Elena Schweitzer / stock.adobe.com

STATUS QUO UND AKTUELLE ENTWICKLUNGEN IN DER STATISTISCHEN GEHEIMHALTUNG

Dipl.Math.oec.Univ. Andreas Nickl und Dipl.Soz.Wiss.Univ. Patrick Rothe

Die statistische Geheimhaltung ist ein zentraler Bestandteil der amtlichen Statistik, mit dem Ziel, den Schutz persönlicher Daten zu gewährleisten und gleichzeitig qualitativ hochwertige Ergebnisse zu liefern. Mit der zunehmenden Digitalisierung sowie der wachsenden Komplexität und Granularität von Daten und deren Darstellungsformen stellt die Geheimhaltung jedoch zugleich eine immer größere Herausforderung methodischer Art dar. Dieser Artikel bietet einen Überblick über die rechtlichen Grundlagen, die aktuell in der deutschen amtlichen Statistik angewandten Geheimhaltungsmethoden und die Rolle neuartiger Ansätze, wie beispielsweise der Verwendung synthetischer Daten, als innovative Lösungen für den Schutz von persönlichen Daten und Betriebsgeheimnissen.

Rechtliche Grundlagen der statistischen Geheimhaltung – Das Statistikgeheimnis und seine Umsetzung von den 80er-Jahren bis heute

Das Recht auf informationelle Selbstbestimmung, erstmals im sogenannten Volkszählungsurteil im Jahr 1983 vom Bundesverfassungsgericht festgehalten, bildet die Grundlage für den Schutz persönlicher Daten in Deutschland. Es leitet sich aus Artikel 2 des Grundgesetzes ab und garantiert jeder Bürgerin und jedem Bürger den Schutz seiner persönlichen Daten. Den Schutz persönlicher Informationen im Allgemeinen regelt dabei – im Zusammenspiel mit der Datenschutzgrundverordnung – das Bundesdatenschutzgesetz. Für die Belange der amtlichen Statistik in Deutschland stellt hierzu jedoch das Bundestatistikgesetz (BStatG) die einschlägige Rechtsgrundlage im Sinne einer „lex specialis" dar. Dies bedeutet, dass die spezialgesetzliche Regelung Vorrang vor den allgemeineren Gesetzen und Verordnungen hat und diese bereichsspezifisch – hier bezogen auf die Arbeit der Statistischen Ämter – umsetzt. Korrespondierende Regelungen für den jeweiligen landesrechtlichen Bereich finden sich in weitgehend identischer Form in den entsprechenden Landesstatistikgesetzen.

Das sogenannte Statistikgeheimnis (§ 16 Abs. 1 Satz 1 BStatG) verpflichtet Amtsträger und andere Verantwortliche dazu, erhobene Einzelangaben über persönliche und sachliche Verhältnisse geheim zu halten. Diese Geheimhaltung ist essenziell dafür, das Vertrauen der Erhebungspflichtigen dauerhaft zu erhalten und hierdurch sowohl die Zuverlässigkeit der gewonnenen Daten sowie die Teilnahmebereitschaft im Falle freiwilliger Erhebungen sicherzustellen. Ausnahmen von der Geheimhaltungspflicht sind abweichend hiervon nur unter bestimmten Bedingungen erlaubt, etwa beim Vorliegen absolut anonymer Einzeldaten oder mit informierter, schriftlicher Einwilligung der Betroffenen. Einzelnen Gruppen von Datenempfängern beziehungsweise Datennutzern räumt der Gesetzgeber jedoch besondere Privilegien hinsichtlich Art und Form des Datenzugangs ein, so beispielsweise der empirisch forschenden Wissenschaft im Rahmen von § 16 Abs. 6 BStatG mit dem sogenannten Wissenschaftsprivileg. Dieses erlaubt es den Statistischen Ämtern über die Forschungsdatenzentren detaillierte Mikrodaten für wissenschaftliche Zwecke nach einem abgestuften Verfahren zur Verfügung zu stellen (Hlawatsch / Meyer / Rothe 2022). Weitere Sonderregelungen existieren unter anderem für die Bereitstellung feingliedriger Tabellen für Planungszwecke an Landes- und Bundesministerien (§ 16 Abs. 4 BStatG) sowie für statistische Zwecke der Kommunen (§ 16 Abs. 5 BStatG) oder aber auch für den erforderlichen Datenaustausch zwischen den Statistischen Landesämtern untereinander sowie mit dem Statistischen Bundesamt zur Erstellung von Bundesstatistiken (§ 16 Abs. 2 und 3 BStatG). Die jeweilige Erlaubnis der Datenweitergabe wird dabei von Regelungen zur Sicherstellung des verantwortungsvollen Umgangs mit den Daten flankiert. So erfolgt diese immer zweckgebunden, die Daten müssen sobald wie nach dem jeweiligen Verwendungszweck möglich gelöscht werden und externe Datenempfänger müssen zur Wahrung der statistischen Geheimhaltung verpflichtet werden.

Andreas Nickl, Dipl.Math.oec.Univ.

Andreas Nickl studierte Diplom-Wirtschaftsmathematik an der Friedrich-Alexander-Universität Erlangen-Nürnberg. Seit 2012 arbeitet er im Bayerischen Landesamt für Statistik, zunächst als Referent im Sachgebiet „Zensus" in der Dienststelle München. Seit 2015 ist er als Referent in der Dienststelle Fürth tätig und verantwortet aktuell als stellvertretender Leiter des Sachgebiets „Statistische Methoden, Digitalisierung und Forschungsdatenzentrum" unter anderem die Bereiche Statistische Geheimhaltung sowie mathematisch-statistische Verfahren. Seit 2022 leitet er das Team „Statistische Methodik und Digitalisierung".

Bildnachweis: Bernhard Lidschreiber – IG Fotografie des SKV

Patrick Rothe, Dipl.Soz.Wiss.Univ.

Patrick Rothe studierte Sozialwissenschaften an der Universität Mannheim. Nach Beendigung seines Studiums arbeitete er als akademischer Mitarbeiter für das Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg, bevor er 2011 ins Bayerische Landesamt für Statistik wechselte. Hier war er – erst in der Dienststelle München, seit 2015 in der Dienststelle Fürth – als Referent unter anderem im Bayerischen Standort des Forschungsdatenzentrums der Statistischen Landesämter sowie im Bereich Methodik tätig. 2018 übernahm er die Leitung des Sachgebiets „Statistische Methoden, Digitalisierung und Forschungsdatenzentrum", wo neben klassischen statistischen Verfahren und neuen digitalen (KI-)Methoden auch das Thema statistische Geheimhaltung und Anonymisierung von (Mikro-)Daten angesiedelt ist.

Bildnachweis: LfStat

Da die Kernbestandteile des BStatG ihren Ursprung Mitte der 80er-Jahre des vergangenen Jahrhunderts haben, ergibt sich zwangsläufig, dass diese mit den seit damals – insbesondere in den letzten Jahren nochmals beschleunigt – stattgefundenen Veränderungen im Bereich der Datenverarbeitung nicht zwingend in allen Bereichen Schritt halten konnten. Zugleich ist das BStatG in jüngerer Zeit unter rechtswissenschaftlichen Gesichtspunkten wieder auf vermehrtes Interesse gestoßen, welches sich unter anderem durch die Veröffentlichung des ersten neuen juristischen Kommentars zum BStatG (Kühling 2023) seit dem Erscheinen des ersten Gesetzeskommentars (Dorer / Mainusch / Tubies 1988) vor mehr als 30 Jahren, aber auch durch verschiedene Veranstaltungen der amtlichen Statistik gemeinsam mit der Wissenschaft als auch Vertretern der Landes- und Bundesverwaltung zur Fortentwicklung des Bundesstatistikrechts[1,2], manifestiert hat.

1 Symposium „Zukunft der amtlichen Statistik – Perspektiven des Bundesstatistikgesetzes" am 25. Januar 2024 in Erfurt, welches vom Thüringer Landesamt für Statistik und dem Bayerischen Landesamt für Statistik ausgerichtet wurde.

2 Symposium zum Bundesstatistikgesetz mit Schwerpunkt Datenschutz und Forschungsdatenzugang, welches am 28. Oktober 2024 vom Statistischen Bundesamt, dem Thüringer Landesamt für Statistik und dem Bayerischen Landesamt für Statistik in Berlin veranstaltet wurde.

Insbesondere die Befassung mit Fragen der rechtlichen Weiterentwicklung hat durch das Inkrafttreten der Überarbeitung der EU-Statistikverordnung im Jahr 2024 oder auch die Bemühungen sowohl der vorhergehenden als auch der jetzigen Bundesregierung um ein Forschungsdatengesetz an Fahrt gewonnen. Nicht zuletzt haben auch die Statistischen Ämter selbst Handlungsfelder identifiziert, in denen ein Bedarf für Anpassungen der bestehenden Rechtslage offenbar wurde, um auch zukünftig sowohl den wachsenden Möglichkeiten im Bereich der zeitnahen Veröffentlichung hochwertiger Daten, aber auch den neuen Herausforderungen an den Schutz vertraulicher Einzelangaben in einer gewandelten gesellschaftlichen und technologischen Umwelt gleichermaßen gerecht werden zu können.

Neben der Klärung offener Rechtsfragen im Rahmen der bereits bestehenden Möglichkeiten und der Änderung bestehender Vorgaben beziehungsweise Schaffung neuer Rechtsgrundlagen durch den Gesetzgeber stellt hier auch die Verpflichtung der amtlichen Statistik zur stetigen Weiterentwicklung der eingesetzten Methodik – gerade auch mit Blick auf europäische und internationale Entwicklungen – ein zentrales Themenfeld dar.

Verfahren der statistischen Geheimhaltung – der methodische Werkzeugkasten

Die amtliche Statistik nutzt eine Vielzahl von Methoden und Verfahren, um Daten zu anonymisieren und Veröffentlichungen vor unerlaubten Rückschlüssen auf die dahinterstehenden statistischen Erhebungseinheiten – beispielsweise Personen, Unternehmen oder Betriebe – zu schützen (u. a. Center of Excellence SDC 2024, Rothe 2015a, Rothe 2015b). Beabsichtigtes Ziel ist es dabei immer zu verhindern, dass Außenstehende korrekte Rückschlüsse auf die ursprünglich von einzelnen Personen oder sonstigen statistischen Einheiten zu einer Veröffentlichung beigetragenen Angaben ziehen können. Hierbei wird versucht zu unterbinden, dass es zur Reidentifizierung von Merkmalsträgern, sei es durch direkte oder durch indirekte Herangehensweisen, kommen kann. Zudem können auch die von den jeweiligen Merkmalsträgern stammenden Angaben im Rahmen der statistischen Geheimhaltung verändert werden, um eine Aufdeckung der exakten oder ungefähren Angabe zu vermeiden. Die hierfür genutzten Verfahren, die auf unterschiedlichen Wirkmechanismen basieren, lassen sich in zwei Hauptkategorien unterteilen:

Pre-tabulare Methoden (Anonymisierung)

Diese Methoden kommen bereits auf der Ebene der originalen Einzeldaten zum Einsatz und entfernen in einem ersten Schritt direkte Identifikatoren wie Name, Adresse oder Matrikelnummer. In einem Folgeschritt werden bei Bedarf gegebenenfalls weitergehende Eingriffe in das Datenmaterial vorgenommen. Es werden dabei verschiedene Stufen der Anonymität unterschieden:

- **Formale Anonymisierung:** Entfernung aller direkten Identifikatoren, sodass ohne weitergehendes Vorwissen zu den in der Datenbasis enthaltenen Einheiten keine Rückschlüsse auf diese mehr möglich sind.

- **Faktische Anonymisierung:** Bearbeitung der Daten, sodass unter als realistisch eingestuften Bedingungen keine Identifikation möglich ist. Dies bedeutet, dass eine korrekte Zuordnung einer Angabe zu einer dahinterstehenden statistischen Einheit zwar nicht mit an absoluter Sicherheit grenzender Wahrscheinlichkeit ausgeschlossen werden kann, die Wahrscheinlichkeit einer Aufdeckung jedoch extrem gering ausfällt. Der Gesetzgeber geht hierbei davon aus, dass ein sehr hoch ausfallender Aufwand, der für eine Aufdeckung mit im Verhältnis dazu geringem Nutzen betrieben werden müsste, einen potenziellen Datenangreifer von entsprechenden Versuchen abhält, sofern dieser in seinem Vorgehen rationalen Überlegungen unterliegt.

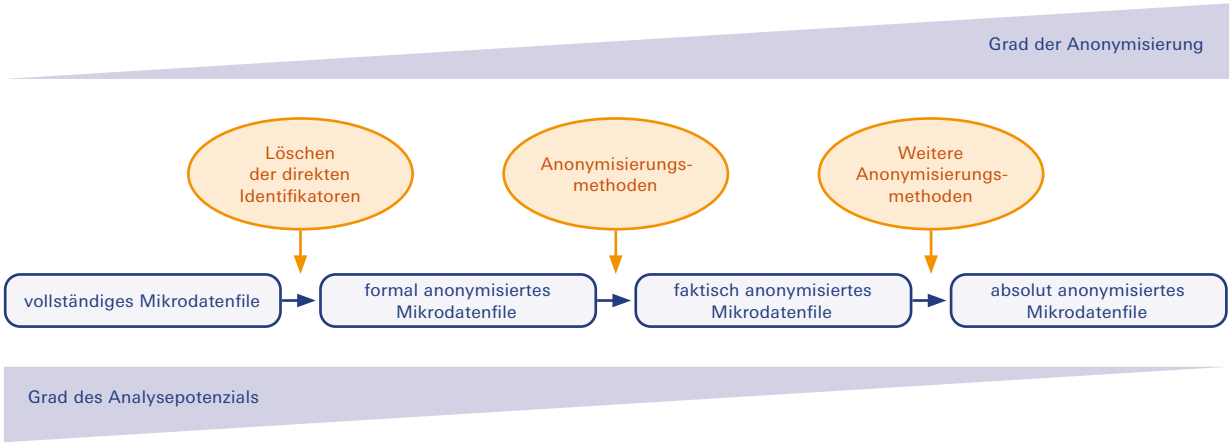
- **Absolute Anonymisierung:** Bei dieser Methode werden Daten so bearbeitet, dass keinerlei Reidentifikation von Angaben einzelner Beitragender mehr möglich ist. Der Nachweis hierfür gestaltet sich jedoch schwierig, was in der Praxis dazu führen kann, dass zu weitreichende Maßnahmen angewandt werden, um die Reidentifizierbarkeit einzelner Einheiten auch tatsächlich sicherzustellen. Rechtlich wird dem Umstand der schwierigen Fassbarkeit Rechnung getragen, in dem hier von einer gesteigerten faktischen Anonymität gesprochen wird, die es zu erreichen gilt.

> *Das sogenannte Statistikgeheimnis (§ 16 Abs. 1 Satz 1 BStatG) verpflichtet Amtsträger und andere Verantwortliche dazu, erhobene Einzelangaben über persönliche und sachliche Verhältnisse geheim zu halten.*

Sofern Daten als absolut anonym eingestuft sind, können sie einer unbegrenzten Öffentlichkeit bereitgestellt werden. Eine Vielzahl der von den Statistischen Ämtern veröffentlichten Inhalte, beispielsweise Ergebnistabellen in Statistischen Berichten oder dem Statistischen Jahrbuch, entsprechen diesem Anonymisierungmaßstab. Auch Public-Use-Files der Forschungsdatenzentren sowie Open-Data-Angebote, die sich an die allgemeine Öffentlichkeit richten, fallen hierunter.

Um für einen Datenbestand den gewünschten Grad an Anonymität zu erreichen, gibt es jeweils unterschiedliche Herangehensweisen und Verfahren, die dafür herangezogen werden können. Dabei gilt immer, dass ein höherer Anomysierungsgrad zwangsläufig mit einem geringeren Informationsgehalt einhergeht und umgekehrt (siehe Abbildung 1).

Abbildung 1
Darstellung des Zusammenhangs zwischen Anonymisierung und Informationsgehalt
Anonymität von Mikrodaten



Hierbei kommen zum einen klassische Vorgehensweisen, wie das Löschen einzelner kritischer Angaben, die Entfernung aller Beiträge einzelner exponierter Einheiten aus dem Datenbestand oder das Abschneiden von Ausreißern in den hohen oder niedrigen Wertebereichen durch ein sogenanntes Top- oder Bottom-Coding zum Einsatz. Zum anderen können hier auch neuere Verfahren wie die Mikroaggregation, die die Angaben mehrerer einzelner Beitragende vereinheitlicht, angewandt werden. Letzteres war beim Zensus 2011 mit dem Einsatz des sogenannten SAFE-Verfahrens der Fall (Tomann / Nickl 2013). Pre-tabulare Methoden haben dabei den Vorteil, in sich konsistente Datenbestände zu erzeugen, die flexibel ausgewertet werden können, ohne dass dabei in der Regel nochmals eine abschließende Geheimhaltungsprüfung vonnöten ist. Demgegenüber steht als Nachteil ein Eingriff in die Daten, der bereits vor der Auswertung Informationen aus dem ursprünglichen Datenmaterial entnimmt, die dann auch nicht mehr für detailliertere Analysen zur Verfügung stehen, ohne dass dies für den Datennutzenden zwangsläufig konkret ersichtlich ist.

Post-tabulare Methoden

In vielen Fällen wird bereits durch das Aggregieren von Einzelangaben zu statistischen Ergebnistabellen die Aufdeckung von Einzelangaben unmöglich gemacht, insbesondere wenn lediglich Gruppen oder Kategorien mit großen Besetzungszahlen ausgewiesen werden. Auch Berechnungsergebnisse, beispielsweise Mittelwerte oder bestimmte Indikatoren, sind tendenziell als weniger kritisch einzustufen. Anders gestaltet sich dies unter anderem bei der Analyse spezifischer Subgruppen, da diese erfahrungsgemäß oftmals nur geringe Fallzahlen aufweisen. In solchen Fällen ist die Anwendung von post-tabularen Geheimhaltungsverfahren notwendig, um den gesetzlichen Anforderungen gerecht zu werden.

Diese Methoden werden nach der Erstellung von Tabellen angewandt und zielen darauf ab, aufdeckungskritische Informationen gänzlich zu unterdrücken beziehungsweise zu verschleiern:

- **Zellsperrung:** Unterdrückung von ausgewählten Werten in Tabellenzellen in Kombination mit der sekundären Sperrung weiterer Angaben, die aufgrund der additiven Zusammenhänge innerhalb einer Tabelle zu einer Aufdeckung des eigentlichen unterdrückten Werts führen könnten. Entscheidungsleitend für die jeweiligen Sperrungen sind dabei statistikspezifisch festgelegte Regelungen zur Gewährleistung von Mindestfallzahlen bei der Besetzung von Tabellenzellen in Häufigkeitstabellen beziehungsweise zur Verhinderung der Aufdeckung von Dominanzfällen in Wertetabellen.

  Es handelt sich hierbei nach wie vor um die am weitesten verbreitete Methode zur statistischen Geheimhaltung von Tabellen in der amtlichen Statistik. Ihr Hauptnachteil liegt jedoch in der, gerade bei größeren oder untereinander zusammenhängenden Tabellen, oftmals manuellen und entsprechend arbeitsintensiven und zeitaufwendigen Umsetzung. Maschinelle Softwarelösungen hierzu sind zwar vorhanden, erfordern jedoch neben umfangreichen Vorarbeiten Expertenwissen bei der Durchführung und sind nach wie vor mit technikbedingten Einschränkungen behaftet. Darüber hinaus gestalten sich die hierfür notwendigen linearen Optimierungsläufe als teilweise äußerst zeitintensiv.

- **Rundung:** Deterministische, zufällige oder kontrollierte Rundung von Werten auf festgelegte Rundungsbasen. Hierbei handelt es sich um eine datenverändernde Gruppe von Geheimhaltungsverfahren, bei denen keine Angaben als kritisch identifiziert und unterdrückt werden, sondern stattdessen wird durch möglichst kleine Veränderungen (Perturbationen) der Originalangaben eine ausreichende Abweichung und damit Unsicherheit seitens eines potentiellen Datenangreifers hinsichtlich der tatsächlichen Ursprungswerte erzeugt.

- **Cell-Key-Methode** (u. a. Höhne / Höninger 2019; Enderle / Kleber 2024): Bei der Cell-Key-Methode – auch als post-tabulare stochastische Überlagerung bekannt – handelt es sich um eine zufallsbasierte Überlagerung von Werten, die besonders für flexible Auswertungsdatenbanken, bei denen eine Durchführung der statistischen Geheimhaltung in Echtzeit benötigt wird, geeignet ist. In ihrer ursprünglichen Form vom Australian Bureau of Statistics entwickelt, findet sie heute in angepasster Form auch in Deutschland, beispielsweise für den Zensus, Verwendung. In Abhängigkeit von vorab festgelegten Verfahrensparametern nimmt der zugrundeliegende Algorithmus für jede Angabe eines Tabellenfelds additive oder multiplikative Veränderungen vor – dem Originalwert wird ein stochastisches Rauschen hinzugefügt, womit eine exakte Offenlegung der Originalangabe nicht mehr möglich ist. Positiv hervorzuheben sind als Vorteile dieses Geheimhaltungsansatzes die gute Steuerbarkeit über die Parametrisierung, der geringe Ressourcenbedarf bei der Durchführung und die hohe Schutzwirkung. Zudem kann – im Gegensatz zu traditionellen Zellsperrverfahren – für jedes Tabellenfeld eine Angabe ausgewiesen werden. Die Auswirkungen des Verfahrens sind jedoch für Außenstehende nicht immer intuitiv erfassbar. Bei der Verwendung und Interpretation der ausgewiesenen Werte und insbesondere bei der vertieften Nutzung, beispielsweise im wissenschaftlichen Kontext in den Forschungsdatenzentren, muss daher mit Vorsicht vorgegangen werden (Rothe et al. 2024, Setzer et al. 2024). Eine entsprechende Kommunikation gegenüber den Nutzerinnen und Nutzern der amtlichen Statistik, die die Bereitstellung entsprechender Informationen und Dokumentationen miteinschließt, ist daher unerlässlich.

Die amtliche Statistik nutzt eine Vielzahl von Methoden und Verfahren, um Daten zu anonymisieren und Veröffentlichungen vor unerlaubten Rückschlüssen auf die dahinterstehenden statistischen Erhebungseinheiten – beispielsweise Personen, Unternehmen oder Betriebe – zu schützen.

Neue Methoden, neue Möglichkeiten, neue Herausforderungen – ein Blick in die nähere und fernere Zukunft

Die amtliche Statistik ist hinsichtlich der sich bietenden Möglichkeiten zur Erhebung, Analyse und Verbreitung statistischer Daten naturgemäß einem steten Wandel unterworfen. Dieser führte unter anderem dazu, dass gedruckte, nicht veränderbare Tabellen, die früher ein Standardergebnis der Arbeit der amtlichen Statistikerinnen und Statistiker waren, sich heute natürlich in Teilen immer noch finden lassen, aber zugleich um eine ganze Reihe moderner Darstellungs- und Veröffentlichungsformen ergänzt wurden. Die Verfügbarmachung von Ergebnissen für die breite Öffentlichkeit über moderne nutzerorientierte Datenbanken – wie GENESIS-Online oder die Zensusdatenbank, flexible Kartendarstellungen in Form von interaktiven Atlanten und Dashboards oder auch der Mikrodatenzugang für die empirische Wissenschaft über die Forschungsdatenzentren – flankieren heutzutage die traditionellen Veröffentlichungsformate. Jedoch erfordern diese neuen Daten- und Darstellungsarten, die ganz individuelle Herausforderungen an die statistische Geheimhaltung mit sich bringen, zumindest in Teilen auch neue Herangehensweisen.

Differential Privacy: Ein anderer Blick auf die Messbarkeit von Anonymität

Zugleich ergeben sich durch den interdisziplinären Austausch, insbesondere zwischen Statistik und Informatik, neue Anstöße und Blickwinkel, die sich für die amtliche Statistik als nutzenbringend erweisen können. Ein Beispiel hierfür ist die Suche nach einem Maß zur möglichst präzisen Quantifizierung von Aufdeckungsrisiken und Informationsgehalt. Diese beiden Aspekte sind untrennbar miteinander verbunden: Steigt der Informationsgehalt einer Datenbasis, so steigt auch die Möglichkeit der Aufdeckung von Einzelangaben, sinken die Aufdeckungsrisiken, so reduziert dies im Gegenzug den Nutzwert der Daten. Aus diesem Grund gilt es, eine für den jeweiligen Verwendungszweck adäquate Balance zwischen Informationsgehalt und Schutz der Daten zu finden. Beim Rückgriff auf traditionelle Methoden zur Anonymisierung ist hierbei neben viel Fingerspitzengefühl und Erfahrungswissen oftmals auch ausgeprägtes Trial-and-Error-Vorgehen vonnöten, um das gewünschte Ergebnis zu erzielen. Hier kommt nun das der Informatik entlehnte Konzept der **Differential Privacy** (Dwork et al. 2006, Garfinkel 2025) ins Spiel, welches eine eindeutige – sehr strikte – Operationalisierung der Anonymität eines gegebenen Datenbestands ermöglicht und sich in konkrete Kennzahlen fassen lässt.

Inwiefern sich dieses Konzept auch für den Bereich der amtlichen Statistik sinnvoll nutzen lässt, wird sich jedoch noch erweisen müssen (Drechsler 2023). Der erste große Einsatz durch ein nationales Statistikamt – in diesem Fall das U.S. Census Bureau für den Zensus 2020 – zeigte sowohl Licht als auch Schatten (Garfinkel 2022), lieferte zugleich aber auch wertvolle Erfahrungen, auf denen andernorts aufgebaut werden könnte.

Das steckt hinter Differential Privacy

Differential Privacy ist ein mathematischer Ansatz, der darauf abzielt, die Identität einzelner Personen in veröffentlichten Daten zu verschleiern. Dies wird in der Regel durch das Hinzufügen von stochastischem Rauschen (kleinen additiven oder subtraktiven Veränderungen gegenüber den Originalwerten) zu den Daten erreicht, um Unsicherheiten über den konkreten Beitrag einzelner Personen (oder anderer statistischer Einheiten) zum Gesamtdatenbestand zu schaffen. Im Gegensatz zu den traditionellen Vorgehensweisen im Bereich der statistischen Geheimhaltung liegt hier kein risikobasierter Ansatz, in dessen Rahmen unterschiedliche Angriffsszenarien, gegen die es sich zu wappnen gilt, spezifiziert werden, zugrunde. Somit ist auch keine Betrachtung des oftmals unbekannten und daher a priori kaum sinnvoll zu kalkulierenden Zusatzwissens eines potenziellen Datenangreifers vonnöten. Stattdessen ergibt sich die Einschätzung, inwiefern eine statistische Einheit durch die Offenlegung ihr zurechenbarer Angaben bedroht ist, alleine aus den vorhandenen Daten. Hierfür wird eine bestehende Datenbasis daraufhin untersucht, ob es durch Hinzunahme einer einzelnen statistischen Einheit zu einer Veränderung der inhaltlichen Aussagekraft kommt. Ist dies der Fall, so könnte die stattgefundene Veränderung – also die Differenz zwischen dem Datenbestand mit und ohne die betreffende Einheit – dieser Einheit eindeutig zugeordnet werden, was deren Privatheit gefährden würde. Differential Privacy verfolgt dabei das Ziel, eine solche Differenzbildung für jede in der Datenbasis enthaltene Einheit zu verhindern.

Kerndichteschätzung (KDE): Wie man informative Karten ohne Geheimhaltungsprobleme darstellen kann

Die Darstellung georeferenzierter Daten erfreut sich innerhalb der amtlichen Statistik zunehmender Beliebtheit, sei es in Form statischer Kartenabbildungen oder in Form interaktiver Atlanten. Doch durch diese neuen Verbreitungsmöglichkeiten entstehen zugleich über die bislang bekannten Fallstricke bei der Veröffentlichung von statistischen Tabellen hinausgehende Geheimhaltungsrisiken, die mitbedacht werden müssen. Zu den hierfür zur Verfügung stehenden Möglichkeiten, wie solchen Risiken begegnet werden kann (STACE 2024), zählt die Kerndichteschätzung (Kernel Density Estimation, KDE) als eine Methode zur Visualisierung georeferenzierter Daten. Sie ermöglicht die Darstellung von regionalen Schwerpunkten, ohne dabei zugleich Rückschlüsse auf dahinterstehende Einzelangaben zuzulassen. Die KDE ist ein nichtparametrisches Schätzverfahren zur Bestimmung der Wahrscheinlichkeitsdichte einer Zufallsvariablen auf Basis einer Stichprobe. In der amtlichen Statistik wird sie zunehmend als methodisch fundierte Alternative zur Zellsperrung eingesetzt, um georeferenzierte Daten kleinräumig darzustellen und gleichzeitig die Anforderungen der statistischen Geheimhaltung gemäß § 16 BStatG zu erfüllen (u. a. Mamonova 2024, Alfken / Rohde o. J.).

Das Verfahren beruht auf der Überlagerung symmetrischer Kernfunktionen (z. B. Gauß- oder Epanechnikov-Kern) über die Positionen einzelner Datenpunkte. Die resultierende Dichtefunktion ist eine geglättete, kontinuierliche Schätzung der Verteilung, die keine exakten Fallzahlen abbildet. Stattdessen ergibt sich für jede Gitterzelle ein aggregierter Dichtewert, der von den umliegenden Datenpunkten beeinflusst wird. Diese Werte sind nicht ganzzahlig und lassen keine Rückschlüsse auf Einzelangaben zu.

Die methodische Kontrolle der Geheimhaltung erfolgt über mehrere Parameter:

- **Bandbreite:** Steuert die Glättung der Dichtefunktion. Eine größere Bandbreite erhöht die Geheimhaltungssicherheit durch stärkere räumliche Verwischung.
- **Kernfunktion:** Die Wahl beeinflusst die Form und Ausdehnung der Dichteverteilung. Der Epanechnikov-Kern minimiert beispielsweise die mittlere quadratische Abweichung.
- **Zellgröße und Klassifikation:** Kleinere Zellen erhöhen die räumliche Auflösung, bergen aber ein höheres Reidentifikationsrisiko. Eine reduzierte Klassenanzahl bei der Farbcodierung unterstützt die Geheimhaltung.
- **Volumenverteilung:** Jeder Datenpunkt trägt anteilig zur Dichte benachbarter Zellen bei, abhängig von seiner Lage im Gitter. Dadurch entsteht eine räumlich geglättete Verteilung, die keine exakten Standorte preisgibt.

Die KDE erfüllt somit die Anforderungen an die faktische Anonymisierung durch mathematische Glättung und räumliche Aggregation. Sie erlaubt eine informationsreiche Visualisierung regionaler Muster, ohne die Vertraulichkeit sensibler Daten zu gefährden. In Kombination mit Geographischen Informationssystemen (GIS) und Open-Source-Software (z. B. R, QGIS, Python) kann die Methode effizient und reproduzierbar umgesetzt werden.

Synthetische Daten: Eine neue Perspektive für den Zugang zu vertraulichen Mikrodaten

Synthetische Daten bieten – zumindest in ausgewählten Nutzungskontexten – eine vielversprechende Alternative zur traditionellen pre-tabularen Geheimhaltung. Sie werden aus Originaldaten generiert und sollen deren wichtigste statistische Eigenschaften bewahren, während sie gleichzeitig das Risiko der Reidentifikation minimieren. Es gibt hierbei zwei unterschiedliche Hauptansätze, wie die Generierung der synthetischen Daten erfolgen kann:

Statistikbasierte Ansätze basieren auf Konzepten wie der multiplen Imputation und fokussieren auf statistische Inferenz. Sie wurden erstmals von Rubin (1993) vorgeschlagen und haben sich in der amtlichen Statistik etabliert. Beispiele für Anwendungen sind die Scottish Longitudinal Study (Nowok et al. 2017) und die EU-Statistiken zu Einkommen und Lebensbedingungen (EU-SILC) (de Wolf 2015).

In der Informatik werden maschinelle Lernverfahren wie Generative Adversarial Networks (GANs) genutzt, um synthetische Daten zu generieren. Diese **informatikbasierten Ansätze** zielen auf Vorhersageprobleme ab und haben sich in der Praxis als besonders effektiv erwiesen (u. a. Goodfellow et al. 2014). Beispiele sind die Anwendung von GANs beim U.S. Census 2020 (Abowd et al. 2022) und dem Israeli National Births Data Registry (Hod / Canetti 2025).

Die Nutzung synthetischer Daten bringt jedoch Herausforderungen mit sich, insbesondere bei der Messung des Analysepotenzials besteht ein Trade-off zwischen Datenschutz und Datenqualität. Methoden wie pMSE (Propensity Mean-Squared Error) oder Confidence Interval Overlap helfen, die Genauigkeit synthetischer Daten zu bewerten. Dennoch bleibt die Frage offen, wie die Validität der Daten für spezifische Analysen sichergestellt werden kann. Zudem bleiben auch bei synthetischen Daten Reidentifikationsrestrisiken teilweise weiterhin bestehen, insbesondere bei vollständig synthetischen Daten.

Fazit und Ausblick

Statistische Geheimhaltung als ein zentrales Grundprinzip der amtlichen Statistik spielt – auch und gerade heute – eine wichtige Rolle, wenn es darum geht, ein qualitativ hochwertiges Datenangebot mit dem Schutz von persönlichen Angaben und Betriebsgeheimnissen in Einklang zu bringen. Insbesondere technologische Weiterentwicklungen bringen dabei die amtliche Statistik unter methodischen Zugzwang, indem sie einerseits neue Möglichkeiten der Verwertung und Darstellung statistischer Daten eröffnen, zugleich aber auch ein wachsendes Potenzial für die Aufdeckung vertraulicher Angaben durch unbefugte Dritte mit sich bringen. Auch unter rechtlichen Gesichtspunkten erscheint es an der Zeit, neben anderen Aspekten des Statistikrechts auch den Themenbereich Geheimhaltung zu evaluieren und gegebenenfalls, wo notwendig, mit zeitgemäßen, an die stattgefundenen Veränderungen angepassten Leitplanken zu versehen.

Neue Elemente im methodischen Werkzeugkasten – in Teilen auch anderen Fachdisziplinen entliehen – dienen dazu, diesen an die amtliche Statistik gerichteten Anforderungen gerecht zu werden, wobei nicht jeder methodische Ansatz zugleich für jeden Verwendungszweck tauglich ist. Stattdessen gibt es eine Reihe verschiedener Ansätze für unterschiedliche Anwendungsszenarien. So haben beispielsweise synthetische Daten das Potenzial, die statistische Geheimhaltung im Bereich Mikrodaten maßgeblich zu beeinflussen. Deren Verwendung ermöglicht eine bessere Nutzung von Mikrodaten, beispielsweise in den Forschungsdatenzentren, ohne dabei die Privatsphäre der ursprünglich in die Datenbasis eingegangenen Personen zu gefährden. Mit der Kerndichteschätzung steht für die Darstellung georeferenzierter Auswertungen in Kartenform ein sowohl anschauliches als auch die statistische Geheimhaltung wahrendes Verfahren zur Verfügung. Und nicht zuletzt kann der verstärkte Austausch mit benachbarten Disziplinen dazu beitragen, dass sich beim allgemeinen Blick auf die praktische Umsetzung des Schutzes von personenbezogenen Daten und Betriebsgeheimnissen, wie im Fall des Differential Privacy-Ansatzes, neue Perspektiven auf alte Probleme ergeben können.

Workshop zur statistischen Geheimhaltung im Bayerischen Landesamt für Statistik am 19. September 2024

Dieser Artikel basiert in Grundzügen auf einem Vortrag im Rahmen eines Workshops zur statistischen Geheimhaltung, der am 19. September 2024 innerhalb einer wissenschaftlichen Veranstaltungsreihe im Bayerischen Landesamt für Statistik in Fürth stattfand. Beteiligt an der Veranstaltung waren hierbei neben Vortragenden des Bayerischen Landesamts für Statistik (LfStat) auch das Statistische Bundesamt (Destatis), Information und Technik Nordrhein-Westfalen (IT.NRW) sowie das Institut für Arbeits- und Berufsforschung (IAB) der Bundesagentur für Arbeit aus Nürnberg. Die behandelten Themen bezogen sich dabei stark auf aktuelle Forschungen zur Weiterentwicklung der statistischen Methodik, unter anderem im Bereich der datenverändernden Geheimhaltungsverfahren, der sicheren Darstellung von Geoinformationen sowie der Entwicklung und Nutzung synthetisch generierter Daten. Daneben wurden auch praxisorientierte Aspekte des Mikrodatenzugangs von den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder sowie des IAB-Forschungsdatenzentrums vorgestellt.

Es ist Bewegung gekommen in den Bereich der statistischen Geheimhaltung in der amtlichen Statistik, und auch wenn man im Einzelfall noch nicht abschätzen kann, welche der neuen Entwicklungen gekommen sind, um dauerhaft in den Statistischen Ämtern zu verbleiben, ist es auf jeden Fall zu begrüßen, dass das Thema sowohl methodisch als auch rechtlich verstärkt in den Fokus gerückt ist. ■

Literatur

Abowd, John M. / Ashmead, Robert / Cumings-Menon, Ryan / Garfinkel, Simson / Heineck, Micah / Heiss, Christine / Johns, Robert / Kifer, Daniel / Leclerc, Philip / Machanavajjhala, Ashwin / Sexton, William / Spence, Matthew / Zhuravlev, Pavel (2022): The 2020 Census Disclosure Avoidance System TopDown Algorithm. In: Harvard Data Science Review, (Special Issue 2), DOI: 10.1162/99608f92.529e3cb9

Alfken, Christoph / Rohde, Johannes (o. J.): Kerndichteschätzer zur Veröffentlichung von Karten mit georeferenzierten Daten der amtlichen Statistik: https://statistik.nrw/service/experimentelle-statistik/kerndichteschaetzer-zur-veroeffentlichung-von-karten-mit-georeferenzierten-daten-der-amtlichen (abgerufen am 22. Juli 2025).

Center of Excellence SDC (2024): Handbook on statistical Disclosure Control – Second edition: sdctools.github.io/HandbookSDC/Handbook-on-Statistical-Disclosure-Control.pdf (abgerufen am 12. Juni 2025).

de Wolf, Peter-Paul (2015): Public Use Files of EU-SILC and EU-LFS data, Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Helsinki, Finland.

Dorer, Peter / Mainusch, Helmut / Tubies, Helga (1988): BStatG. Bundesstatistikgesetz mit Erläuterungen. München: C. H. Beck.

Drechsler, Jörg (2023): Differential Privacy for Government Agencies – Are We There Yet? In: Journal of the American Statistical Association, 118: 541, 761–773, DOI: 10.1080/01621459.2022.2161385

Dwork, Cynthia / McSherry, Frank / Nissim, Kobbi / Smith, Adam (2006): Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi, Shai / Rabin, Tal: Theory of Cryptography, S. 265–284. Berlin-Heidelberg: Springer.

Enderle, Tobias / Kleber, Birgit (2024): Geheimhaltung mit der Cell-Key-Methode im Zensus 2022. In: Wirtschaft und Statistik 06/2024, S. 82–91.

Garfinkel, Simson (2022). Differential Privacy and the US Census 2020: mit-serc.pubpub.org/pub/differential-privacy-2020-us-census/release/2 (abgerufen am 15. Juli 2025).
DOI: 10.21428/2c646de5.7ec6ab93

Garfinkel, Simson (2025). Differential Privacy. Cambridge/London: The MIT Press.

Goodfellow, Ian / Pouget-Abadie, Jean / Mirza, Mehdi / Xu, Bing / Warde-Farley, David / Ozair, Sherjil / Courville, Aaron / Bengio, Yoshua (2014): Generative adversarial nets. In: Advances in Neural Information Processing Systems,
DOI: 10.1145/3422622

Hlawatsch, Anja / Meyer, Karen / Rothe, Patrick (2022): 20 Jahre Forschungsdatenzentrum der Statistischen Ämter der Länder – Das Daten- und Dienstleistungsangebot für wissenschaftliche Nutzungen von Mikrodaten der amtlichen Statistik. In: Bayern in Zahlen 11/2022, S. 25–33.

Hod, Shlomi / Canetti, Ran (2025): Differentially Private Release of Israel's National Registry of Live Births, arXiv: 2405.00267v2

Höhne, Jörg / Höninger, Julia (2019): Die Cell-Key-Methode – ein Geheimhaltungsverfahren. In: Zeitschrift für amtliche Statistik Berlin Brandenburg. Ausgabe 3+4/2018, S. 14–19.

Kühling, Jürgen (2023): BStatG. Bundesstatistikgesetz. Kommentar. C. H. Beck: München.

Mamonova, Swetlana (2024): Die Kerndichteschätzung als eine innovative Visualisierungsmethode georeferenzierter Daten in der amtlichen Statistik. In: Statistisches Monatsheft Baden-Württemberg 5/2024, S. 41–45.

Nowok, Beata / Raab, Gillian / Dibben Chris (2017): Providing bespoke synthetic data for the UK longitudinal studies and other sensitive data with the synthpop package for R. In: Statistical Journal of the IAOS, 33(3):785–796, DOI: 10.3233/SJI-150153

Rothe, Patrick (2015 a): Statistische Geheimhaltung – Der Schutz vertraulicher Daten in der amtlichen Statistik. Teil 1: Rechtliche und methodische Grundlagen. In: Bayern in Zahlen 05/2015, S. 294–303.

Rothe, Patrick (2015 b): Statistische Geheimhaltung – Der Schutz vertraulicher Daten in der amtlichen Statistik. Teil 2: Herausforderungen und aktuelle Entwicklungen. In: Bayern in Zahlen 08/2015, S. 482–489.

Rothe, Patrick / Güttgemanns, Volker / Rohde, Johannes / Setzer, Stefanie (2024): Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. Teil 2: Auswirkungen des neuen Geheimhaltungsverfahrens. In Wirtschaft und Statistik 03/2024, S. 45–54.

Rubin, Donald (1993). Discussion: Statistical disclosure limitation. In: Journal of Official Statistics, 9: 462–468.

Setzer, Stefanie / Rohde, Johannes / Güttgemanns, Volker / Rothe, Patrick (2024): Die Cell-Key-Methode in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. Teil 1: Vorstellung des neuen Geheimhaltungsverfahrens. In: Wirtschaft und Statistik 03/2024, S. 31–44.

STACE (2024): Guidelines for Statistical Disclosure Control Methods Applied on Geo-Referenced Data: github.com/sdcTools/GeoSpatialGuidelinesSources/releases/download/v1.0/Geo_SDC_Guidelines.pdf (abgerufen am 12. Juni 2025).

Tomann, Jörg / Nickl, Andreas (2013): ZENSUS 2011: Die Zensusdatenbank. In: Bayern in Zahlen 04/2013, S. 186–189.