Statistische Geheimhaltung – Der Schutz vertraulicher Daten in der amtlichen Statistik

Teil 1: Rechtliche und methodische Grundlagen

Dipl.-Soz. 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]

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 bei-

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.

spielsweise 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,

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).

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 Kombination 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 (2^{10}) 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 Merk-

mal 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 ($10^3 = 1000$) 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 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 Viel-

4 Dies gilt analog für Betriebe und Unternehmen, die anhand von brancheninternem 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.

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.

Übersicht: **Die unterschiedlichen Stufen der Anonymität und deren Zielgruppen**

Grad der Anonymität	Zielgruppe	Mögliche Produktform, z. B.	Informationsgehalt
absolut	- breite Öffentlichkeit	- Statistisches Jahrbuch - Statistische Berichte - GENESIS Online - Statistikatlas - Public-Use-Files	niedrig bis mittel
faktisch	- unabhängige wissenschaftliche Forschung	- Forschungsdatenzentren-Gastwissenschaftlerarbeitsplatz - Scientific-Use-Files	mittel bis hoch
formal	- Angehörige der Statistischen Ämter - Empfangsberechtigte nach Ausnahmeregelungen des BStatG	- nur zur Verarbeitung innerhalb der Statistischen Ämter	maximal

zahl 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 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

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).

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 inner-

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.

halb 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 Spaltensummen 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 kon-

sistent 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):

Tab. 5 **Beispiel für die Berücksichtigung des Randwertkriteriums**
Prüfungserfolg nach Geschlecht

	Weiblich	Männlich	Insgesamt
Bestanden	3	0	3
Nicht bestanden	7	10	17
Insgesamt	10	10	20

Tab. 6 **Beispiel für Sperrungen bei Berücksichtigung des Randwertkriteriums**
Prüfungserfolg nach Geschlecht

	Weiblich	Männlich	Insgesamt
Bestanden	•	•	3
Nicht bestanden	•	•	17
Insgesamt	10	10	20

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 Rah-

menbedingungen 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 For-

schungsdateninfrastruktur 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. I S. 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.