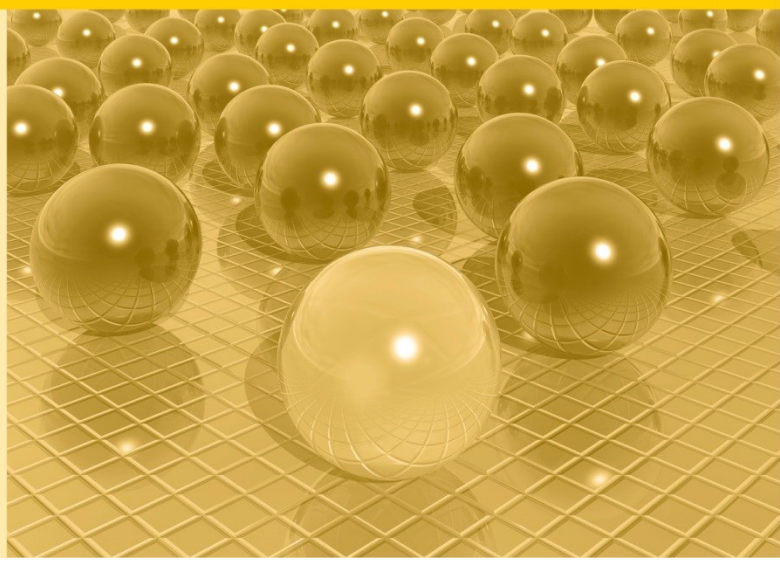
FDZ-Arbeitspapier
Nr. 50



Statistische Geheimhaltung – Der Schutz vertraulicher Daten in der amtlichen Statistik
Teil 1: Rechtliche und methodische Grundlagen
Teil 2: Herausforderungen und aktuelle Entwicklungen

Dipl.-Soz. Patrick Rothe

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

2019

Impressum

Herausgeber: Statistische Ämter des Bundes und der Länder

Herstellung: Information und Technik Nordrhein-Westfalen
Telefon 0211 9449-01 • Telefax 0211 442006
Internet: www.forschungsdatenzentrum.de
E-Mail: forschungsdatenzentrum@it.nrw.de

Fachliche Informationen
zu dieser Veröffentlichung:

Forschungsdatenzentrum der
Statistischen Ämter der Länder
– Geschäftsstelle –
Tel.: 0211 9449-2873
Fax: 0211 9449-8087
forschungsdatenzentrum@it.nrw.de

Informationen zum Datenangebot:

Statistisches Bundesamt
Forschungsdatenzentrum

Tel.: 0611 75-2420
Fax: 0611 72-3915
forschungsdatenzentrum@destatis.de

Forschungsdatenzentrum der
Statistischen Ämter der Länder
– Geschäftsstelle –
Tel.: 0211 9449-2873
Fax: 0211 9449-8087
forschungsdatenzentrum@it.nrw.de

Erscheinungsfolge: unregelmäßig

Erschienen im Februar 2019

Diese Publikation wird kostenlos als **PDF-Datei** zum Download unter www.forschungsdatenzentrum.de angeboten.

FDZ-Arbeitspapier
Nr. 50

Statistische Geheimhaltung – Der Schutz vertraulicher Daten in der amtlichen Statistik
Teil 1: Rechtliche und methodische Grundlagen
Teil 2: Herausforderungen und aktuelle Entwicklungen

Dipl.-Soz. Patrick Rothe

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

2019

Statistische Geheimhaltung – Der Schutz vertraulicher Daten in der amtlichen Statistik

Teil 1: Rechtliche und methodische Grundlagen

Patrick Rothe

Angesichts zahlreicher Enthüllungen über die missbräuchliche Datennutzung durch Geheimdienste, des Datenhungers millionenfach genutzter Webseiten und Internetdienste sowie des Zukunftstrends „Big Data“, ist der Schutz der Privatsphäre des Einzelnen wieder verstärkt in den Fokus der öffentlichen Diskussion gerückt. Die amtliche Statistik als einer der wichtigsten Datenproduzenten in Deutschland ist hiervon maßgeblich betroffen. Der vorliegende, zweiteilig konzipierte Beitrag trägt diesem Umstand Rechnung und setzt sich mit der Sicherstellung des Schutzes vertraulicher Daten innerhalb der amtlichen Statistik auseinander. Er bietet einen Überblick über die rechtlichen und methodischen Grundlagen der Geheimhaltungspraxis in den Statistischen Ämtern. Neben den einschlägigen gesetzlichen Regelungen werden die Grundzüge der gebräuchlichsten Geheimhaltungsverfahren und deren Auswirkungen auf die Veröffentlichungen der amtlichen Statistik vorgestellt.

1. Warum statistische Geheimhaltung?

Eines der verfassungsgemäß garantierten Grundrechte aller Bürger stellt das Recht auf informationelle Selbstbestimmung[1] dar. Dieses wurde erstmalig im für die Belange des Datenschutzes wegweisenden Volkszählungsurteil des Bundesverfassungsgerichts von 1983 festgehalten und leitet sich aus Artikel 2 des Grundgesetzes ab. Die statistische Geheimhaltungspflicht setzt dieses – vergleichbar mit den Regelungen des Datenschutzgesetzes in anderen gesellschaftlichen Bereichen – für die amtliche Statistik um. So unterliegen die für statistische Zwecke erhobenen Daten einer engen Zweckbindung, von der nur in gesetzlich geregelten Sonderfällen abgewichen werden darf. Abgesehen von diesen besonderen Ausnahmen gilt grundsätzlich § 16 Abs. 1 BStatG (Bundesstatistikgesetz), der besagt: „Einzelangaben über persönliche und sachliche Verhältnisse, die für eine Bundesstatistik gemacht werden, sind von den Amtsträgern und für den öffentlichen Dienst besonders Verpflichteten, die mit der Durchführung von Bundesstatistiken betraut sind, geheim zu halten (…)“. Das bedeutet, dass die mit der Arbeit mit vertraulichen statistischen Daten betrauten Personen besondere Sorgfalt beim Umgang mit diesen üben müssen. Ausgehend von den Veröffentlichungen der amtlichen Statistik darf es nicht möglich sein, konkrete Rückschlüsse auf einzelne Erhebungspflichtige zu ziehen, indem diesen durch Dritte zuvor unbekannte Informationen zugeordnet werden können. Dabei wird keine inhaltliche Unterscheidung zwischen sensiblen und nicht-sensiblen Merkmalen vorgenommen, d. h. alle Angaben werden als gleichermaßen schutzbedürftig angesehen, unabhängig vom möglichen Schaden, der einem Betroffenen durch Bekanntwerden einer ihm zugehörigen Angabe entstehen könnte.[2]

1 „Das Grundrecht gewährleistet (...) die Befugnis des Einzelnen, grundsätzlich selbst über die Preisgabe und Verwendung seiner persönlichen Daten zu bestimmen. Einschränkungen dieses Rechts auf informationelle Selbstbestimmung sind nur im überwiegenden Allgemeininteresse zulässig.“ (Auszug aus dem „Volkszählungsurteil“ von 1983).

2 Unter analytischen Gesichtspunkten kann es jedoch auch im Kontext der amtlichen Statistik in Deutschland sinnvoll sein, zwischen sensiblen Merkmalen als denjenigen Angaben, die das Ziel eines Enthüllungsversuchs darstellen könnten, und nicht-sensiblen, aber identifizierenden Merkmalen, die die Identifizierung eines Merkmalsträgers und somit den Rückschluss auf dessen sensible Angaben erlauben, zu unterscheiden.

Zusätzlich zu den rechtlichen Regelungen und generellen ethischen Überlegungen zu Privatheit und Selbstbestimmung verfügt die amtliche Statistik auch unter rein rationalen Gesichtspunkten über ein starkes Eigeninteresse, die Angaben der einzelnen Befragten vor deren Offenlegung zu schützen, denn das Vertrauensverhältnis zwischen den Befragten und der amtlichen Statistik stellt eine unerlässliche Arbeitsgrundlage dar: Nur wenn die Erhebungspflichtigen mit Sicherheit davon ausgehen können, dass ihre Angaben vertraulich behandelt werden, ist im Gegenzug mit verlässlichen Antworten auf die gestellten Fragen – insbesondere in Bezug auf subjektiv als sensibel empfundene Angaben, wie beispielsweise Informationen zu Einkommens- und Vermögensverhältnissen oder zum Gesundheitszustand – zu rechnen. Im Fall von Erhebungen, bei denen eine Teilnahmepflicht besteht, wäre bei fehlendem Vertrauen ein höherer Anteil an falschen oder ungenauen Angaben bzw. gänzlich fehlenden Angaben (Item-Nonresponse) zu erwarten. Bei freiwilligen Erhebungen würde sich dies hingegen negativ auf die generelle Teilnahmebereitschaft auswirken, bei der von einem deutlichen Rückgang auszugehen wäre (Unit-Nonresponse). Infolgedessen entstünde zwangsläufig ein deutlich höherer Aufwand, um angestrebte Stichprobengrößen oder Quotenvorgaben zu erreichen und die Repräsentativität der Erhebungsergebnisse zu gewährleisten. In Zeiten tendenziell sinkender Teilnahmebereitschaft an freiwilligen Befragungen würde dies eine deutliche Erschwernis für die erfolgreiche Gewinnung einer hochwertigen Datenbasis darstellen.

Ausnahmen, in denen von der Geheimhaltungspflicht abgesehen werden kann

Von der allgemeingültigen Pflicht zur Geheimhaltung darf daher nur abgewichen werden, wenn hierfür auf gesetzlichem Wege besondere Ausnahmen definiert wurden: Solche Ausnahmen existieren unter anderem für die Übermittlung nicht-anonymisierter Einzeldaten an das Statistische Bundesamt oder andere Statistische Landesämter zur Produktion von Statistiken und deren Vorbereitung (§ 16 Abs. 2 BStatG) oder aber zur methodischen Weiterentwicklung (§ 3 Abs. 2 BStatG). Zudem dürfen Tabellen, die auch Einsen beinhalten können, ausschließlich für Planungszwecke an oberste Bundes- und Landesbehörden weitergegeben werden (§ 16 Abs. 4 BStatG). Die verwaltungstechnische Regelung von Einzelfällen ist den Datenempfängern hingegen untersagt. Ebenfalls sind Gemeinden dazu berechtigt, sofern sie über eine kommunale Statistikstelle verfügen, in rechtlich geregelten Fällen die sie betreffenden Einzeldaten zu erhalten und eigene statistische Auswertungen mit diesen durchzuführen (§ 16 Abs. 5 BStatG). Von diesem Recht wurde beispielsweise im Rahmen des Zensus 2011 Gebrauch gemacht. Ein besonderes Datenzugangsrecht genießt die unabhängige empirische Forschung in Form des sogenannten „Wissenschaftsprivilegs“ (§ 16 Abs. 6 BStatG). Dieses ermöglicht Angehörigen von Hochschulen und anderen vergleichbaren Forschungseinrichtungen die Arbeit mit faktisch anonymen Datenbeständen zur projektbezogenen Durchführung wissenschaftlicher Vorhaben.

Darüber hinaus entfällt die Pflicht zur Geheimhaltung, wenn es sich bei den betreffenden Informationen um Angaben über öffentliche Einrichtungen handelt, die bereits auf anderem Wege allgemein zugänglich gemacht wurden (§ 16 Abs. 1 S. 2 Nr. 2 BStatG). Dies gilt jedoch nicht für Angaben über private Merkmalsträger. Mit ausdrücklicher schriftlicher Einwilligung des Auskunftspflichtigen darf zudem gänzlich auf die Geheimhaltung verzichtet werden (§ 16 Abs. 1 S. 2 Nr. 1). Voraussetzung hierfür ist, dass der Auskunftspflichtige zuvor ausreichend über die Auswirkungen dieses Vorgehens informiert wurde. Auch Informationen, die bereits zu statistischen Ergebnissen aggregiert wurden (§ 16 Abs. 1 S. 2 Nr. 3 BStatG) – was den Regelfall in den Veröffentlichungen der amtlichen Statistik darstellt – und bei denen daher kein Rückschluss mehr auf die dahinter stehenden statistischen Einheiten möglich ist (§ 16 Abs. 1 S. 2 Nr. 4 BStatG), unterliegen grundsätzlich nicht der Geheimhaltungspflicht.

Bei der Verpflichtung zur statistischen Geheimhaltung handelt es sich übrigens um keine nationale Besonderheit, sondern diese stellt auch international ein grundlegendes Prinzip der amtlichen Statistik dar und wird entsprechend unter anderem im Rahmen des Verhaltenskodex des Europäischen Statistischen Systems der der „Fundamental Principles of Official Statistics" der Vereinten Nationen (United Nations Economic and Social Council 2014) thematisiert. Dabei wird ausdrücklich betont, dass es sich bei der Wahrung der statistischen Geheimhaltung – nicht zu Unrecht auch als „Statistikgeheimnis"[3] bezeichnet – um den Schutz eines grundlegenden Bürgerrechts handelt, welches auch angesichts des weit verbreiteten sorglosen Umgangs mit persönlichen Daten, beispielsweise im Internet in sozialen Netzwerken, nicht eingeschränkt werden darf – auch wenn diese Auffassung in der aktuellen Diskussion von verschiedener Seite wiederholt geäußert wurde (u. a. Krämer 2014, Rendtel 2014). Gerade angesichts der Auswirkungen der NSA-Affäre ist es umso mehr von Bedeutung für die Statistischen Ämter, sich von geheimdienstlichen Tätigkeiten abzugrenzen und den Schutz vertraulicher Angaben zu gewährleisten (Sarreither 2015).

2. Herausforderungen der statistischen Geheimhaltung in der Praxis

Das Ziel aller Maßnahmen zur statistischen Geheimhaltung ist es, zu verhindern, dass ein Außenstehender (auch etwas drastisch „Datenangreifer" genannt) durch Veröffentlichungen der amtlichen Statistik Informationen über einzelne, konkret identifizierbare statistische Einheiten – Personen, Unternehmen, Betriebe oder sonstige von den Statistischen Ämtern erfasste Merkmalsträger – gewinnen kann.

Ein besonderes Augenmerk sollte vor diesem Hintergrund darauf gerichtet werden, dass die amtliche Statistik in Deutschland ihre Daten heutzutage über verschiedenste Wege zugänglich macht (Leitner 2013): Neben der traditionellen Veröffentlichung von Tabellen in gedruckter oder digitaler Form sind Daten ebenfalls über statische oder flexible Datenbankanwendungen – beispielsweise GENESIS-Online (Carle 2005) oder die Zensusdatenbank (Tomann/Nickl 2013) –, in Form interaktiver Kartendarstellungen wie dem Statistikatlas (Kobl 2014) oder aber über die Forschungsdatenzentren (Rothe 2012) auch als faktisch anonymisierte Einzeldaten für wissenschaftliche Auswertungen beziehbar. Hinzu kommen Sonderauswertungen und Auftragsarbeiten, die auf kundenspezifischen Auftrag hin von den Statistischen Ämtern übernommen werden und von den regulären Standardveröffentlichungen abweichen. Darüber hinaus werden deutsche Mikrodaten auch an Eurostat übermittelt und dort unter anderem international zur Nutzung für Forschungszwecke zur Verfügung gestellt (Bujnowska 2013). Aus diesen modernen Informationsangeboten resultiert für die Nutzer der Daten der amtlichen Statistik eine Vielzahl neuer Anwendungsmöglichkeiten, zugleich bringen sie aber auch neue Herausforderungen für die Sicherstellung der statistischen Geheimhaltung mit sich.

3. Wann sind Daten wirklich anonym?

Oftmals wird, wenn es um den Schutz persönlicher Daten geht, auf die Anonymität der Datenverarbeitung verwiesen, die schon allein dadurch gewährleistet sei, dass keine identifizierenden Merkmale wie Name oder Adresse mehr in den Daten vorhanden wären. Es konnte jedoch wiederholt nachgewiesen werden, dass auch ohne das Vorhandensein solcher direkter Identifikatoren mit geringem Aufwand und anhand von nur wenigen vorliegenden Angaben Personen in Datenbeständen zweifelsfrei zu identifizieren sind und diesen die korrekten Daten zugeordnet werden können. So konnte beispielsweise Sweeney (2000) zeigen, dass es anhand einer Veröffentlichung vermeintlich anonymer Patientendaten von Krankenhäusern eines US-Bundesstaats – es handelte sich lediglich um die Merkmale Postcode, Geschlecht und Geburtsdatum – möglich war, rund drei Viertel der betroffenen Personen als einzigartige Kom-

3 Vergleichbar mit der Verletzung der ärztlichen oder anwaltlichen Schweigepflicht wird auch ein Bruch des Statistikgeheimnisses mit entsprechenden strafrechtlichen Sanktionen in Form von Geld- oder Freiheitsstrafen geahndet (§ 203 StGB).

bination dieser drei Merkmale darzustellen. Erforderlich hierfür war lediglich ein Abgleich mit anderen von öffentlichen Stellen verbreiteten Daten, in diesem Fall des von jedem erwerbbaren Wählerverzeichnisses. Ähnliches konnte jüngst für angeblich anonyme Daten, die bei der Benutzung von Kreditkarten erhoben werden, nachgewiesen werden, wobei in vielen Fällen bereits das bloße Vorliegen der Transaktionsdaten zu lediglich vier Einkäufen ausreichte, um anhand des hieraus resultierenden Profils valide Rückschlüsse auf 90 % der tatsächlich dahinterstehenden Personen zu ziehen (Montjoye et al. 2015). Vergleichbares gelang zuvor bereits anhand von durch Metadaten abbildbaren Mobilitätsmustern, wie sie bei der Nutzung von Mobiltelefonen anfallen (Montjoye et al. 2013).

Aber warum ist es überhaupt möglich, dass es mit so wenigen Daten gelingt, ohne Vorliegen direkter Identifikatoren eindeutige Zuordnungen der Daten zu den betreffenden Personen vorzunehmen? Die Erklärung verbirgt sich in den individuellen Ausprägungen von Merkmalskombinationen, die schon bei nur wenigen vorliegenden Merkmalen und Ausprägungen, eine Vielzahl unterschiedlichster Kombinationen ergeben können. So ergeben sich beispielsweise bei zehn Merkmalen, die lediglich zwei unterschiedliche Ausprägungen – im Falle des Geschlechts beispielsweise „weiblich“ und „männlich“ – annehmen können, 1024 (210) unterschiedliche Merkmalskombinationen, denen die einzelnen Merkmalsträger zugeordnet werden können. Geht man nun davon aus, dass es sich bei der Vielzahl der erfassten Merkmale nicht um binäre Variablen handelt, sondern dass jedes Merkmal unter Umständen dutzende oder sogar hunderte verschiedener Ausprägungen annehmen kann, so vervielfacht sich die Zahl der möglichen individuellen Merkmalskombinationen. Verfügt jedes Merkmal beispielsweise über zehn unterschiedliche Ausprägungen, so reichen bereits drei Merkmale aus, um auf annähernd dieselbe Zahl an Merkmalskombinationen (103 = 1000) wie im ersten Beispiel zu gelangen. Mit jedem hinzugenommenen Merkmal steigt die Wahrscheinlichkeit, dass ein einzelner Merkmalsträger eine individuelle, nur einmal vorkommende Merkmalskombination (auch als Uniqueness bezeichnet) aus für sich genommen unverdächtig erscheinenden Angaben aufweist, sprunghaft an. Die Individualität der einzelnen Merkmalsträger lässt diese aus der Masse hervorstechen. Dies wird auch von dem Umstand, dass viele der theoretisch möglichen Kombinationen empirisch nicht in Erscheinung treten, zumeist nur wenig abgemildert. Mit ein wenig entsprechendem Vorwissen – beispielsweise wenn es sich um Nachbarn, Bekannte, Kollegen oder aber auch um Prominente handelt[4] – ist es somit möglich, diese individuellen Einzelfälle zu identifizieren, sofern keine weitergehende Bearbeitung der Daten zu deren Schutz erfolgt. Hierdurch wird es einem Datenangreifer ermöglicht, sein Vorwissen, das er zur Identifizierung eingesetzt hat, um weitere, ihm zuvor unbekannte Informationen zu erweitern.

Aus diesem Grund ist das Löschen der direkten Identifikatoren aus dem vorliegenden Datenmaterial zwar eine zwingend notwendige, aber keineswegs hinreichende Voraussetzung für eine wirksame Anonymisierung statistischer Daten. Anonymität ist dementsprechend erst dann gegeben, wenn in den betreffenden Daten entweder keine einzigartigen, individuellen Kombinationen von Merkmalsausprägungen mehr vorliegen, beziehungsweise dann, wenn es unmöglich ist, korrekte Rückschlüsse auf die sich dahinter verbergenden, tatsächlichen statistischen Einheiten zu ziehen.

Unterschiedliche Formen der Anonymität

Das Ziel jeder Geheimhaltungsmaßnahme ist folglich die Herstellung von Anonymität. Hierbei wird zwischen verschiedenen Abstufungen unterschieden (vgl. Übersicht): So bezeichnet absolute Anonymität die Tatsache, dass es unter keinen Umständen möglich ist, anhand vorliegender Daten auf den dahinter stehenden individuellen

4 Dies gilt analog für Betriebe und Unternehmen, die anhand von branchenintern Wissen oder auf anderen Wegen veröffentlichten Angaben identifizierbar sein können. Auch Verzeichnisse und Datenbanken aller Art können, sofern sie Angaben zu einzelnen Merkmalsträgern enthalten, als potentielles Angriffswissen dienen.

Merkmalsträger zu schließen. Daten, die dieses Kriterium erfüllen, können ohne Einschränkung veröffentlicht und an Dritte weitergegeben werden. Dies gilt sowohl für die Veröffentlichung statistischer Ergebnisse als auch für entsprechend bearbeitete Mikrodaten (Public-Use-Files).

Weniger streng gefasst wird diese Anforderung bei der faktischen Anonymität, wie sie die Zielvorgabe für Daten darstellt, die der wissenschaftlichen Forschung bereitgestellt werden dürfen. Diese basiert nicht auf der Anforderung, eine mögliche Enthüllung unter allen nur denkbaren Umständen zu verhindern, sondern auf einer Risikoabschätzung anhand eines Kosten-Nutzen-Modells. Davon ausgehend werden Daten so bearbeitet, dass diese nur noch mit einem unverhältnismäßig hohen Aufwand an Zeit und Arbeitskraft einem konkreten Merkmalsträger zugeordnet werden können, sodass sich aus der Sicht eines rational agierenden Datenangreifers ein Enthüllungsversuch als nicht lohnenswert erweist. Durch dieses Vorgehen wird der notwendige Eingriff in die Daten vergleichsweise gering gehalten, ohne dass hierdurch unkalkulierbare Risiken hinsichtlich des Schutzes der Daten in Kauf genommen werden müssten. Mit berücksichtigt werden bei dieser Abwägung darüber hinaus nicht nur die Eigenschaften der Daten, sondern auch rechtliche, technische und organisatorische Regelungen, die dazu dienen können, eine missbräuchliche Verwendung der Daten zu verhindern. Dabei kann es sich um Maßnahmen wie das Schließen eines Nutzungsvertrags, die Verpflichtung der Datenempfänger zur statistischen Geheimhaltung nach § 16 Abs. 7 BStatG, die Ahndung von Zuwiderhandlungen mit Geld- und Freiheitsstrafen nach § 203 StGB, die technische Abschottung von Arbeitsplätzen und Ähnliches handeln. Im Gegenzug ist es dafür möglich, die notwendigen Eingriffe in die Daten zu reduzieren und den Datennutzern hierdurch ein Mehr an Analysepotential zur Verfügung stellen zu können. Die Anwendung dieses Konzepts bezieht sich jedoch ausschließlich auf Mikrodaten, nicht aber auf Auswertungstabellen.[5]

Für die Arbeit der Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder stellt die faktische Anonymität daher eine zentrale Grundlage dar, die es erlaubt, empirisch Forschenden eine Vielzahl statistischer Einzeldaten zu Analysezwecken bereitzustellen.[6]

Die formale Anonymisierung schließlich bezeichnet den geringsten Grad der Geheimhaltung; hierbei werden lediglich die direkten Identifikatoren wie Name, Adresse, Matrikelnummer oder Ähnliches aus dem Datenmaterial entfernt; weitergehende Geheimhaltungsmaßnahmen kommen dabei nicht zum Einsatz. Aus diesem Grund ist diese Form der Anonymisierung nicht ausreichend, wenn Daten an Externe weitergegeben werden sollen.

4. Geheimhaltungsverfahren

Um die statistische Geheimhaltung zu gewährleisten, steht den amtlichen Statistikern eine Reihe unterschiedlicher Verfahren zur Verfügung. Anhand des Zeitpunkts der Anwendung – vor oder nach Erstellung der Auswertungstabellen (pre-tabular oder post-tabular) – und der Art des Eingriffs (informationsreduzierend oder datenverändernd) – lassen sich hierbei die unterschiedlichen Methoden klassifizieren:

Pre-tabulare Verfahren setzen dabei bereits auf Ebene der Original-Einzeldaten einer Statistik an, wohingegen post-tabulare Verfahren erst nach Erstellung der Auswertungsergebnisse auf die fertigen Tabellen angewandt werden. Pre-tabulare Geheimhaltung wird auch als Anonymisierung bezeichnet.

Die zweite Unterscheidung bezieht sich auf die Art und Weise, auf die die statistische Geheimhaltung sichergestellt wird: Informationsreduzierende Verfahren stellen dabei den meistgenutzten Ansatz dar. Mittels Löschung von Merkmalen oder

5 Das Konzept faktisch anonymer Tabellen wurde in der Vergangenheit zwar vereinzelt auf dessen Umsetzbarkeit in der Praxis hin untersucht (Hochgürtel/Weiss 2011; Hochgürtel 2013), wurde aber letztlich nicht weiterverfolgt.

6 Als Basis diente hierfür insbesondere ein gemeinsam von Wissenschaft und amtlicher Statistik durchgeführtes Forschungsprojekt, bei dem die Realisierbarkeit einer rechtskonformen faktischen Anonymisierung anhand der Einzeldaten des Mikrozensus in der Praxis erprobt wurde (Müller et al. 1991).

auch ganzer Merkmalsträger, der Zusammenfassung von Kategorien oder der Unterdrückung von Angaben wird das Auftreten kritischer Fälle reduziert beziehungsweise gänzlich verhindert. Auch die Zensierung von Werten, die einen bestimmten Schwellenwert übersteigen (Top-Coding) oder unterschreiten (Bottom-Coding), fällt in diese Verfahrensgruppe. Ebenfalls informationsreduzierend wirkt sich die Durchführung einer Stichprobenziehung aus. Hieraus resultiert, dass alle Erhebungen, bei denen es sich ursprünglich um Stichprobenerhebungen handelt – beispielsweise beim Mikrozensus oder der Einkommens- und Verbrauchsstichprobe –, sich unter Geheimhaltungsgesichtspunkten deutlich unkritischer darstellen als dies bei Vollerhebungen der Fall ist, da das Auftreten einer einzigartigen Merkmalskombination in einer Stichprobe nicht zwingend bedeutet, dass es sich auch in der Gesamtpopulation um eine solche handelt. Das Auffinden eines Merkmalsträgers mit einer bestimmten Merkmalskombination reicht aus Sicht eines Datenangreifers in diesem Fall also nicht aus; er benötigt darüber hinausgehend weitere Informationen, um sich sicher sein zu können, dass es sich wirklich um den gesuchten Merkmalsträger handelt und nicht um einen statistischen Doppelgänger.

Eine grundlegend andere Herangehensweise verfolgen die datenverändernden Geheimhaltungsverfahren: Mittels möglichst geringer Eingriffe in die Daten – entweder auf Basis der ursprünglichen Mikrodaten oder der bereits fertiggestellten Auswertungstabellen – werden diese so verändert, dass möglichst keine geheimhaltungsrelevanten Problemfälle mehr im Datenmaterial beziehungsweise in den daraus erzeugten Ergebnistabellen auftauchen. Pre-tabular kommen hierfür beispielsweise die Vertauschung von Merkmalsausprägungen zwischen ähnlichen Merkmalsträgern (Swapping) oder Mikroaggregation zum Einsatz. Ein Beispiel für die letztgenannte Verfahrensgruppe stellt das SAFE-Verfahren (Höhne 2003) dar, das unter anderem im Rahmen der Veröffentlichung der Ergebnisse des Zensus 2011 zum Einsatz kam (Giessing et al. 2014). Post-tabular können hingegen beispielsweise Rundungs- oder Zufallsüberlagerungsverfahren eingesetzt werden, um Tabellen, die Aufdeckungsrisiken enthalten, nachträglich geheimhaltungskonform zu machen. Die Löschung von Informationen ist hierbei nicht notwendig; stattdessen wird durch die Veränderung gegenüber den Echtwerten das potentiell vorhandene Angriffswissen eines Dritten, das zur Identifikation einzelner statistischer Einheiten eingesetzt werden könnte, entwertet. Selbst im Falle einer geglückten Identifikation würde auf Seiten des Datenangreifers Unsicherheit darüber bestehen, ob es sich bei der zugeordneten Information tatsächlich um den echten Wert handelt – und wenn nicht, wie stark er von diesem abweicht.

5. Prototypischer Ablauf einer Geheimhaltungsprüfung am Beispiel einer Häufigkeitstabelle

Bei der Durchführung der statistischen Geheimhaltung, wie sie in der amtlichen Statistik im Regelfall ausgehend von einer erstellten Auswertungstabelle erfolgt, handelt es sich um einen zweistufigen Prozess, in dessen Verlauf zuerst die in der betreffenden Tabelle möglicherweise enthaltenen kritischen Felder identifiziert und in einem Folgeschritt geheim gehalten werden. Im Beispiel wird von der Anwendung eines post-tabularen, informationsreduzierenden Geheimhaltungsverfahrens ausgegangen, wie es heute den Regelfall in den meisten Statistikbereichen darstellen dürfte.

Schritt 1:
Die Identifikation potentieller Risiken

Als Beispiel hierfür dient im Folgenden eine fiktive, aus Gründen der besseren Verständlichkeit möglichst einfach gehaltene Tabelle, die die Merkmalsträger – beispielsweise die Einwohner einer Gemeinde – nach Altersgruppen und Geschlecht ausweist (vgl. Tabelle 1). Die entsprechenden Arbeitsschritte lassen sich jedoch selbstverständlich analog auf komplexere Tabellen übertragen.

Tab. 1 Beispiel für eine fiktive Häufigkeitstabelle

Bevölkerung nach Alter und Geschlecht

Alter	Weiblich	Männlich	**Insgesamt**
0 bis 14	3	3	**6**
14 bis 49	8	9	**17**
50 bis 75	12	9	**21**
75 oder älter	4	1	**5**
Insgesamt	**27**	**22**	**49**

In einem ersten Schritt wird anhand statistikspezifischer Regeln festgestellt, ob ein Aufdeckungsrisiko in der zu veröffentlichten Tabelle gegeben ist und welche konkreten Tabellenfelder hiervon betroffen sind. Die innerhalb der amtlichen Statistik verbreitetste Regel zur Identifizierung solcher kritischer Fälle stellt die Mindestfallzahlregel dar. Diese legt fest, dass innerhalb einer Fallzahltabelle die in einem Tabellenfeld ausgewiesene Häufigkeit nicht geringer als ein festgelegter Wert n sein darf. Für gewöhnlich wird in der amtlichen Statistik von n = 3 ausgegangen, d. h. dass alle ausgewiesenen Fallzahlen mindestens dem Wert 3 entsprechen müssen, um in einer Veröffentlichung als unkritisch zu gelten.[7] Alle Angaben, die die festgesetzte Mindestfallzahl unterschreiten, müssen hingegen geheim gehalten werden.

Schritt 2:
Anwendung eines Geheimhaltungsverfahrens
Hat man nun mögliche Aufdeckungsrisiken identifiziert, so wird in einem zweiten Schritt ein auf die jeweilige Fachstatistik, die Art der Daten und der Veröffentlichung sowie die Nutzergruppe abgestimmtes Geheimhaltungsverfahren auf die betreffenden Daten angewendet. Bei der grundsätzlichen Entscheidung für oder gegen ein bestimmtes Verfahren müssen dabei im Vorfeld verschiedene Aspekte gegeneinander abgewogen werden: So muss ein Geheimhaltungsverfahren in allererster Linie Einzelangaben zuverlässig vor einer potentiellen Aufdeckung schützen, soll aber zugleich nur so wenig wie möglich in den informativen Gehalt der Daten eingreifen, um deren Qualität möglichst wenig zu beeinträchtigen – zwangsläufig ergibt sich hieraus ein Konflikt zwischen den zwei sich widersprechenden Zielen des Schutzes der Daten auf der einen und des Erhalts der Datenqualität auf der anderen Seite. Hinzu kommen Aspekte wie die möglichst einfache praktische Integration der Verfahren in die Abläufe innerhalb der Statistischen Ämter und die Verständlichkeit des Verfahrens und seiner Auswirkungen für die Nutzer der Daten.

Im nachfolgenden Beispiel wird anhand der fiktiven Ergebnistabelle aus dem vorigen Abschnitt die Anwendung des Zellsperrverfahrens, bei dem es sich um das meistverwendete Geheimhaltungsverfahrens innerhalb der amtlichen Statistik handelt, auf Basis der Mindestfallzahlregel (mit n = 3) demonstriert (vgl. Tabelle 2). Zu sperrende Werte sind rot markiert; Sperrungen werden durch einen ebenfalls roten Punkt dargestellt.

Tab. 2 Beispiel für die Primärsperrung

Bevölkerung nach Alter und Geschlecht

Alter	Weiblich	Männlich	**Insgesamt**
0 bis 14	3	3	**6**
14 bis 49	8	9	**17**
50 bis 75	12	9	**21**
75 oder älter	4	1	**5**
Insgesamt	**27**	**22**	**49**

In der Beispieltabelle findet sich nur ein Tabellenfeld, das eine Häufigkeit ausweist, die den Wert 3 unterschreitet, und aus diesem Grund primär gesperrt werden muss.

Ziel ist neben der Sperrung des eigentlichen kritischen Tabellenfelds (Primärsperrung) die Verhinderung der Rückrechenbarkeit der vorgenommenen Löschung durch die Vornahme weiterer Sperrungen (Sekundärsperrungen). Dies ist notwendig, da Tabellen mit Randsummen zwangsläufig ein lineares Gleichungssystem darstellen, bei dem sich die Innenfelder zu Zeilen- und Spalten-

7 Die Mindestfallzahl von n = 3 ergibt sich dabei folgendermaßen: Wird in einem Innenfeld einer Tabelle die Häufigkeit n = 1 ausgewiesen, so ist offensichtlich, dass es sich hierbei um einen Einzelfall handelt. Beträgt die ausgewiesene Anzahl hingegen n = 2, so bedeutet dies, da jeder Merkmalsträger seine eigene Ausprägung kennt, dass jeder der beiden mit diesem Vorwissen Rückschlüsse auf den jeweils anderen ziehen kann. Erst ab einer Häufigkeit von drei Merkmalsträgern ist dies nicht mehr möglich, sofern davon ausgegangen wird, dass nicht n – 1 Merkmalsträger ihr Vorwissen teilen und so gemeinsam Rückschlüsse auf den verbleibenden Merkmalsträger ziehen können.

summen aufaddieren. Wird nun lediglich ein einzelnes Tabellenfeld gesperrt, so wäre es ohne weiteres möglich, durch Subtraktion der Werte in den verbliebenen Tabellenfeldern derselben Zeile oder Spalte von der jeweiligen Randsumme, den gesperrten Wert rückzurechnen. Um dies zu verhindern, müssen daher mindestens ein Tabellenfeld in derselben Zeile, ein weiteres in derselben Spalte sowie dasjenige Tabellenfeld, in dem die Zeile und die Spalte der beiden zuvor genannten Felder aufeinander treffen, ebenfalls gesperrt werden. Die Anordnung der Sperrpartner bildet dabei ein Viereck (vgl. Tabellen 3 und 4). Grundsätzlich sollten aufgrund des daraus resultierenden hohen Informationsverlusts nach Möglichkeit keine Zellen, die Randsummen beinhalten, sondern ausschließlich Innenfelder einer Tabelle gesperrt werden.

Tab. 3 Beispiel für die Sekundärsperrung

Bevölkerung nach Alter und Geschlecht

Alter	Weiblich	Männlich	**Insgesamt**
0 bis 14	3	3	**6**
14 bis 49	8	9	**17**
50 bis 75	12	9	**21**
75 oder älter	4	•	**5**
Insgesamt	**27**	**22**	**49**

Tab. 4 Beispiel für die geheimgehaltene Tabelle mit Primär- und Sekundärsperrung

Bevölkerung nach Alter und Geschlecht

Alter	Weiblich	Männlich	**Insgesamt**
0 bis 14	•	•	**6**
14 bis 49	8	9	**17**
50 bis 75	12	9	**21**
75 oder älter	•	•	**5**
Insgesamt	**27**	**22**	**49**

Die Vornahme der Sekundärsperrung erweist sich dabei oftmals als deutlich anspruchsvoller als die Umsetzung der primären Geheimhaltung, da aus Gründen der Datenqualität eine sorgfältige Auswahl der jeweiligen Sperrpartner vonnöten ist. Auch entsteht durch die Sekundärsperrung zumeist eine deutlich stärkere Beeinträchtigung des informativen Gehalts einer Tabelle als dies durch die vorgenommene Primärsperrung der Fall ist. Erschwerend kommt hinzu, dass die Realisierung der Zellsperrung in den meisten Fällen weitgehend manuell durchgeführt wird und bislang nur in bestimmten Fällen automatisiert werden kann. Für die computergestützte Durchführung von primärer und sekundärer Zellsperrung einsetzbare Programme wie Tau-Argus (De Wolf 2013; Hundepool et al. 2010: 131ff.) oder sdcTables (Templ 2008) kommen innerhalb der amtlichen Statistik in Deutschland bislang nur selten zum Einsatz. Damit einher geht ein insbesondere in umfangreichen und komplexen Tabellen prinzipbedingtes Fehlerrisiko, dem durch die Anwendung des Vier-Augen-Prinzips, d. h. der Prüfung durch mindestens zwei unterschiedliche Bearbeiter, versucht wird entgegenzuwirken.

Darüber hinaus müssen im Rahmen einer tabellenübergreifenden Geheimhaltung Sperrungen über das gesamte Tabellenprogramm einer Statistik konsistent vorgenommen werden. Es ist folglich notwendig, identische Tabellenfelder, die in einer Tabelle gesperrt wurden, auch in allen anderen Tabellen, zu unterdrücken – unabhängig davon, ob es sich dabei um ein primär oder sekundär geheim gehaltenes Tabellenfeld handelt. Unterbleibt dies, so ist es gegebenenfalls möglich, einer Tabelle Angaben zu entnehmen, diese in eine geheim gehaltene Tabelle zu übertragen und anhand der Additivität von Tabellen die gesperrten Felder wiederherzustellen. Gerade bei umfangreichen Veröffentlichungen und besonders auch im Fall von individuellen Sonderauswertungen kann es für die Verantwortlichen eine große Herausforderung und einen hohen Arbeitsaufwand darstellen, dies zu verhindern. Auch durch die unabgestimmte Veröffentlichung von Tabellen zu denselben Merkmalen durch unterschiedliche Stellen kann es zum Auftreten von Enthüllungsrisiken kommen, wenn die Sperrungen unterschiedlich umgesetzt werden. Ein möglicher Ausweg hierzu wird in einer verbesserten Abstimmung unter den Akteuren innerhalb des Statistischen Verbundes sowie in der Anwendung datenverändernder Geheimhaltungsverfahren gesehen.

Exkurs: Das Randsummenkriterium

Eine weitere Regel, die jedoch nur vergleichsweise selten Anwendung findet, stellt das sogenannte Randsummenkriterium – auch als Randwertregel bezeichnet – dar. Durch dieses wird dem Umstand Rechnung getragen, dass auch wenn ein Tabellenfeld keine Anzahl kleiner n aufweist, bei bestimmten Tabellenkonstellationen dennoch ein Aufdeckungsrisiko gegeben sein kann. Ein solches liegt dann vor, wenn innerhalb einer Tabellenzeile oder -spalte alle Merkmalsträger in dieselbe Kategorie fallen. Somit ist es möglich, ohne genauere Kenntnis des individuellen Merkmalsträgers ein Zusatzwissen über diesen zu erhalten, wofür man lediglich über die Kenntnis verfügen muss, dass dieser einer bestimmten Gruppe von Merkmalsträgern angehört. Man spricht in diesem Fall vom Vorliegen eines Randwertproblems.

Im dargestellten Beispiel (vgl. Tabelle 5) wird das geschlechtsspezifische Prüfungsergebnis innerhalb eines fiktiven Studiengangs dargestellt. Das Enthüllungsrisiko im vorliegenden Fall liegt darin, dass alle männlichen Studierenden des Faches die abgelegte Prüfung nicht bestanden haben, wohingegen die weiblichen Studierenden sich auf beide mögliche Prüfungsresultate verteilen. Hieraus folgt, dass allein anhand der Kenntnis des Geschlechts über jeden männlichen Studierenden mit Sicherheit die Aussage gemacht werden kann, dass dieser die Prüfung nicht bestanden hat, ohne sonstige individuelle Informationen über diesen zu benötigen. Darüber hinaus ist bereits an der Information „Prüfung bestanden“ im Gegenzug ersichtlich, dass die Prüfung von einer Frau abgelegt worden sein muss. In diesem Fall würde die Durchführung der Geheimhaltung zu den im Folgenden dargestellten Sperrungen führen (vgl. Tabelle 6):

Im Vergleich zur Mindestfallzahlregel wird die Randwertregel nur selten angewandt, obwohl sie als Alternative zur Mindestfallzahlregel einen wichtigen Beitrag zur Sicherstellung der statistischen Geheimhaltung leisten kann, indem sie kritische Fälle, die durch Anwendung der Mindestfallzahlregel nicht erkannt werden würden, identifizierbar macht und im Gegenzug unnötige Sperrungen verhindern kann. Wichtig ist dabei zu beachten, dass Randwertprobleme immer unter inhaltlichen Gesichtspunkten betrachtet werden müssen: So gibt es zahlreiche Konstellationen, unter denen aus logischen Gründen nur bestimmte Randwerte überhaupt möglich sind. Eine Sperrung ist in diesen Fällen daher weder notwendig noch zielführend.

6. Zusammenfassung und Ausblick

Im Rahmen des vorliegenden ersten Teils des Beitrags wurden die rechtlichen Grundlagen und Rahmenbedingungen der statistischen Geheimhaltung dargestellt. Darüber hinaus wurde ein kurzer Überblick über die beiden unterschiedlichen Gruppen von Verfahren, die zur Sicherstellung der statistischen Geheimhaltung zur Verfügung stehen, gegeben, sowie die Geheimhaltung von Häufigkeitstabellen ausführlicher dargestellt. In einem Folgebeitrag soll darauf aufbauend die Geheimhaltung von Wertetabellen vorgestellt sowie auf aktuelle Entwicklungen und zukünftige Herausforderungen im Bereich der statistischen Geheimhaltung, mit denen sich die amtliche Statistik konfrontiert sieht, eingegangen werden.

Literaturangaben

Bundesverfassungsgerichts-Urteil vom 15. Dezember 1983, 1 BVR 209/83, 1 BVR 269/83, 1 BVR 362/83, 1 BVR 420/83, 1 BVR 440/83, 1 BVR 484/83.

Bujnowska, A. (2013), Modes of access to EU microdata in the new legal frameworks. Working paper. Joint UNECE/Eurostat work session on statistical data confidentiality, 28-30. Oktober 2013, Ottawa.

Carle, M. (2005), GENESIS-Online (Bayern) – Das statistische Informationssystem im Internet. Bayern in Zahlen 11/2005, S. 444-450.

de Wolf, P.-P. (2013), Open source software Argus. Working paper. Joint UNECE/Eurostat work session on statistical data confidentiality, 28-30. Oktober, Ottawa 2013.

Europäisches Statistisches System (2011), Verhaltenskodex für europäische Statistiken für die nationalen und gemeinschaftlichen statistischen Stellen, verbesserte Auflage.

Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 2 des Gesetzes vom 9. Juni 2005 (BGBl. I S. 1534).

Giessing, S./Heinzl, F./Kleber, B./Wilke, A. (2014), Geheimhaltung beim Zensus 2011. Bayern in Zahlen 11/2014, S. 673-681.

Hochgürtel, T./Weiss, E. (2011), De facto anonymity in results. Working paper. Joint UNECE/Eurostat work session on statistical data confidentiality, 26.-28. Oktober 2011, Tarragona.

Hochgürtel, T. (2013), Die Messung der Enthüllungsrisiken von Ergebnissen statistischer Analysen. Arbeitspapier Nr. 3. Institut für Diskrete Mathematik und Angewandte Statistik der Hochschule für Technik und Wirtschaft des Saarlandes.

Höninger, J. (2015), Mindestfallzahlregel versus Randwertregel – Eine Betrachtung der Enthüllungsrisiken. Zeitschrift für amtliche Statistik Berlin Brandenburg 02/2015 (im Erscheinen).

Höhne, J. (2003), SAFE – ein Verfahren zur Geheimhaltung und Anonymisierung statistischer Einzelangaben. Berliner Statistik Monatsschrift 03/2003, S. 96-107.

Kobl, D. (2014), Der neue Statistikatlas Bayern. Bayern in Zahlen 4/2014, S. 156-163.

Krämer, W. (2014), Kommentar zu Ulrich Rendtel – Vom Datenangreifer zum zertifizierten Wissenschaftler. AStA Wirtschafts- und sozialstatistisches Archiv Vol 8. (4), S. 203-204.

Leitner, C. (2013), Daten der Statistischen Ämter des Bundes und der Länder. In: Arbeitsgruppe Regionale Standards (Hg.): Regionale Standards. Ausgabe 2013. Eine gemeinsame Empfehlung des ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V., der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e. V. (ASI) und des Statistischen Bundesamtes. GESIS-Schriftenreihe Band 12. Mannheim/Köln: GESIS, S. 269-277.

Montjoye de, Y.-A./Hidalgo C. A./Verleysen, M./Blondel, V. D. (2013), Unique in the Crowd: The privacy bounds of human mobility. Science Reports 3: 1376.

Montjoye de, Y.-A./Radaelli, L./Singh, V. K./Pentland, A. (2015), Unique in the shopping mall – On the reidentifiability of credit card metadata. Science Vol. 347, Issue 6221, S. 536-539.

Müller, W./Blien, U./Knoche, P./Wirth, H. (1991), Die faktische Anonymität von Mikrodaten. Stuttgart: Metzler/Poeschel.

Rendtel, U. (2014), Vom potenziellen Datenangreifer zum zertifizierten Wissenschaftler – Für eine Neugestaltung des Wissenschaftsprivilegs beim Datenzugang. AStA Wirtschafts- und sozialstatistisches Archiv Vol 8. (4), S. 183-197.

Rothe, P. (2012), 10 Jahre Forschungsdatenzentrum der Statistischen Ämter der Länder – Ein Blick auf Vergangenheit, Gegenwart und Zukunft der Forschungsdateninfrastruktur der amtlichen Statistik in Deutschland. Bayern in Zahlen 7/2012, S. 492-500.

Sarreither, D. (2015), Amtliche Statistik wird sich behaupten. Ein Plädoyer für Professionalität. Wirtschaft und Statistik 1 (2015), S. 9-17.

Strafgesetzbuch (StGB) in der Fassung der Bekanntmachung vom 13. November 1998 (BGBl. IS. 3322), das zuletzt durch Artikel 1 des Gesetzes vom 21. Januar 2015 (BGBl. I S. 10) geändert worden ist.

Sweeney, L. (2000), Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh.

Templ, M. (2008), Statistical Disclosure Control for Microdata Using the R-Package sdcMicro. Transactions on data Privacy 1, S. 67-85.

Tomann, J./Nickl, A. (2013), Zensus 2011: Die Zensusdatenbank. Bayern in Zahlen 4/2013, S. 186-189.

United Nations Economic and Social Council (2014), Fundamental Principles of Official Statistics. Download unter http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf, abgerufen am 23. März 2015.

Statistische Geheimhaltung – Der Schutz vertraulicher Daten in der amtlichen Statistik

Teil 2: Herausforderungen und aktuelle Entwicklungen

Patrick Rothe

Im ersten Teil des Beitrags wurden die rechtlichen Grundlagen der statistischen Geheimhaltung vorgestellt und ein grundsätzlicher Überblick über die verschiedenen Verfahren, mit denen der Geheimhaltungspflicht nachgekommen werden kann, gegeben. Zudem wurde der Umgang mit Häufigkeitstabellen detaillierter vorgestellt. Als Fortsetzung hierzu widmet sich Teil 2 der Geheimhaltung von Wertetabellen. Darüber hinaus soll ein kurzer Ausblick auf aktuelle Entwicklungen und momentane sowie zukünftige Herausforderungen im Bereich der amtlichen Statistik, wie datenverändernde Geheimhaltungsverfahren, den Umgang mit entstehendem Informationsverlust oder die Veröffentlichung georeferenzierter Daten, gegeben werden.

Geheimhaltung in Wertetabellen

Insbesondere im Bereich der Wirtschaftsstatistiken sind Wertetabellen eine weit verbreitete Darstellungsform, die eine gegenüber Häufigkeitstabellen abweichende Prüfung von möglichen Enthüllungsrisiken erfordert. Sie dienen dazu, beispielsweise Umsätze, Steuerbeträge oder Einkommen und Verdienste in aggregierter Form darzustellen. Das größte Problem liegt hierbei im Vergleich zu Häufigkeitstabellen nicht so sehr in der Aufdeckung exakter Werte durch einen Außenstehenden – wobei diese natürlich ebenfalls verhindert werden muss –, sondern vielmehr darin, dass bereits die Ermittlung ungefährer Angaben, die dem echten Beitrag eines Auskunftgebenden vergleichsweise nahe kommen, ein Aufdeckungsrisiko darstellen kann. So könnte beispielsweise einem Mitbewerber schon die Enthüllung eines ungefähren Schätzwerts des Konkurrenten einen unzulässigen Vorteil verschaffen, ohne dass für diesen hierfür die Kenntnis des exakten Werts vonnöten wäre.

Eine solche Gefahr ist besonders dann gegeben, wenn ein Merkmalsträger einen überproportional großen Anteil zu einem aggregierten Wert beiträgt und dieser Beitrag dadurch relativ nahe beim veröffentlichten Gesamtwert liegt. Dies könnte beispielsweise der Fall sein, wenn in einem Wirtschaftszweig nahezu alle Unternehmen sehr wenig, ein einziges Unternehmen jedoch den Großteil zur Gesamtsumme der Umsätze beiträgt. Man spricht in diesem Fall vom Vorliegen eines Dominanzfalls. In einer solchen Konstellation ist es für jedes der Unternehmen möglich, eine Schätzung des Umsatzes des dominierenden Unternehmens vorzunehmen, indem der eigene Umsatz vom ausgewiesenen Gesamtwert subtrahiert wird. Die beste Schätzung erzielt dabei das Unternehmen mit dem zweitgrößten Umsatz. Zugleich wird die Schätzung umso genauer ausfallen, je geringer die Summe der Umsätze der verbleibenden Unternehmen ist. Um derartige Fallkonstellationen zu erkennen, werden in der amtlichen Statistik sogenannte Dominanz- beziehungsweise Konzentrationsregeln angewandt. Heutzutage kommt dabei vorzugsweise die p%-Regel zum Einsatz. Die ebenfalls noch im Einsatz befindlichen (1, k)- und (2, k)-Regeln gelten als veraltet und sollten gemäß Beschluss der Leiter der Statistischen Ämter

nicht mehr angewandt werden, weshalb auf diese im Folgenden nicht näher eingegangen wird.[1]

Bei Anwendung der p%-Regel muss Folgendes gelten: Der betreffende Zellwert wird geheim gehalten, wenn die Differenz zwischen dem Zellwert X und dem größten und zweitgrößten Beitrag X_1 und X_2 nicht mindestens p Prozent vom größten Beitrag X_1 beträgt:

$$X - x_2 - x_1 < \frac{p}{100} * x_1$$

Anders ausgedrückt: Die p%-Regel prüft, ob der sich ergebende Schätzfehler desjenigen Beitragenden mit dem zweitgrößten Wert mindestens p Prozent beträgt. Sofern dies der Fall ist, kann der entsprechende Gesamtwert veröffentlicht werden. Ist dem jedoch nicht so, so müssen Geheimhaltungsmaßnahmen durchgeführt werden. Als positiver Nebeneffekt der Anwendung der p%-Regel ist zu vermerken, dass diese implizit auch die Vorgaben der Mindestfallzahlregel mit n = 3 berücksichtigt, so dass bei positiver Dominanzprüfung zugleich sichergestellt ist, dass mindestens drei unterschiedliche Erhebungseinheiten zum Gesamtwert beitragen. Eine darüber hinaus gehende Fallzahlprüfung ist somit nicht nötig. Der Parameter p muss von den jeweiligen Fachabteilungen statistikspezifisch festgelegt werden und darf nicht veröffentlicht werden, um hierdurch mögliche Rückschlüsse zu vermeiden.

Im nachfolgend geschilderten fiktiven Beispiel soll in einer Wertetabelle die Summe der Umsätze innerhalb einzelner Wirtschaftszweige abgebildet werden. Um das Vorliegen von Dominanzfällen prüfen zu können, ist darüber hinaus die Kenntnis der entsprechenden Einzelangaben der zum jeweiligen Wert beitragenden Unternehmen – im folgenden Beispiel A, B und C genannt – notwendig: Unternehmen A weist genauso wie Unternehmen B einen Umsatz von 15 000 Euro aus; im Fall von Unternehmen C beträgt er hingegen 200 000 Euro. Der entsprechend aggregierte Gesamtwert für den Wirtschaftszweig beträgt folglich 230 000 Euro.

Wenn A nun versuchen sollte, den gemeldeten Umsatz von C zu schätzen, würde A hierzu seinen eigenen Beitrag vom Gesamtumsatz abziehen. A erhält dadurch in Abwesenheit weiteren Vorwissens den bestmöglichen Schätzwert für den Umsatz von C. Je kleiner zudem der Beitrag von B zum Gesamtumsatz ausfällt, desto näher befindet sich der Schätzwert am wahren Wert von C. Durch das Vorhandensein brancheninternen Wissens kann A die Schätzung dabei gegebenenfalls noch weiter verbessern. Um dem entgegenzuwirken, soll in diesem Beispiel unter Verwendung der p%-Regel mit p = 10 geprüft werden, ob im vorliegenden Fall eine Veröffentlichung des Gesamtumsatzes gefahrlos möglich ist, oder ob dieser geheim gehalten werden muss. Hierbei ergibt sich folgendes Bild:

$$230\,000 - 15\,000 - 200\,000 < \frac{10}{100} * 200\,000$$

Der geforderte Mindestabstand zum Wert des größten Beitragenden C beträgt 20 000 Euro und übersteigt somit die tatsächliche Differenz von 15 000 Euro. Die Schätzung von A für den Umsatz von C beträgt 215 000 Euro und fällt damit zu präzise aus. Somit liegt ein Dominanzfall vor; der Gesamtumsatz im vorliegenden Beispiel ist geheimhaltungsbedürftig und darf nicht ausgewiesen werden. Werden in einer Tabelle die Umsätze mehrerer Wirtschaftszweige samt Summen dargestellt, müssen dementsprechend, analog zum geschilderten Beispiel zum Umgang mit Häufigkeitstabellen, Primär- und Sekundärsperrungen vorgenommen werden, um die Rückrechenbarkeit zu verhindern.

Datenverändernde Geheimhaltungsverfahren

Angesichts einer zunehmenden Anzahl neuer Möglichkeiten zur flexiblen Auswertung und Darstellung statistischer Datenbestände – beispielsweise mithilfe eines sogenannten Data-Warehouses – stößt die Umsetzung traditionell verwendeter informationsreduzierender Geheimhaltungsverfahren wie der Zellsperrung zunehmend an Grenzen. Schließlich war diese ursprünglich als Instrument zur Bearbeitung fixer Tabellen, die in gedruckter

1 Erläuterungen zur (1, k)- und (2, k)-Regel finden Interessierte u.a. in Hundepool et al. 2010 (S. 117 ff.).

Form veröffentlicht wurden, konzipiert worden. Für eine flexible Kombination von Merkmalen und die dynamische Erstellung von Auswertungstabellen außerhalb des Standardveröffentlichungsprogramms der Statistischen Ämter, wie sie durch moderne Auswertungsdatenbanken ermöglicht wird, ist sie hingegen ungeeignet. Als ein möglicher Ausweg hierfür wird der Rückgriff auf datenverändernde Verfahren gesehen, die entweder pre-tabular zur Erzeugung eines anonymisierten Mikrodatenbestands oder post-tabular bei der Veränderung der erzeugten Tabellen zum Einsatz kommen. Diese Ansätze versprechen eine Lösung von Problemen wie dem der tabellenübergreifenden Geheimhaltung und ermöglichen eine flexible Generierung geheim gehaltener Ergebnisse in Echtzeit, zugleich erfordern sie aber ein Umdenken bei Anwendern und Nutzern innerhalb und außerhalb der amtlichen Statistik. Eine der nachvollziehbarsten Verfahrensgruppen, die hierbei in Frage kommen, stellen Rundungsverfahren dar, auf die exemplarisch für diese Form der Geheimhaltung näher eingegangen werden soll:

Rundungsverfahren können dabei für sich den Vorteil verbuchen, dass sie – zumindest in den einfacheren Ausprägungen – leicht umsetzbar sowie relativ einfach nachvollziehbar und daher auch Außenstehenden gegenüber inhaltlich gut vermittelbar sind. Ganz besonders gilt das für die deterministische Rundung, bei der die Werte einer Tabelle in Abhängigkeit von der gewählten Rundungsbasis – beispielsweise 3, 5 oder 10 – auf- oder abgerundet werden. Die Umsetzung eines solchen Verfahrens gestaltet sich aus technischer Sicht einfach, allerdings ist das bei Verwendung einer kleinen Rundungsbasis erreichbare Schutzniveau gering. Im Gegenzug fällt bei der Wahl einer größeren Rundungsbasis zwar der Schutz der Angaben stärker aus, da sich prinzipbedingt größere Abweichungen gegenüber den Originalwerten ergeben, zugleich kommt es dadurch aber zum Auftreten größerer Verzerrungen, die den inhaltlichen Gehalt der Daten beeinträchtigen können. Besonders gilt dies natürlich für vergleichsweise niedrige Werte und Fallzahlen, wohingegen die entstehenden Abweichungen bei großen Werten oder Häufigkeiten tendenziell weniger ins Gewicht fallen. Aus diesem Grund werden in der Regel die Innenfelder einer Tabelle separat von den Randfeldern behandelt. Wäre dies nicht der Fall, so würden sich die Abweichungen gegenüber den Echtwerten aufaddieren und zu gegebenenfalls deutlich verzerrten Randsummen führen. Durch die getrennte Behandlung wird demgegenüber sichergestellt, dass alle Tabellenfelder – auch die dargestellten Summen – maximal um denselben Wert von der Originalangabe abweichen. Hieraus ergibt sich jedoch ein weiteres Problem, das die meisten Rundungsverfahren mit sich bringen: Innerhalb der Tabelle ist keine Additivität mehr gegeben. Das bedeutet, dass sich die Innenfelder nicht zwangsläufig zur Randsumme aufaddieren lassen, so wie man es normalerweise von Tabellen gewohnt ist. Die Konsistenz ist jedoch über alle Tabellen hinweg, in denen ein bearbeitetes Tabellenfeld in Erscheinung tritt, gegeben.

Alternative Varianten, wie das zufällige Runden, bei dem in Abhängigkeit von einer Zufallsentscheidung ab- oder aufgerundet wird, können das Schutzniveau erhöhen, das andere genannte Problem jedoch nicht lösen. Einen Ausweg hierfür bietet das sogenannte kontrollierte Runden, bei dem versucht wird, die durch Rundung erzeugte Tabelle im Rahmen eines aufwendigen Verfahrens so zu optimieren, dass auch weiterhin eine weitgehende Additivität gegeben ist. Gegen die Anwendung dieses Verfahrens sprechen jedoch vor allem die große Komplexität bei der programmiertechnischen Implementierung sowie die damit einhergehende hohe Rechenintensivität, die insbesondere die Bearbeitung sehr umfangreicher Tabellen einschränkt.

Darüber hinaus existieren weitere, gegenüber dem Runden komplexere Verfahren, wie die stochastische Zufallsüberlagerung – beispielsweise das vom nationalen Statistikamt in Australien eingesetzte ABS-Verfahren –, die mit einem höheren erzielbaren Sicherheitsniveau einhergehen. Alle diese post-tabularen Geheimhaltungsverfah-

Tab. 1 Beispiel für eine fiktive Häufigkeitstabelle

Bevölkerung nach Alter und Geschlecht

Alter in Jahren	Weiblich	Männlich	Insgesamt
unter 14	3	3	6
14 bis 49	8	9	17
50 bis 75	12	9	21
75 oder älter	4	1	5
Insgesamt	**27**	**22**	**49**

Tab. 2 Beispiel für die Geheimhaltung durch deterministische 3er-Rundung

Bevölkerung nach Alter und Geschlecht

Alter in Jahren	Weiblich	Männlich	Insgesamt
unter 14	3	3	6
14 bis 49	9	9	18
50 bis 75	12	9	21
75 oder älter	3	0	6
Insgesamt	**27**	**21**	**48**

ren beinhalten jedoch den prinzipiellen Nachteil, dass je nach angewandtem Verfahren erstellte Ergebnistabellen nicht additiv, nicht konsistent oder im schlechtesten Fall weder additiv noch konsistent sind. Auch für die Anonymisierung von Mikrodaten existieren datenverändernde Verfahren, wie PRAM (Hundepool et al. 2010: 69 ff.) oder SAFE (Höhne 2003), auf die hier jedoch nicht näher eingegangen werden soll.

Wichtig ist, zu beachten, dass datenverändernde Verfahren in der Regel grundsätzlich auf alle Angaben – egal ob diese als unter Geheimhaltungsaspekten kritisch einzustufen sind oder nicht – angewandt werden, was dazu führt, dass ein Großteil der dargestellten Informationen gegenüber den Originalwerten verändert sein kann. Eine gezielte, regelbasierte Prüfung auf das Vorhandensein von Aufdeckungsrisiken entfällt dabei. Je nach Verfahren können aber, wie in Tabelle 2 zu sehen ist, durchaus auch gegenüber der Originaltabelle (Tabelle 1) unveränderte Zellen in den Veröffentlichungen vorhanden sein. Aus Sicht des Betrachters ist jedoch die Unsicherheit darüber, ob die dargestellte Angabe tatsächlich der Wirklichkeit entspricht oder doch um einen unbekannten Wert davon abweicht, jederzeit gegeben, und stellt somit den wesentlichen Schutzmechanismus dar.

Informationsverlust und Datenqualität

Zugleich steht und fällt die Sinnhaftigkeit eines jeden Verfahrens mit den Einschränkungen, die dieses für den Informationsgehalt und somit die Nutzbarkeit der Daten mit sich bringt. Die Anwendung von Geheimhaltungsverfahren wirkt sich zwangsläufig immer auf die Qualität und das Analysepotential der betroffenen Statistikdaten aus: Durch gezielte Eingriffe in die Daten – sei es nun durch informationsreduzierende oder datenverändernde Methoden – werden bewusst Veränderungen gegenüber den Originaldaten herbeigeführt, um die Anonymität der gemachten Einzelangaben zu gewährleisten. Dieser Eingriff in die Daten ist der zentrale Wirkmechanismus, mit dessen Hilfe das Ziel der statistischen Geheimhaltung erreicht wird. Jede Abweichung gegenüber den Originaldaten stellt jedoch zugleich unweigerlich einen Verlust an Datenqualität dar. Ein wichtiges Ziel muss es demgemäß bei der Anwendung von Geheimhaltungsverfahren sein, die Vertraulichkeit der Angaben einzelner Erhebungseinheiten zu garantieren, ohne dass hierfür ein unverhältnismäßig großer Informationsverlust in Kauf genommen werden muss. In der Vergangenheit spielten Aspekte der Datenqualität und der Nutzbarkeit der Daten bei der Auseinandersetzung mit dem Thema statistische Geheimhaltung jedoch oftmals nur eine nachgeordnete Rolle. In jüngerer Zeit hat diesbezüglich ein Umdenken stattgefunden, das unter anderem zur Beschäftigung mit geeigneten Maßen und Kennzahlen für den Informationsverlust, den ein bestimmtes Anonymisierungs- oder Geheimhaltungsverfahren mit sich bringt, geführt hat. Was die Erfassung des Informationsverlusts angeht, so werden diverse Möglichkeiten diskutiert, wie dieser konkret gemessen werden kann (u. a. Hundepool et al. 2010: S. 96 ff.; Rosemann 2007). Diese reichen von der Erfassung des Anteiles und der Stärke der vorgenommenen Veränderungen oder der Unterschiede, die sich bei der Durchführung bestimmter statistischer Analysen ergeben, bis hin zur Berechnung von informationstheoretischen Entropiemaßen, wie beispielsweise der Hellinger Distanz, oder der Möglichkeit eines Echtzeitvergleichs von anonymisierten und

nicht-anonymisierten Auswertungsergebnissen innerhalb eines Remote-Access-Systems (Höninger 2012).

Bei all dem steht außer Frage, dass die Einhaltung der gesetzlichen Verpflichtung zum Schutz vertraulicher Angaben immer über die oberste Priorität verfügen muss. Aber genauso offensichtlich ist, dass mit maximal geschützten Daten, die nur noch einen minimalen Informationsgehalt aufweisen, nicht mehr viel anzufangen ist, weshalb es unerlässlich ist, nach einer möglichst optimalen Balance zwischen diesen beiden Anforderungen zu suchen.

Aktuelle Entwicklungen und kommende Herausforderungen

Aktuelle und sich abzeichnende zukünftige Entwicklungen sorgen dafür, dass es sich bei der statistischen Geheimhaltung nicht um einen statischen Themenkomplex mit einem fixen Instrumentarium an anzuwendenden Verfahren handelt, sondern es ist eine permanente Überprüfung und Anpassung der genutzten Methoden notwendig. Vor allem neue Veröffentlichungsformen sind es, die eine Herausforderung für die Sicherstellung der statistischen Geheimhaltung mit sich bringen können, schließlich sollen die sich bietenden Potentiale genutzt werden können, ohne dabei jedoch den Schutz vertraulicher Daten zu vernachlässigen. Bei all diesen Betrachtungen sollte beachtet werden, dass die grundlegenden gesetzlichen Regelungen zur statistischen Geheimhaltung noch aus einer Zeit vor dem beginnenden Siegeszug der modernen Computertechnik stammen. Personal Computer fristeten damals noch ein Nischendasein; Laptops, Tablets oder Smartphones existierten höchstens in Science-Fiction-Filmen und auch das Internet existierte nur in Form seiner frühesten Vorläufer. An die Art und Weise, auf die IT innerhalb weniger Jahre Wirtschaft und Gesellschaft durchdringen würde, war zum damaligen Zeitpunkt noch nicht zu denken. Dementsprechend treffen die Bestimmungen aus dem von 1987 stammenden Bundesstatistikgesetz auf eine gegenüber dem Entstehungszeitpunkt stark veränderte Rahmensituation, wobei die damals festgelegten Grundsätze und Ziele jedoch nichts von ihrer Bedeutung verloren haben – sondern eher im Gegenteil angesichts der Möglichkeiten der modernen Datenverarbeitung[2] noch zusätzlich an Gewicht gewonnen haben. Als Beispiele für Entwicklungen aus neuerer Zeit soll im Folgenden exemplarisch auf die Arbeit mit georeferenzierten Statistikdaten sowie den Zugang der empirischen Forschung zu statistischen Mikrodaten eingegangen werden.

Geheimhaltung georeferenzierter Daten

Zu einer der mithin interessantesten Entwicklungen zählt sicherlich die Veröffentlichung von mit Geokoordinaten angereicherten Daten durch die Statistischen Ämter. Durch die Verbindung von amtlichen Statistikdaten mit konkreten Raumbezügen eröffnen sich neue Möglichkeiten für regionalisierte Analysen, die nicht mehr an die bislang verfügbaren Verwaltungsgliederungen (Gemeinden, Kreise etc.) gebunden sind, und neue Arten der Visualisierung statistischer Ergebnisse in Form von (interaktiven) Kartendarstellungen möglich machen. Die Statis-tischen Ämter selbst, die empirisch arbeitenden Raumwissenschaften – vor allem die Geographie, aber auch die Wirtschafts- und Sozialwissenschaften – sowie die neue Form des datenbasierten Journalismus werden zukünftig in voraussichtlich immer stärkerem Maße auf die Nutzung und Präsentation von georeferenzierten Informationen zurückgreifen. Aus diesen neuen Darstellungsoptionen ergeben sich jedoch zugleich neue Anforderungen an die Praxis der statistischen Geheimhaltung; liefert die Möglichkeit einer kleinräumigen geografischen Zuordnung doch zusätzliche, möglicherweise individuierende Informationen, die von einem potentiellen Datenangreifer gewinnbringend bei der Reidentifikation einzelner Merkmalsträger eingesetzt werden könnten. Die Sicherstellung des Schutzes der Daten, die oftmals zusätzlich parallel in traditioneller

2 Bereits in der Begründung des Volkszählungsurteils von 1983 wurde dies als ein maßgeblicher Grund für die Bedeutung des Datenschutzes benannt, wobei man aus heutiger Sicht unter technischen Gesichtspunkten damals in Bezug auf IT und elektronische Datenverarbeitung noch relativ am Anfang der Entwicklung stand.

Tabellenform verfügbar sind, ohne dass die neuen Möglichkeiten des georeferenzierten Arbeitens zugleich wieder beschnitten werden, stellt dabei das zentrale Ziel dar. Vor allem das Nebeneinander unterschiedlich gegliederter Darstellungen muss berücksichtigt werden, da es hier durch Differenzbildungen zur Entstehung von Aufdeckungsrisiken kommen könnte. Um dem entgegenzuwirken, existiert eine Reihe von Geheimhaltungsmöglichkeiten (Höhne/Höninger 2014): Dabei bietet sich insbesondere die Vergröberung des verwendeten Rasters bei kartographischen Darstellungen an – eine Maßnahme, die auf Tabellen bezogen einer Umgestaltung durch die Vergröberung von Kategorien entsprechen würde. Statt auf Gitterzellen mit einer Größe von 1 km x 1 km beziehen sich die Zuordnungen dann beispielsweise auf Bereiche mit einer Größe von 5 km x 5 km. Auch variable Rastergrößen innerhalb ein- und derselben Abbildung sind aus Geheimhaltungssicht denkbar. Aber auch die Möglichkeit, analog zur Zellsperrung in Tabellen kritische Rasterzellen schlichtweg zu unterdrücken und die Sperrung durch die Unterdrückung weiterer Zellen sekundär abzusichern, stellt eine gangbare Alternative dar.

Zugang der Wissenschaft zu Mikrodaten der amtlichen Statistik

Eine weitere nachhaltig wirksame Entwicklung stellt die stattgefundene Öffnung der amtlichen Statis-tik für die Belange der empirisch orientierten, wissenschaftlichen Forschung dar. Das in § 16 Abs. 6 des Bundesstatistikgesetzes (BStatG) festgeschriebene „Wissenschaftsprivileg“ eröffnet den Statistischen Ämtern die Möglichkeit, Angehörigen von Hochschulen und anderen unabhängigen wissenschaftlichen Forschungseinrichtungen faktisch ano-nymisierte Einzelangaben für die Durchführung von zeitlich begrenzten Forschungsprojekten bereitzustellen. Besonders die modernen Sozial- und Wirtschaftswissenschaften sind für ihre Analysen in hohem Maße auf hochwertige Sekundärdaten mit möglichst vielen Detailinformationen angewiesen. Eine Vielzahl von Forschungsfragen lässt sich überhaupt nur durch den Rückgriff auf Mikrodaten anhand statistischer Analyseverfahren adäquat untersuchen und sinnvoll beantworten. Von Seiten der Statistischen Ämter wurde mit den Forschungsdatenzentren des Bundes und der Länder daher eine deutschlandweite Infrastruktur geschaffen, die der empirisch arbeitenden Wissenschaft einen Zugang zu mittlerweile rund 120 unterschiedlichen Erhebungen aus allen Bereichen der amtlichen Statistik ermöglicht. Innerhalb von nur wenig mehr als zehn Jahren gelang es hierdurch, das zuvor vorhandene Defizit nachhaltig zu beheben.[3]

Allerdings ist die internationale Entwicklung im Bereich Forschungsdaten schon wieder einen Schritt weiter: Remote Access lautet das Schlagwort, hinter dem sich aus Sicht vieler Wissenschaftler die Erfüllung langgehegter Wünsche in Sachen Datenzugang verbirgt. Die Möglichkeit, per Fernzugriff – idealerweise vom eigenen Arbeitsplatz aus – auf Mikrodaten der amtlichen Statistik zuzugreifen und Analysen durchzuführen, stellt sich aus Nutzersicht verlockend dar (Desai 2003). Für die Datennutzer entfallen zeit- und möglicherweise auch kostenintensive Gastaufenthalte in den Statistischen Ämtern; zugleich müssen keine derartigen Einschränkungen der Datenqualität und des Analysepotentials in Kauf genommen werden, wie dies bei Verwendung der für die externe Nutzung gedachten Scientific-Use-Files oft der Fall ist. Außerdem würden auch die Mitarbeiter der Statistischen Ämter vor Ort potenziell entlastet. Demgegenüber steht auf Seiten der Statistischen Ämter jedoch das Problem, im Einklang mit den einschlägigen gesetzlichen Regelungen die Wahrung der statistischen Geheimhaltung sicherzustellen. Werden Daten innerhalb eines statistischen Amtes zur Verfügung gestellt, so erfolgt der Zugang unter unmittelbar kontrollierbaren technischen und organisatorischen Rahmenbedingungen. Bei einer Datennutzung per Remote Access ist dies nicht oder nur eingeschränkt gegeben. Dabei ist es in erster Linie der direkte Blick auf die Mikrodaten, der sowohl das spontane Erkennen von Merkmalsträgern als auch die gezielte Suche

3 Diese Entwicklung betraf nicht nur die Statistischen Ämter des Bundes und der Länder, sondern auch die großen öffentlichen Datenproduzenten als Ganzes (Bender 2014).

nach solchen ermöglicht, der ein Risiko aus Geheimhaltungssicht darstellt, wenn es im Gegensatz zu den abgeschotteten Gastarbeitsplätzen in den Statistischen Ämtern nicht möglich ist, zu verhindern, dass möglicherweise Zusatzinformationen, die der Reidentifizierung dienen könnten, genutzt werden. Darüber hinaus ist es – vergleichbar zum Datenzugang innerhalb der Ämter – die Erstellung deskriptiver Ergebnistabellen und die Erzeugung von Grafiken, die ein Risiko beinhalten kann. Die Ergebnisse statistischer Auswertungsverfahren, beispielsweise von Regressionsanalysen, sind hingegen aus Geheimhaltungssicht weitgehend unbedenklich, auch wenn sich unter spezifischen Bedingungen ebenfalls Enthüllungsrisiken ergeben können (Hochgürtel 2013; Ronning et al. 2011; Vogel 2011).

Die technische Umsetzbarkeit einer solchen Remote-Access-Lösung wurde im europäischen Kontext im Rahmen des ESSNet-Projekts „Decentralised And Remote Access to Confidential Microdata in the ESS (DARA)“ (Essnet DARA 2014) erfolgreich getestet. Eine Realisierung des dabei erprobten Sys-tems innerhalb des Europäischen Statistischen Sys-tems ist angedacht. Manche Statistikämter anderer Staaten bieten ihren Datennutzern auch bereits heute die Möglichkeit, zu festgelegten – teilweise recht restriktiven und relativ kostspieligen – Bedingungen per Fernzugriff mit amtlichen Daten zu arbeiten oder sie arbeiten daran, eine solche Nutzungsoption zu implementieren (Le Gléau/Royer 2011; Schulte-Nordholt 2013). Allerdings sind die dabei geltenden gesetzlichen Rahmenbedingungen in der Regel nicht oder nur eingeschränkt mit den entsprechenden Regelungen in Deutschland vergleichbar.

In jüngerer Zeit werden zudem zunehmend organisatorische Möglichkeiten in Form von Akkreditierungs-, Lizenzierungs- und Zertifizierungsverfahren für Wissenschaftler und Forschungseinrichtungen (Rendtel 2014, Tubaro et al. 2011) oder das Bilden eines sogenannten „Circle of Trust“ (OECD 2014) zwischen Datenproduzenten und Datennutzern, basierend auf vertraglichen Regelungen und vertrauensbildenden Maßnahmen, als Ergänzung oder teilweisen Ersatz der bisherigen Geheimhaltungspraxis diskutiert, um damit eine Erleichterung – insbesondere aber nicht ausschließlich, für den transnationalen Forschungsdatenzugang – zu erreichen.

Fazit

Das Ziel der statistischen Geheimhaltung ist der grundgesetzlich verbriefte Schutz vertraulicher Daten. Die amtliche Statistik hat Sorge dafür zu tragen, dass dieser Verpflichtung Genüge getan wird, und entsprechende Risiken, die aus ihren Veröffentlichungen resultieren könnten, auszuschließen. Um dies zu erreichen, ist es notwendig zu beachten, dass es sich bei statistischer Geheimhaltung um ein vielschichtiges Thema handelt, welches sich nicht alleine auf einzelne Teilaspekte rechtlicher, methodischer, technischer und organisatorischer Art beschränken lässt. Es ist notwendig, diese Aspekte im Zusammenspiel zu betrachten, um zu praxisgerechten Lösungen zu gelangen und das angestrebte Ziel, sichere Daten in der bestmöglichen Qualität veröffentlichen zu können, zu erreichen.

Im Rahmen des Beitrags wurde versucht, eine Einführung in die rechtlichen und methodischen Grundlagen der statistischen Geheimhaltung zu geben und dabei die wichtigsten Verfahren zur Geheimhaltung von Häufigkeits- und Wertetabellen vorzustellen. Natürlich ist es nicht möglich, dabei mehr als einen knappen Überblick über die vielschichtige Thematik zu geben. Interessierte, die sich intensiver mit dem Thema an sich oder aber mit einzelnen Verfahren oder Teilaspekten beschäftigen möchten, finden eine Vielzahl an Informationen beispielsweise in Höhne 2010, Hundepool et al. 2010, Hundepool/De Wolf 2012 oder Ronning et al. 2005. Darüber hinaus ist zu Einzelaspekten eine Vielzahl von zumeist englischsprachigen Fachartikeln und Arbeitspapieren von Experten aus amtlicher Statistik und universitärer Forschung verfügbar.

Literatur

Bender, S. (2014), Datenzugang in Deutschland: Der Paradigmenwechsel hat bereits stattgefunden. AStA – Wirtschafts- und Sozialstatistisches Archiv, Vol. 8 (4), S. 237-248.

Desai, T. (2003), Proving Remote Access to Data: The Academic Perspective. In: Monographs of Official Statistics 1. Work session on statistical data confidentiality. Luxembourg, 7 to 9 April 2003. Part 1. Luxembourg: Office for Official Publications of the European Communities, S. 151–159.

Essnet DARA (2014). Final Report. Downloadunter: www.safe-centre.info/wp-content/uploads/2012/01/final_report_ESSnet_DARA_20131204_publishable_version.pdf, abgerufen am 30. Juni 2015.

Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz BStatG) vom 22. Januar 1987. (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 2 des Gesetzes vom 9. Juni 2005 (BGBl. I S. 1534).

Hochgürtel, T. (2013), Die Messung der Enthüllungsrisiken von Ergebnissen statistischer Analysen. Arbeitspapier Nr. 3. Institut für Diskrete Mathematik und Angewandte Statistik der Hochschule für Technik und Wirtschaft des Saarlandes.

Höhne, J. (2003), SAFE – ein Verfahren zur Geheimhaltung und Anonymisierung statistischer Einzelangaben. Berliner Statistik. Monatsschrift, 03/2003, S. 96–107.

Höhne, J. (2010), Verfahren zur Anonymisierung von Einzeldaten. Statistik und Wissenschaft, Band 16. Wiesbaden: Statistisches Bundesamt.

Höhne, J./Höninger, J. (2014), Statistische Geheimhaltung bei der Auswertung georeferenzierter Daten. Zeitschrift für amtliche Statistik Berlin-Brandenburg 03/2014, S. 54–61.

Höninger, J. (2012), Morpheus – An innovative approach to remote data access. Journal of the International Association for Official Statistics, Vol. 28 (3/4), S. 151–157.

Hundepool, A./Domingo-Ferrer, J./Franconi, L./Gies-sing, S./Lenz, R./Naylor, J./Schulte-Nordholt, E./Seri, G./De Wolf, P. (2010), Handbook on Statistical Confidentiality. Version 1.2. Download unter: http://neon.vb.cbs.nl/casc/.%5CSDC_Handbook.pdf, abgerufen am 30. Juni 2015.

Hundepool, A./De Wolf, P.-P. (2012), Statistical disclosure control. Method series 12. Den Hague/Heerlen: Statistics Netherlands.

Le Gléau, J.-P./Royer, J.-F. (2011), Le centre d'accès sécurisé aux données de la statistique publique française: un nouvel outil pour les chercheurs. Courrier des statistiques, n° 130, May 2011, INSEE. Download unter: www.insee.fr/fr/ffc/docs_ffc/cs130e.pdf, abgerufen am 30. Juni 2015.

OECD (2014), OECD Expert Group for International Cooperation on Microdata Access. Final Report. Download unter: www.oecd.org/std/microdata-access-final-report-OECD-2014.pdf, abgerufen am 30. Juni 2015.

Rendtel, U. (2014), Vom potenziellen Datenangreifer zum zertifizierten Wissenschaftler – Für eine Neugestaltung des Wissenschaftsprivilegs beim Datenzugang. AStA Wirtschafts- und sozialstatistisches Archiv, Vol 8. (4), S. 183–197.

Ronning, G./Sturm, R./Höhne, J./Lenz, R./Rosemann, M./Scheffler, M./Vorgrimler, D. (2005), Handbuch zur Anonymisierung wirtschaftsstatis-tischer Mikrodaten. Statistik und Wissenschaft. Band 4. Wiesbaden: Statistisches Bundesamt.

Ronning, G./Bleninger, P./ Drechsler, J./Gürke, C. (2011), Remote Access. Eine Welt ohne Mikrodaten?? FDZ-Arbeitspapier Nr. 33. Download unter: www.forschungsdatenzentrum.de/publikationen/veroeffentlichungen/33.asp, abgerufen am 30. Juni 2015.

Rosemann, M. (2007), Auswirkungen von stochas-tischer Überlagerung und Mikroaggregation auf die Schätzung linearer und nichtlinearer Modelle. Wirtschaft und Statistik, 04/2007, S. 417–432.

Schulte-Nordholt (2013), Access to microdata in the Netherlands: from a cold war to cooperation projects. Workingpaper zur Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa, 28.–30. Oktober 2013. Download unter: www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_3_Schulte_Nordholt.pdf, abgerufen am 30. Juni 2015.

Tubaro, P./Cros, M./Silberman, R. (2012), Access to Official Data and Researcher Accreditation in Europe: existing Barriers and a Way forward. IASSIST Quarterly, Spring 2012, S. 22–27.

Vogel, A. (2011), Enthüllungsrisiko beim Remote Access: Die Schwerpunkteigenschaft der Regressionsgerade, Working Paper Reihe des Rates für Sozial- und Wirtschaftsdaten, Nr. 174. Download unter: www.ratswd.de/download/RatSWD_WP_2011/RatSWD_WP_174.pdf, abgerufen am 30. Juni 2015.

Statistische Ämter des Bundes und der Länder,
FDZ-Arbeitspapier Nr. 50; Statistische Geheimhaltung – Der Schutz vertraulicher Daten in der amtlichen Statistik