Statistische Geheimhaltung – Der Schutz vertraulicher Daten in der amtlichen Statistik

Teil 2: Herausforderungen und aktuelle Entwicklungen

Dipl.-Soz. Patrick Rothe

Im in Ausgabe 05/2015 der Monatszeitschrift „Bayern in Zahlen“ enthaltenen 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

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

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 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.

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**

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 Geheimhaltungsverfahren 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 grundle-

genden 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 Statistischen 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 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 Statistik 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 anonymisierte 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.

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.

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 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 Systems innerhalb des Europäischen Statistischen Systems 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.

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

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. Download-unter: 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./Giessing, 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 wirtschaftsstatistischer 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 stochastischer 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.