

Chancen und Herausforderungen – Neue Entwicklungen im Bereich der statistischen Geheimhaltung: Ein Bericht von der UNECE Worksession on Statistical Confidentiality in Helsinki

Dipl.-Soz. Patrick Rothe

Genauso wie die amtliche Statistik als Ganzes mit dem stetigen technologischen und gesellschaftlichen Wandel konfrontiert ist, trifft dies auch für den Teilbereich der statistischen Geheimhaltung zu. Gewandelte Rahmenbedingungen und veränderte technische Möglichkeiten erfordern eine kontinuierliche Modernisierung und Anpassung an neue Gegebenheiten, Risiken und Chancen. Im Folgenden soll anhand ausgewählter Vorträge der UNECE Worksession on Statistical Confidentiality ein Überblick über die aktuell wichtigsten internationalen Entwicklungen in den Bereichen Mikrodatenzugang und statistische Geheimhaltung gegeben werden.

Veranstaltungsort der alle zwei Jahre von der Wirtschaftskommission für Europa der Vereinten Nationen (UNECE) in Kooperation mit Eurostat ausgetragenen Konferenz war vom 5. bis 7. Oktober 2015 Helsinki. Als Gastgeber fungierte Statistics Finland mit seinem Hauptgebäude im Zentrum der finnischen Hauptstadt. Ziel der Veranstaltung ist es, Geheimhaltungs- und Mikrodatenexperten aus amtlicher Statistik und wissenschaftlicher Forschung auf internationaler

Ebene miteinander ins Gespräch zu bringen und den Austausch von Erfahrungen und Best-Practice-Beispielen sowie einen vertieften fachlichen Diskurs zu ermöglichen. Dementsprechend drehte sich an den zweieinhalb Veranstaltungstagen mit mehr als 40 Vorträgen und Präsentationen alles um das Thema statistische Geheimhaltung und seine zahlreichen Facetten. Aus Deutschland waren neben Vertretern des Statistischen Bundesamts und des Bayerischen Landesamts für Statistik auch Angehörige des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit in Nürnberg sowie Wissenschaftler der Hochschule für Technik und Wirtschaft des Saarlands in Saarbrücken als Tagungsteilnehmer vor Ort.

Zugang zu amtlichen Mikrodaten

Einen der inhaltlichen Schwerpunkte stellte der Zugang zu amtlichen Mikrodaten für wissenschaftliche Forschungszwecke dar: Als Einstieg in dieses Thema wurde von Aleksandra Bujnowska (Eurostat) ein Überblick über die verschiedenen derzeit vorhandenen Möglichkeiten zur wissenschaftlichen Nutzung von Mikrodaten der europäischen Gemeinschaftsstatistiken gegeben (Bujnowska 2015). Diese werden – insbesondere seit der 2013 erfolgten Neufassung der zugrunde liegenden EU-Verordnung – sehr rege genutzt, wobei zukünftig mit weiterhin ansteigenden Nutzungszahlen zu rechnen ist. Auch



Als Veranstaltungsort diente das Gebäude von Statistics Finland in der Työpajankatu in Helsinki.
Foto: Takayuki Ito, National Statistics Center Japan

plant Eurostat mittelfristig eine Ausweitung sowohl des Datenangebots als auch der Möglichkeiten zur Datennutzung. Die als Scientific-Use-Files¹ beziehbaren oder in einem Safe-Centre bei Eurostat in Luxemburg vor Ort nutzbaren Daten basieren auf den von den EU-Mitgliedstaaten übermittelten Einzeldaten und bieten durch die vorgenommene Harmonisierung eine der aktuell besten Möglichkeiten, sich international vergleichend mit sozial- und wirtschaftswissenschaftlichen Fragestellungen auseinanderzusetzen. Besonderer Beliebtheit unter den wissenschaftlichen Nutzern erfreuen sich dabei die Daten der EU-SILC-Erhebung (Gemeinschaftsstatistik über Einkommen und Lebensbedingungen) sowie der LFS-Erhebung (Arbeitskräftestichprobe der EU).

Neben Eurostat stellte eine Reihe nationaler Statistikämter den aktuellen Stand ihrer Zugangsangebote für die wissenschaftliche Nutzung von Mikrodaten vor, wobei besonders die länderübergreifende Kooperation der skandinavischen Staaten (Nielsen/Thaulow 2015) und das norwegische Projekt zum Remote-Access auf Registerdaten (Heldal 2015) mit einer eigens entwickelten Fernzugriffsumgebung inklusive eigener Programmiersprache hervorstachen. Dass eine Kooperation nicht nur über Ländergrenzen, sondern über ganze Kontinente hinweg ebenfalls möglich und sinnvoll ist, unterstrich die Vorstellung des Austauschs von Wissen und Software zwischen dem Australian Bureau of Statistics (ABS) und Statistics Canada beziehungsweise Statistics Sweden. Konkret handelt es sich dabei im kanadischen Fall um die Übernahme und Anpassung des vom ABS eingesetzten Systems DataAnalyzer, das weitgehende Analysemöglichkeiten innerhalb einer Auswertungsdatenbank eröffnet, an die eigenen Bedürfnisse (Simard/Lalor 2015). Auch Statistics Sweden versucht derzeit, auf den Vorarbeiten des ABS aufzubauen, und stellte erste Ergebnisse eines noch in einem frühen Anfangsstadium befindlichen Geheimhaltungsprojektes vor (Andersson et al. 2015). Dieses hat die Implementierung eines datenverändernden, auf einer post-tabularen stochastischen Überlagerung basierenden Geheimhaltungsverfahrens, wie es vom ABS zur Geheimhaltung von Zensusergebnissen eingesetzt wird, zum Ziel.

Einen Blick in eine mögliche Zukunft der Arbeit mit amtlichen Mikrodaten bot die Vorstellung sogenannter virtueller Forschungsumgebungen. Diese werden als ein wichtiger Schritt angesehen, um die enge Zusammenarbeit zwischen Wissenschaftlern verschiedener Einrichtungen an gemeinsamen Forschungsprojekten mit vertraulichen Daten zu erleichtern. Vor dem Hintergrund, dass wissenschaftliche Forschung heute zunehmend dezentral, international und interdisziplinär stattfindet, ergeben sich sowohl für die akademische Forschung als auch für die amtliche Statistik neue Herausforderungen. Wie eine infrastrukturelle Antwort hierauf in Form einer virtuellen Forschungsumgebung aussehen könnte, wurde von David Schiller vom IAB in Nürnberg skizziert (Schiller 2015). Als Ziel wird dabei angestrebt, die physische Forschungsumgebung – bestehend aus Computerhardware, Datenbanken und Speichersystemen – mit der intellektuellen Umgebung in Form der beteiligten Projektpartner, von Expertennetzwerken und des dort vorhandenen immateriellen Wissens zu verknüpfen. Grundvoraussetzung hierfür ist das Vorhandensein einer gesicherten Verbindung zwischen diesen beiden Bereichen, die die Forschungsumgebung unter Datenschutz- und Sicherheitsgesichtspunkten wirksam nach außen abschottet. Außerdem muss eine solche Lösung in der Lage sein, den gesamten Zyklus, den ein Forschungsprojekt durchläuft, abzudecken, um über sämtliche Phasen hinweg ein erfolgreiches Arbeiten zu ermöglichen und damit eine tatsächliche Zukunftsalternative darzustellen.

Weniger virtuell, sondern vielmehr von physischer Natur stellt sich der an der University of Edinburgh entwickelte SafePOD dar (Dibben 2015). Dieser beinhaltet im Inneren eines würfelförmigen Containers im Format 2m x 2m x 2m einen standardisierten Computerarbeitsplatz für Gastwissenschaftler, der in einer vollständig kontrollier- und überwachbaren, nach außen abgeschirmten Arbeitsumgebung per Remote Access den Zugriff auf vertrauliche Mikrodaten ermöglicht. Insbesondere für den schnellen Aufbau von standardisierten Forschungsdatennetzwerken soll er hierdurch geeignet sein. Ob es sich bei dem SafePOD jedoch wirklich wie erhofft um eine für die Datennutzer angenehme Arbeitsatmosphäre handelt –

¹ Scientific-Use-Files sind amtliche Mikrodatenbestände, zu denen ausschließlich wissenschaftliche Nutzer Zugang erhalten können. Es handelt sich dabei um faktisch anonyme Daten, d. h. diese wurden im Vorfeld derart bearbeitet, dass eine Zuordnung von enthaltenen Daten zu einer konkret identifizierbaren Erhebungseinheit, z. B. einer befragten Person, nur unter der Inkaufnahme eines irrational hohen Aufwands in Form von Zeit und Ressourcen möglich ist. In der Regel wird die Anonymisierung noch durch vertragliche Regelungen ergänzt.

und inwiefern die ständige Überwachbarkeit durch die im Inneren montierte Videokamera dieser abträglich sein könnte – war unter den Zuhörern alles andere als unumstritten.

Neben den Möglichkeiten zur technischen Umsetzung spielen auch organisatorisch-institutionelle Punkte beim Thema Mikrodatenzugang eine wichtige Rolle. Maurice Brandt vom Statistischen Bundesamt lenkte in seiner Präsentation den Fokus verstärkt auf diese Aspekte des Themas und stellte das Konzept des sogenannten „Circle of Trust“ vor (Brandt 2015). Dieses wurde in der Paris Group der OECD, die sich mit Best-Practice-Beispielen für den Zugang zu Mikrodaten beschäftigt hat, erarbeitet und setzt auf die Schaffung gegenseitigen Vertrauens, das als Basis für eine zunehmend vertiefte Kooperation zwischen amtlichen und wissenschaftlichen Akteuren dienen soll. Festgelegte Mindeststandards – unter anderem die statistische Geheimhaltung, aber auch die generell vorhandene fachliche Expertise betreffend –, zu denen sich jedes Mitglied verpflichtet, sowie eine je nach beteiligtem Akteur – Datenproduzent, Datenarchiv etc. – angepasste Abstufung in verschiedene Vertrauenszonen stellen dabei die Arbeitsgrundlage dar. Dabei fußt das gegenseitige Vertrauen auf rechtlichen Vereinbarungen und organisatorischen Festlegungen, die den Rah-

men für alle darauf aufbauenden Folgeschritte bilden und eine dauerhafte Evolution der Zusammenarbeit möglich machen sollen.

Dass die Interessen amtlicher Datenproduzenten und der Datennutzer aus Wissenschaft und Forschung zwangsläufig nicht immer deckungsgleich ausfallen, wurde bei der Präsentation von Felix Ritchie von der University of the West of England in Bristol deutlich, der vehement einen Bewusstseinswandel auf Seiten der amtlichen Statistik einforderte. Das traditionelle, an der Vermeidung des „worst case“ ausgerichtete Sicherheitsverständnis von statistischer Geheimhaltung kritisierte er als nicht mehr zeitgemäß, zu theoriegläubig, zu datenzentriert und zu wenig nutzerorientiert. Die gängige Vorstellung eines böswilligen Datenangreifers sei empirisch nicht haltbar, zudem verbaue sich die amtliche Statistik mit ihrer restriktiven Datenfreigabe („default-closed“-Einstellung) die Möglichkeit, selbst von den Ergebnissen einer verstärkten Nutzung ihrer Daten zu profitieren. Der Öffentlichkeit werde hierdurch ein großer gesamtgesellschaftlicher Nutzen vorenthalten. Die Lösung hierfür sei ein Schwenk hin zu einer „default-open“-Strategie, die so viele Daten wie möglich verfügbar macht, der Erkenntnis, dass Wissenschaftler in der Realität nicht böswillig handeln, sondern unbewusst Fehler machen, sowie der Annahme, dass Zugangseinrichtungen grundsätzlich unsicher sind und dass Risikoeinschätzungen immer subjektiven Verzerrungen unterliegen. Zugleich sollte – wann immer möglich – nicht-statistischen Sicherungsvorkehrungen der Vorzug gegenüber datenbezogenen Maßnahmen gegeben werden. Außerdem sei es dringend notwendig anzuerkennen, dass Fehler niemals völlig auszuschließen sein werden, dementsprechend sollten Planungen zum Schutz der Daten auch nicht auf dieses Ziel ausgerichtet werden (Hafner et al. 2015).

Verfahren zur statistischen Geheimhaltung

Den zweiten Schwerpunkt neben dem Zugang zu Mikrodaten für wissenschaftliche Zwecke bildete die Auseinandersetzung mit unterschiedlichen Verfahren zur statistischen Geheimhaltung. Besonders der Rückgriff auf die Generierung (teil-)synthetischer Daten, um hiermit Geheimhaltungsprobleme zu umgehen, nahm einen breiten Raum ein.



Viele der Vorträge beschäftigten sich mit der Generierung synthetischer Daten, so auch diese Präsentation von Prof. Dr. Kiyomi Shirakawa von der Hitotsubashi University in Tokio.
Foto: Takayuki Ito, National Statistics Center Japan

Spezielle Softwarepakete können dabei die Erstellung möglichst realistischer künstlicher Daten unterstützen, indem versucht wird, wichtige strukturelle Zusammenhänge, die in den Originaldaten enthalten sind, in einem synthetischen Datenbestand nachzubilden (Nowok 2015). Wie die Verwendung synthetisch erstellter Daten in der Praxis aussehen kann und welche Anwendungsgebiete hierfür denkbar sind, wurde von Peter-Paul DeWolf von Statistics Netherlands in seiner Vorstellung der Methodik hinter den kommenden Public-Use-Files² zu EU-SILC und LFS, die zukünftig von Eurostat bereitgestellt werden sollen, präsentiert (De Wolf 2015). Mit der Verwendung synthetischer Daten im Fall von EU-SILC ist dabei zum einen jederzeit der Schutz der vertraulichen Angaben gewährleistet. Zugleich ermöglicht die durch die vorgenommene Modellierung erzielte Realitätsnähe der Daten es den Wissenschaftlern, bereits vor Durchlaufen des Beantragungsprozesses und vor Erhalt der eigentlichen Forschungsdaten Berechnungen durchzuführen und anhand derer die Tauglichkeit der Daten für die gewünschten Auswertungen zu beurteilen. Dem gegenübergestellt wurde das Vorgehen bei der Anonymisierung der LFS-Daten, bei der auf traditionelle Methoden aus dem informationsreduzierenden Bereich zurückgegriffen wurde. Auch hiermit wird ein Sicherheitsniveau erreicht, das es ermöglicht, diese Daten als absolut anonymes Public-Use-File der Öffentlichkeit zur Verfügung zu stellen. Zum selben Oberthema wurde eine ganze Reihe weiterer Forschungsarbeiten vorgestellt, wobei auch mögliche Nachteile und Probleme bei der Arbeit mit (teil-)synthetischen Datenbeständen aufgezeigt wurden. So ging Jörg Drechsler vom IAB in seinem Vortrag detailliert auf Untersuchungen zur Verwendung synthetisch erzeugter Geodaten und die dabei zutage getretenen Einschränkungen hinsichtlich der Balance aus Datenqualität und Schutz der Daten ein (Drechsler/Hu 2015).

Um eine effiziente Durchführung der Geheimhaltung zu ermöglichen, ist eine möglichst nutzerfreundliche Integration in den Arbeitsablauf notwendig. Je geringer die Hürden bei Bedienbarkeit und Zugänglichkeit ausfallen, desto höher dürfte die Wahrscheinlichkeit eines erfolgreichen Einsatzes neuartiger Lösungen sein. Wie dies mit dem Geheimhaltungstool Tau-Argus³ in der Praxis aussehen kann, zeigte der Kon-



Kaffeepausen boten genauso wie ein Sektempfang und ein Conference Dinner die Möglichkeit zum gegenseitigen Kennenlernen und vertieften Erfahrungsaustausch.
Foto: Takayuki Ito, National Statistics Center Japan

ferenzbeitrag von Sarah Giessing und Sven Grunwald vom Statistischen Bundesamt. Besonders bei der sogenannten tabellenübergreifenden Geheimhaltung, bei der nicht nur eine einzelne Tabelle, sondern eine Vielzahl inhaltlich verwandter Tabellen betrachtet werden muss, stößt man bei der Nutzung von Tau-Argus unter Bedienungsgesichtspunkten auf merkliche Einschränkungen. Abhilfe schafft hier ein vom Statistischen Bundesamt programmiertes SAS-Makro, über das mittels einer grafischen Nutzeroberfläche eine komfortablere Steuerung der maschinellen Geheimhaltung ermöglicht wird (Giessing/Grunwald 2015).

In der abschließenden Podiumsdiskussion am Ende der Konferenz wurde der Frage nachgegangen, auf welche zukünftigen Herausforderungen die statistische Geheimhaltung sich – insbesondere auch durch das Internet, soziale Medien und Big Data – einstellen muss. Mark Elliot (University of Manchester) und Josep Domingo-Ferrer (Universität Rovira et Virgili Barcelona) – beide seit vielen Jahren forschend im Bereich der statistischen Geheimhaltung tätig – sowie Felix Ritchie, der wiederum mit Nachdruck die Interessen der wissenschaftlichen Community vertrat, setzten sich teilweise kontrovers mit dem Thema auseinander. Trotz der unterschiedlichen Sichtweisen gab es zentrale Punkte, auf die

² Bei Public-Use-Files handelt es sich um absolut anonyme Datenbestände, die öffentlich verfügbar gemacht werden können. Oftmals werden diese auch als sogenannte CAMPUS-Files zu Übungszwecken in der universitären Lehre eingesetzt.

³ Bei Tau-Argus handelt es sich um ein im Rahmen eines Eurostat-Projekts entwickeltes Programm zur Automatisierung von Geheimhaltungsprozessen mit Schwerpunkt auf der Durchführung von Zellsperungsverfahren.

sich alle Diskutanten einigen konnten: Erklärtes Ziel bei der statistischen Geheimhaltung müsse immer die Suche nach der richtigen Balance zwischen Datenschutz und Qualität der Daten sein. Dabei dürfe das eine Ziel nicht zugunsten des jeweils anderen aufgegeben werden. Als eine der wichtigsten Fähigkeiten in der Informationsgesellschaft wurde die Kompetenz jedes Einzelnen für einen souveränen Umgang mit seinen Daten angesehen. Nur dann ist – auch jenseits der statistischen Geheimhaltung – ein umfassender Schutz der Privatsphäre in der Zukunft überhaupt möglich.

Schlussbetrachtung

Als Ergebnis der Veranstaltung kann festgehalten werden, dass derzeit auf internationaler Ebene eine Vielzahl unterschiedlicher Verfahren und Lösungen im Bereich der Geheimhaltung erprobt wird. Die Probleme, mit denen sich amtliche Datenproduzenten konfrontiert sehen, ähneln sich dabei weltweit, auch wenn die rechtlichen und kulturellen Bedingungen variieren mögen. Dabei kommen zunehmend elaboriertere Verfahren, beispielsweise aus dem Bereich der Datenveränderung und der Erzeugung syn-

thetischer Daten zum Einsatz. Aber auch traditionelle Anonymisierungs- und Geheimhaltungsmethoden, wie Aggregation oder Zellsperren, haben nach wie vor ihre Daseinsberechtigung. Dabei spielt die Suche nach der Balance zwischen dem Schutz vertraulicher Daten und dem beinhalteten informativen Gehalt eine bedeutsame Rolle. Doch das beste Verfahren nützt nichts, wenn es unkomfortabel zu bedienen ist und nicht in den Arbeitsalltag integriert werden kann. Daher kommt auch der Weiterentwicklung in diesem Bereich eine große Bedeutung zu. Zu guter Letzt sind es aber sowohl die Datennutzer aus der Wissenschaft als auch die breite Öffentlichkeit, für die die Ergebnisse der amtlichen Datenproduzenten erstellt werden; dementsprechend ist es wichtig, deren individuelle Bedürfnisse zu berücksichtigen und diese – unter Wahrung der statistischen Geheimhaltung – bei der Nutzung der vorhandenen Daten zu unterstützen.

Der Veranstaltungsort der nächsten Konferenz wird auf Einladung des dortigen nationalen Statistikamts im Jahr 2017 die mazedonische Hauptstadt Skopje sein.

Literaturangaben

- Andersson, K./Jansson, I./Kraft, K. (2015), Protection of frequency tables – current work at Statistics Sweden. Workingpaper zur Joint UNECE/Eurostat worksession on statistical confidentiality vom 5. bis 7. Oktober 2015 in Helsinki.
- Brandt, M. (2015), Circle of Trust. Workingpaper zur Joint UNECE/Eurostat worksession on statistical confidentiality vom 5. bis 7. Oktober 2015 in Helsinki.
- Bujnowska, A. (2015), Access to EU microdata for research purposes. Workingpaper zur Joint UNECE/Eurostat worksession on statistical confidentiality vom 5. bis 7. Oktober 2015 in Helsinki.
- De Wolf, P.-P. (2015), Public Use Files of EU-SILC and EU-LFS data. Workingpaper zur Joint UNECE/Eurostat worksession on statistical confidentiality vom 5. bis 7. Oktober 2015 in Helsinki.
- Dibben, C. (2015), Micro, remote, safe settings (safe-PODS) – extending a safe setting network across a country. Workingpaper zur Joint UNECE/Eurostat worksession on statistical confidentiality vom 5. bis 7. Oktober 2015 in Helsinki.
- Drechsler, J./Hu, J. (2015), Generating synthetic geocoding information for public release. Workingpaper zur Joint UNECE/Eurostat worksession on statistical confidentiality vom 5. bis 7. Oktober 2015 in Helsinki.
- Giessing, S./Grunwald, S. (2015), A Graphical User Interface to Manage Cell Suppression on Sets of Linked Tables using SAS and Tau-Argus. Workingpaper zur Joint UNECE/Eurostat worksession on statistical confidentiality vom 5. bis 7. Oktober 2015 in Helsinki.
- Hafner, H.-P./Lenz, R./Ritchie, F./Welpton, R. (2015), Evidence-based, context-sensitive, user-centred, risk-managed SDC planning: designing data access solutions for scientific use. Workingpaper zur Joint UNECE/Eurostat worksession on statistical confidentiality vom 5. bis 7. Oktober 2015 in Helsinki.

- Heldal, J. (2015), RAIRD – Remote Access Infrastructure for Register Data. Workingpaper zur Joint UNECE/Eurostat worksession on statistical confidentiality vom 5. bis 7. Oktober 2015 in Helsinki.
- Nielsen, C./Thaulow, I. (2015), Joint access to nordic microdata. Workingpaper zur Joint UNECE/Eurostat worksession on statistical confidentiality vom 5. bis 7. Oktober 2015 in Helsinki.
- Nowok, B. (2015), synthpop: An R package for generating synthetic versions of sensitive microdata for statistical disclosure limitation. Workingpaper zur Joint UNECE/Eurostat worksession on statistical confidentiality vom 5. bis 7. Oktober 2015 in Helsinki.
- Schiller, D. (2015), Virtual Research Environments (VREs) to enable access to confidential data for scientific purposes. Workingpaper zur Joint UNECE/Eurostat worksession on statistical confidentiality vom 5. bis 7. Oktober 2015 in Helsinki.
- Simard, M./Lalor, T. (2015), Confidentialized Analysis of Microdata CSPA project. Workingpaper zur Joint UNECE/Eurostat worksession on statistical confidentiality vom 5. bis 7. Oktober 2015 in Helsinki.