wöchentlich in der Anstalt?
9. Wie viel Stunden ertheilen Sie außerhalb der Anstalt?
10. Sind Sie verheirathet?
11. Wie viel haben Sie Kinder?
12. Wie hoch beläuft sich Ihr Gehalt?
13. In welchen Fächern ertheilen Sie Unterricht an der Anstalt?
14. In welchen Klassen (mit Angabe der Abtheilungen) geben Sie Unterricht?

3. Klassen- resp. Klassenabtheilungskarte.

Name der Schule.
Straße und No. ihrer Lage.

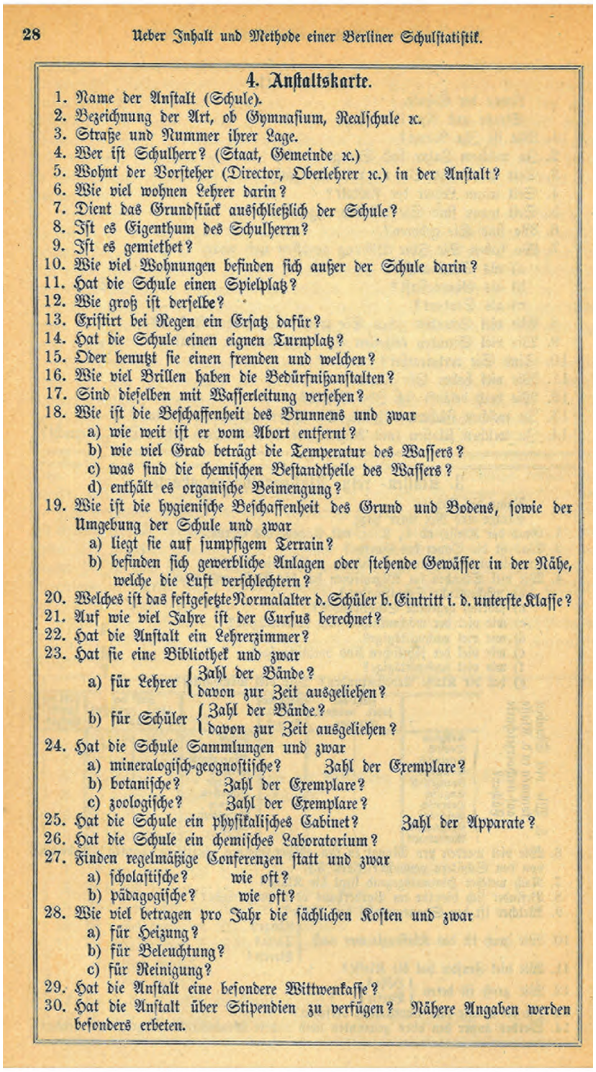
1. Grad der Klasse, ob I., 2. ꝛc. mit Angabe des Cötus.
2. Was ist die Dauer des Cursus?
3. In wie viel Abtheilungen (Cötus) zerfällt die Klasse?
4. Wie viel Stunden im Allgemeinen hat die Klasse und zwar
   a) wöchentliche Schulstunden?
   b) jährlich gegebene Schulstunden?
   c) wie viel der wöchentlichen sind vormittägige?
   d) wie viel nachmittägige?
   e) wie viel der jährlichen sind vormittägige?
   f) wie viel nachmittägige?
   g) hat die Klasse Arbeitsstunden? Wie viel wöchentlich?

Wie viel Stunden kommen in d. Klasse auf nachstehende Fächer:		Zahl der jährl. wöchentl. Stunden.	Zahl der schriftl. Aufgab.		Zahl der jährl. wöchentl. Stunden.	Zahl der schriftl. Aufgab.
	Religion			Physik		
	Deutsch			Chemie		
	Lateinisch			Mineralogie		
	Griechisch			Botanik		
	Französisch			Zoologie		
	Englisch			Schreiben		
	Hebräisch			Zeichnen		
	Geschichte			Modelliren		
	Geographie			Singen		
	Mathematik			Turnen		

6. Wie viel werden pro Monat in den obigen Fächern schriftliche, häusliche Arbeiten von den Schülern gemacht? (Cel. 3.)
7. Nach welcher Himmelsgegend liegt die Klasse?
8. Befindet sich dieselbe im Vorderhaus oder Hofgebäude?
9. Welches ist der Sonnenwinkel der Klasse?
10. Wie groß ist das Klassenzimmer nach { Länge? / Tiefe? / Breite?
11. Wie viel Fenster hat die Klasse?
12. Wie groß ist deren { Höhe? / Breite?
13. Ist die Klasse mit Ventilation versehen?
14. Werden außer den oben genannten noch andere Gegenstände gelehrt, und welche?

viel, als die einseitige Rechnung ergiebt." So wird etwa ausgeführt, dass große Klassen neben den offensichtlichen Nachteilen für die Schüler auch zu einer stärkeren Beanspruchung des Lehrers führen, der so früher in den Ruhestand geht, „während er in einer kleinen Klasse vielleicht noch Jahre lang wirken könnte." Angesichts des auch heute großen Anteils von Lehrkräften, die frühzeitig in Pension gehen, wäre eine genauere volkswirtschaftliche Analyse dieser Überlegung möglicherweise durchaus nutzbringend.

In einem längeren Abschnitt werden dann die Fragen der günstigsten Schulform diskutiert, zum Teil im Zusammenhang mit der Bildung und dem Einkommen der Eltern; aber auch, wie man die Armen und Begabten am besten beschulen solle. Letztendlich gelangen die Autoren zu der heute nicht mehr zeitgemäßen Schlussfolgerung: „Wir halten es für unmöglich, Kinder aus ganz verschiedenen Ständen und Vermögensklassen in derselben Schule zu vereinigen und ihnen einen wahrhaft erziehenden Unterricht zu erteilen. Denn die Verschiedenheiten des Standes und des Vermögens gehören zu den wesentlichen Momenten der Individualität." Interessanterweise wird hier, wie auch heute noch, die Individualität als das Hauptargument des Für und Wider jedweder Schulform benutzt. Es wird aber keine durchgängige Lösung vorgestellt – die bis heute auf sich warten lässt. Allerdings wird die alte Erkenntnis, dass „eine Hauptaufgabe des Unterrichts das Lernen lehren ist", in diesem Zusammenhang mehrfach hervorgehoben. Für die Schulstatistik folgern die Autoren, dass Angaben zu Stand und Vermögen der Eltern darin berücksichtigt werden müssten.



28 Ueber Inhalt und Methode einer Berliner Schulstatistik.

4. Anstaltskarte.

1. Name der Anstalt (Schule).
2. Bezeichnung der Art, ob Gymnasium, Realschule ꝛc.
3. Straße und Nummer ihrer Lage.
4. Wer ist Schulherr? (Staat, Gemeinde ꝛc.)
5. Wohnt der Vorsteher (Director, Oberlehrer ꝛc.) in der Anstalt?
6. Wie viel wohnen Lehrer darin?
7. Dient das Grundstück ausschließlich der Schule?
8. Ist es Eigenthum des Schulherrn?
9. Ist es gemiethet?
10. Wie viel Wohnungen befinden sich außer der Schule darin?
11. Hat die Schule einen Spielplatz?
12. Wie groß ist derselbe?
13. Existirt bei Regen ein Ersatz dafür?
14. Hat die Schule einen eignen Turnplatz?
15. Oder benutzt sie einen fremden und welchen?
16. Wie viel Brillen haben die Bedürfnißanstalten?
17. Sind dieselben mit Wasserleitung versehen?
18. Wie ist die Beschaffenheit des Brunnens und zwar
    a) wie weit ist er vom Abort entfernt?
    b) wie viel Grad beträgt die Temperatur des Wassers?
    c) was sind die chemischen Bestandtheile des Wassers?
    d) enthält es organische Beimengung?
19. Wie ist die hygienische Beschaffenheit des Grund und Bodens, sowie der Umgebung der Schule und zwar
    a) liegt sie auf sumpfigem Terrain?
    b) befinden sich gewerbliche Anlagen oder stehende Gewässer in der Nähe, welche die Luft verschlechtern?
20. Welches ist das festgesetzte Normalalter d. Schüler b. Eintritt i. d. unterste Klasse?
21. Auf wie viel Jahre ist der Cursus berechnet?
22. Hat die Anstalt ein Lehrerzimmer?
23. Hat sie eine Bibliothek und zwar
    a) für Lehrer { Zahl der Bände? / davon zur Zeit ausgeliehen?
    b) für Schüler { Zahl der Bände? / davon zur Zeit ausgeliehen?
24. Hat die Schule Sammlungen und zwar
    a) mineralogisch-geognostische? Zahl der Exemplare?
    b) botanische? Zahl der Exemplare?
    c) zoologische? Zahl der Exemplare?
25. Hat die Schule ein physikalisches Cabinet? Zahl der Apparate?
26. Hat die Schule ein chemisches Laboratorium?
27. Finden regelmäßige Conferenzen statt und zwar
    a) scholastische? wie oft?
    b) pädagogische? wie oft?
28. Wie viel betragen pro Jahr die sächlichen Kosten und zwar
    a) für Heizung?
    b) für Beleuchtung?
    c) für Reinigung?
29. Hat die Anstalt eine besondere Wittwenkasse?
30. Hat die Anstalt über Stipendien zu verfügen? Nähere Angaben werden besonders erbeten.

Dann wandte man sich dem Ziel des Unterrichtes zu. Immer mit dem Hintergedanken: „Vor Allem wird nun unserem supponirten Vater im Bezug auf den Unterricht bei der Wahl einer Schule die Art derselben zu wissen nötig sein. Diese wird aber charakterisirt durch das Ziel des Unterrichts." Entsprechend sollten neben den allgemeinen Unterrichtsmethoden und Klassenzielen auch die praktischen Mittel wie Lehr- und Lernbücher, aber auch Art und Umfang der Bibliotheken, Sammlungen etc. bekannt sein.

Eine Schlüsselrolle nehmen, wenig überraschend, die Lehrer ein, die „mit pädagogischer Wissenschaft und Kunst ausgerüstet sein müssen". Zur Beurteilung ihrer Eignung sollten für die Schulstatistik etliche Merkmale erfasst werden, die heutigen Datenschützern wohl umfassender als unbedingt nötig erscheinen dürften. Dazu zählen neben Angaben zum Alter, Geburts- und Ausbildungsort, der praktischen Erfahrung und Art und Umfang des erteilten Unterrichts auch Angaben zur Familie und zum Gehalt. Hierbei muss berücksichtigt werden, dass eine Schule oft nur einen Lehrer hatte; sollten mehrere Lehrer an einer Schule arbeiteten, sollte auch noch die Zuordnung von Lehrern und Klassen erfasst werden.

Auch die Ausbildung der Lehrer wird kritisiert: „Die Vorbildung, welche Universitäten und Seminare geben, ist im Durchschnitt wahrhaft kläglich." Umso wichtiger wäre der Austausch der Lehrer untereinander, der im Regelfall in pädagogischen Konferenzen stattfand. Ziel dieser Konferenzen war eine laufende Fortbildung der Lehrer und auch eine Einheit des Kollegiums an den Schulen. Die Anzahl solcher Konferenzen sollte Eingang in die Schulstatistik finden.

Ein weiterer Faktor, der in der Schulstatistik nicht unberücksichtigt bleiben durfte, war der sittliche Zustand der Schule, auf den Rückschlüsse aus der Zahl der Schulvergehen gezogen werden können. Schulvergehen „sind etwa Unordnung, Ungehorsam, Faulheit und Lüge". Auch entsprechende Kennzahlen werden durch die Autoren vorgeschlagen:

„Die Unordnung kann gemessen werden
1) durch das Zuspätkommen,
2) durch das Vergessen eines Lehr= oder Lernmittels oder einer Arbeit,
3) durch die Zahl der ohne Entschuldigung versäumten Schulstunden,
4) durch die Zahl der Unordnungen in Bezug auf Reinlichkeit des Körpers und der Kleidung;

die Faulheit durch die Zahl
1) der nicht gefertigten,
2) der nicht sorgfältig gemachten Arbeiten."

Die Lüge aber nimmt eine Sonderrolle ein: „Das Hauptkennzeichen für die Sittlichkeit einer Schule aber ist die Abwesenheit der Lüge, denn da sie mit allen sittlichen Ideen in den härtesten Conflict geräth und die Gelegenheiten, ihr zu verfallen, nicht gar selten sind, so weist ihre Abwesenheit darauf hin, daß die sittlichen Ideen in dem Schulleben eine nicht unbedeutende Herrschaft erlangt haben." Freilich ist die „Abwesenheit der Lüge" nur schlecht statistisch zu fassen.

Interessant ist, dass die Autoren – entgegen dem Eindruck, den man durch den vorigen Absatz gewinnen konnte – beim sittlichen Zustand in Bezug auf Unordnung und Trägheit durchaus auch die Lehrer in der Pflicht sehen: „Man darf z. B. den Zögling nicht etwa bei Kopfweh noch arbeiten, lernen, sich üben lassen; muß man doch schon beim bloßen Gähnen, wenn es sich wiederholt, mit dem Gegenstande seiner Behandlung wechseln. Wo die Schüler in einzelnen Stunden regelmäßig frisch, in anderen regelmäßig stumpf und matt sind, da ist etwas Falsches in Lehrplan und Methode." Sind etwa Schüler mit dem Stoffe überfordert, muss ihnen mehr Zeit gegeben werden; gegebenenfalls sind auch Wissenslücken in Kauf zu nehmen, denn: „Wo solche und ähnliche Rücksichten nicht genommen werden, ist es die Schule selbst, welche zu Trägheit und Unordnung erzieht und das Ringen und Streben des schwächeren Schülers beeinträchtigt und unterdrückt."

Abschließend: „Zu diesen Punkten kommen endlich noch einige an sich äußerliche Verhältnisse: Kosten, Höhe des Schulgeldes, Zahl der Freistellen, Abgang der Schüler und ihre Bestimmung." Nicht im Artikel diskutiert, aber statistisch erfasst werden sollte auch die Religion der Eltern und Kinder. Nicht erwähnt wurde das Thema, um schon damals „gewissen Streitpunkten aus dem Wege zu gehen". Wir erinnern an dieser Stelle an das später im *Statistisches Jahrbuch der Stadt Berlin 1882* diskutierte Problem mit einem „frechen Angriff" von antisemitischer Seite gegen das Statistische Jahrbuch und gegen einige Veröffentlichungen, die mit den amtlichen Zahlen arbeiteten. Als statistisches Desideratum wurde noch eine Erfassung der „psychologischen Eigenthümlichkeit" der Kinder ausgemacht, wofür die Grundlage aber erst durch eine bessere psychologische Ausbildung der Lehrer gelegt werden müsste.

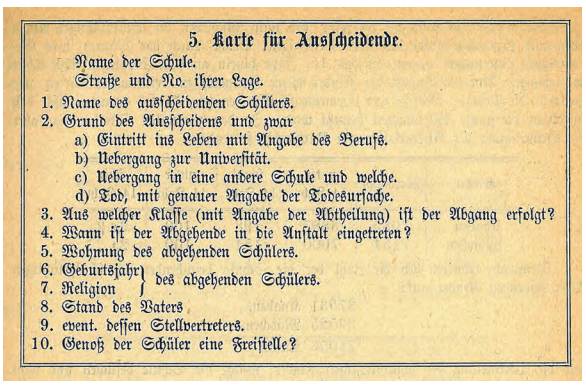
3. Die Methode der Erhebung

Zuerst stand der Vorschlag, die bewährte und heute noch im Kern so praktizierte Methode der Bevölkerungsstatistik – Erhebung und Fortschreibung –, auch auf die Schulstatistik anzuwenden. Nach der Volkszählung von 1867 wurde aber die „sogenannte Kartenmethode" favorisiert, die deutlich detailliertere Ergebnisse liefern könnte und „dadurch die gesellschaftlichen Institutionen ungleich vollständiger und schärfer zu charakterisiren vermag". Es würden entsprechend fünf Karten zur Schule selbst, zu Schülern, Lehrern, Klassen und den Ausscheidenden eingesetzt werden müssen.

Fortsetzung folgt

Dr. Jürgen Hübner war bis zu seinem Ausscheiden im Mai 2014 verantwortlich für die Zeitschrift für amtliche Statistik Berlin Brandenburg.

Dr. Holger Leerhoff ist Referent für Bildungsanalysen beim Amt für Statistik Berlin-Brandenburg.



5. Karte für Ausscheidende.

Name der Schule.
Straße und No. ihrer Lage.

1. Name des ausscheidenden Schülers.
2. Grund des Ausscheidens und zwar
   a) Eintritt ins Leben mit Angabe des Berufs.
   b) Uebergang zur Universität.
   c) Uebergang in eine andere Schule und welche.
   d) Tod, mit genauer Angabe der Todesursache.
3. Aus welcher Klasse (mit Angabe der Abtheilung) ist der Abgang erfolgt?
4. Wann ist der Abgehende in die Anstalt eingetreten?
5. Wohnung des abgehenden Schülers.
6. Geburtsjahr } des abgehenden Schülers.
7. Religion }
8. Stand des Vaters
9. event. dessen Stellvertreters.
10. Genoß der Schüler eine Freistelle?

Statistische Woche 2015 in Hamburg

Die jährliche gemeinsame Tagung der Deutschen Statistischen Gesellschaft (DStatG) und des Verbandes Deutscher Städtestatistiker (VDSt), die Statistische Woche, findet in diesem Jahr vom **15. bis 18. September** an der **Helmut-Schmidt-Universität** in Hamburg statt. Die programmatischen Schwerpunkte sind:

Statistische Indikatoren für das politische Monitoring,
Statistical Surveillance,
Statistical Analysis of Network Data.

Der diesjährige Redner der Heinz-Grohmann-Vorlesung ist Joachim Wagner, Professor für Volkswirtschaftslehre mit dem Schwerpunkt Empirische Wirtschaftsforschung an der Universität Lüneburg. Die Gumbel-Vorlesung hält Christoph Rothe von der Columbia University (New York City).

Bei der Statistischen Woche 2015 wird es außerdem eine Session der Deutschen Gesellschaft für Demographie (DGD), der Spanischen Gesellschaft für Statistik und OR (SEIO) zum Thema *Data depth and classification* sowie eine Session der Italienischen Statistischen Gesellschaft geben.

Wie auch in den letzten Jahren ist das Amt für Statistik Berlin-Brandenburg (AfS) aktiv an der Veranstaltung beteiligt. Hartmut Bömermann (Leiter der Abteilung *Bevölkerung und Regionalstatistik*) ist Organisator der gemeinsamen Sitzung des Ausschusses für Regionalstatistik der DStatG und des VDSt. Dr. Holger Leerhoff (Referent im Bereich *Schule Berlin*, *Bildungsanalysen*) wird mit einem Beitrag zur Berliner Schulstatistik um 1870 vertreten sein. Über die Bereitstellung von Mikrodaten im Statistischen Informationssystem Berlin-Brandenburg (StatIS-BBB) sowie über die durchgeführten Nutzerschulungen referieren Ramona Voshage und Katja Baum (beide Referat *Mikrodaten, Analysen, Forschungsdatenzentrum*).

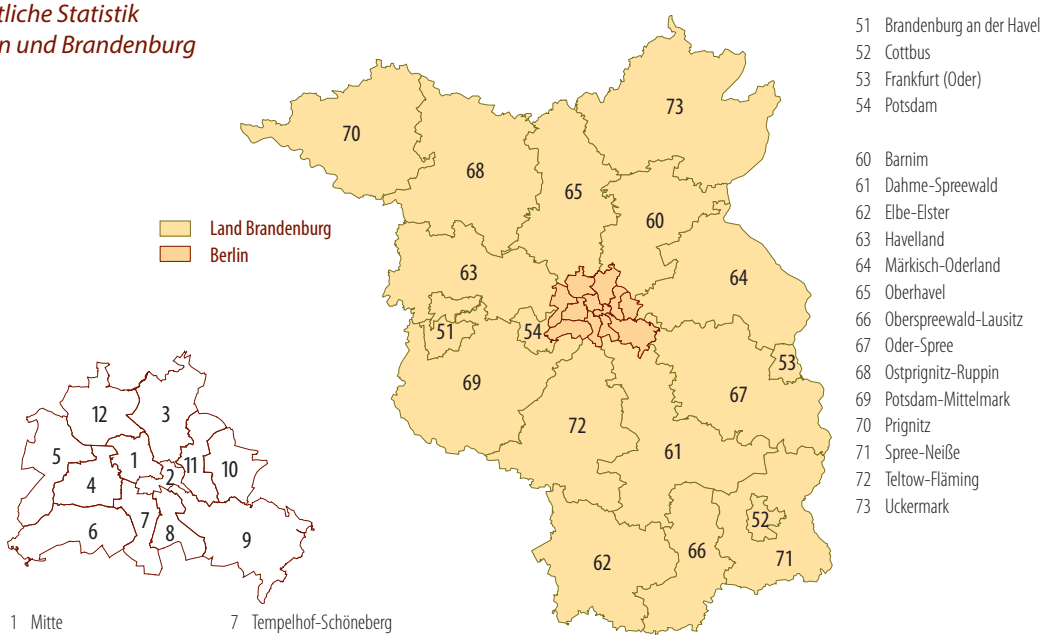
Die Statistische Woche ist eine hervorragende Gelegenheit zum Erfahrungsaustausch mit Kolleginnen und Kollegen aus dem Bereich Statistik, Datennutzerinnen und -nutzern verschiedener Institute und Verbände sowie aus Politik und Wissenschaft. Zudem ermöglicht die Veranstaltung, politische und soziale Entwicklungen zu diskutieren und Kontakte für künftige Kooperationen zu knüpfen.

Weitere Informationen zum Programm der Statistischen Woche 2015
und zur Anmeldung finden Sie unter: www.statistische-woche.de

statistik Berlin Brandenburg

- Wir berichten fachlich unabhängig, neutral und objektiv über die Ergebnisse der amtlichen Statistik.
- Wir haben den gesetzlichen Auftrag zur Datenerhebung mit der Möglichkeit zur Auskunftsverpflichtung.
- Wir garantieren die Einhaltung des Datenschutzes.
- Wir wenden adäquate statistische Methoden und Verfahren an und erhöhen kontinuierlich das erreichte Qualitätsniveau.
- Wir gewährleisten regionale und zeitliche Vergleichbarkeit unserer Statistiken durch überregionale Kooperation.
- Wir ermöglichen jedermann Zugang zu statistischen Ergebnissen.

Wir sind der führende Informationsdienstleister für amtliche Statistik in Berlin und Brandenburg



51 Brandenburg an der Havel
52 Cottbus
53 Frankfurt (Oder)
54 Potsdam

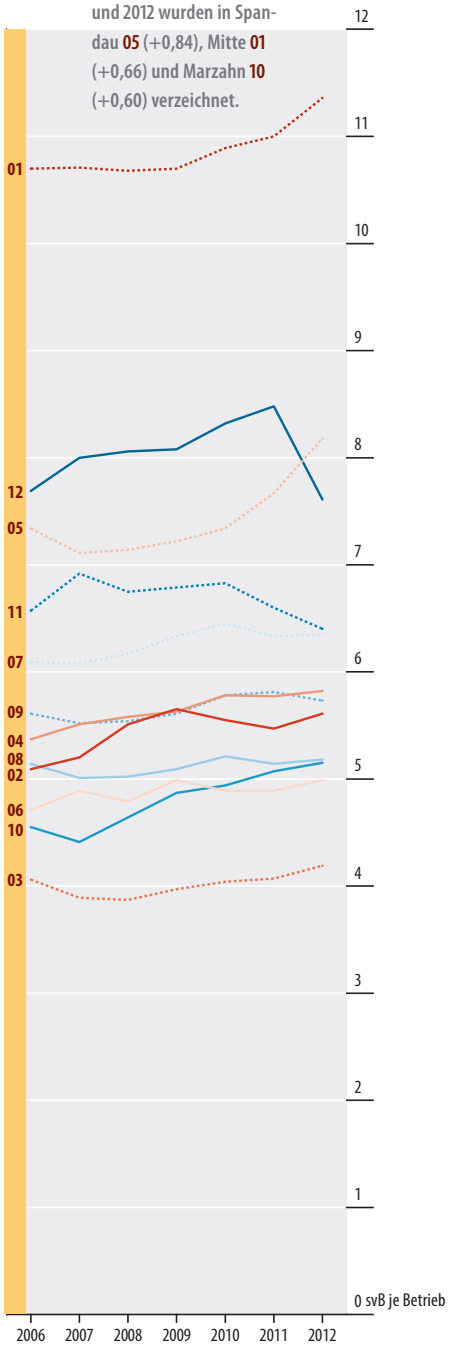
60 Barnim
61 Dahme-Spreewald
62 Elbe-Elster
63 Havelland
64 Märkisch-Oderland
65 Oberhavel
66 Oberspreewald-Lausitz
67 Oder-Spree
68 Ostprignitz-Ruppin
69 Potsdam-Mittelmark
70 Prignitz
71 Spree-Neiße
72 Teltow-Fläming
73 Uckermark

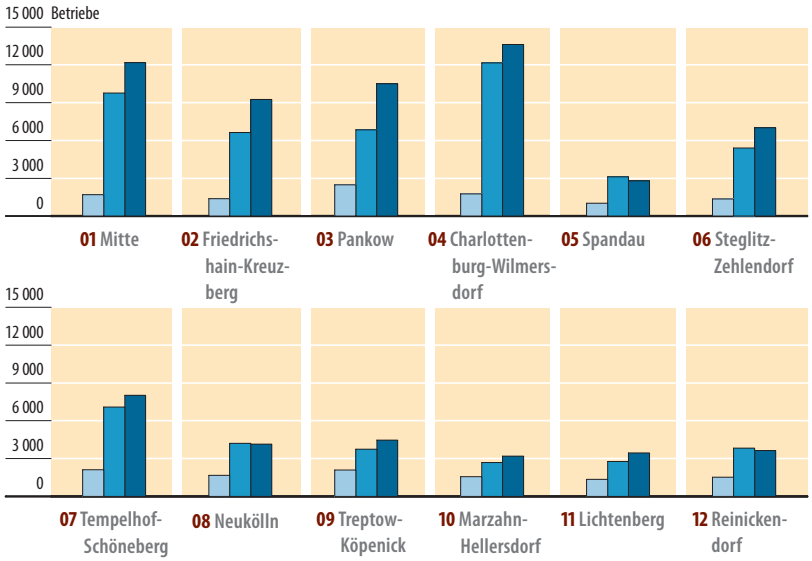
1 Mitte
2 Friedrichshain-Kreuzberg
3 Pankow
4 Charlottenburg-Wilmersdorf
5 Spandau
6 Steglitz-Zehlendorf
7 Tempelhof-Schöneberg
8 Neukölln
9 Treptow-Köpenick
10 Marzahn-Hellersdorf
11 Lichtenberg
12 Reinickendorf

Unter
www.statistik-berlin-brandenburg.de
finden Sie einen Überblick über das gesamte Leistungsspektrum des Amtes mit aktuellen Daten, Pressemitteilungen, Statistischen Berichten, regionalstatistischen Informationen, Wahlstatistiken und -analysen.

Betriebe in Berlin 2012

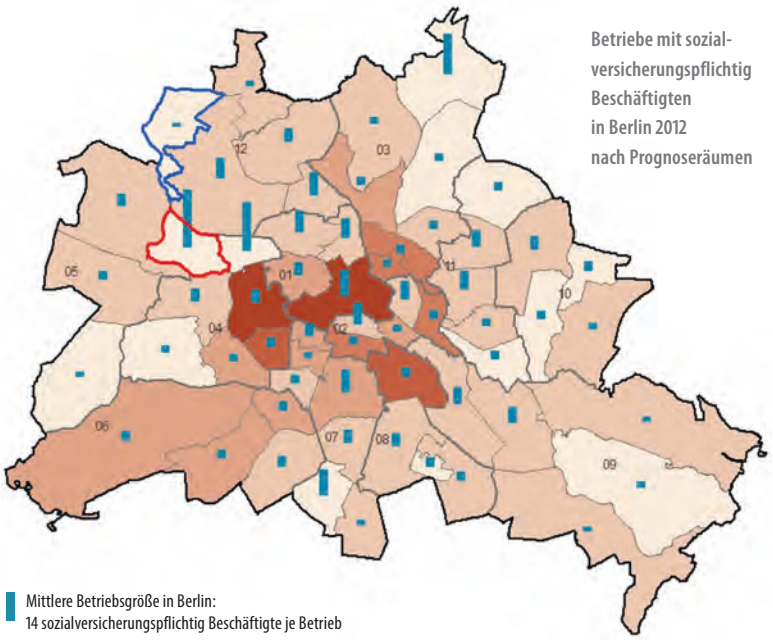
Am 31. Dezember 2012 gab es in Berlin **174 654 Betriebe** mit sozialversicherungspflichtig Beschäftigten (svB) und/oder steuerbaren Umsätzen aus Lieferungen und Leistungen. Die höchsten Zuwächse der Betriebsgröße (svB je Betrieb) zwischen 2006 und 2012 wurden in Spandau **05** (+0,84), Mitte **01** (+0,66) und Marzahn **10** (+0,60) verzeichnet.





Produzierendes Gewerbe (WZ[1]-Abschnitte B-F)
Verteilung und Versorgung (WZ-Abschnitte G-L)
übrige Dienstleistungen (WZ-Abschnitte M, N, P-S)

Betriebe mit sozialversicherungspflichtig Beschäftigten in Berlin 2012 nach Prognoseräumen



Mittlere Betriebsgröße in Berlin:
14 sozialversicherungspflichtig Beschäftigte je Betrieb

Minimalwert:
1,77 svB (Heiligensee/Konradshöhe)

Maximalwert:
29,0 svB (Spandau 3)

Anteil der sozialversicherungspflichtig Beschäftigten an den svB des Landes in %

- unter 0,5
- 0,5 bis unter 1,0
- 1,0 bis unter 1,5
- 1,5 bis unter 2,0
- 2,0 bis unter 3,0
- 3,0 und mehr
- Landesgrenze
- Bezirksgrenze
- Grenze Prognoseraum

Am 31. Dezember 2012 hatte Berlin insgesamt **1 126 499 sozialversicherungspflichtig Beschäftigte**, die einer Arbeit in einem Betrieb in Berlin nachgingen. Ein Jahr zuvor waren es 1 092 483 sozialversicherungspflichtig Beschäftigte.

1 Klassifikation der Wirtschaftszweige 2008, Herausgeber: Statistisches Bundesamt 2008