Fachgespräch mit Oberregierungsrätin Sarah Giessing

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

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

Womit beschäftigt sich die Bund-Länder-Arbeitsgruppe Geheimhaltung?

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

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

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

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

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

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

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

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

Gibt es „das eine" perfekte Geheimhaltungsverfahren?

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

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

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

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

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

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

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

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