Maurice Brandt, M. A., Dr. Markus Zwick

infinitE – Eine informationelle Infrastruktur für das E-Science Age

Verbesserung des Mikrodatenzugangs durch „Remote-Access“

Nach ihrer Einrichtung im Jahr 2001 haben sich die Forschungsdatenzentren (FDZ) der Statistischen Ämter des Bundes und der Länder inzwischen so etabliert, dass sie aus der empirisch arbeitenden Wissenschaft nicht mehr wegzudenken sind. Die Nutzung von Mikrodaten aus der amtlichen Statistik erfreut sich einer breiten und weiter zunehmenden Beliebtheit.[1])

Dabei stellen die Forschungsdatenzentren einen fundamentalen Wandel der Nachfrage nach ihren Produkten fest: Die kontrollierte Datenfernverarbeitung wie auch der Gastwissenschaftlerarbeitsplatz haben sich mittlerweile zu den gefragtesten Zugangsformen zu den wirtschaftsstatistischen Einzeldaten entwickelt.

Die Nachfrage nach einer On-Site-Nutzung statistischer Mikrodaten ist mittlerweile so groß geworden, dass die Datenwünsche von den jeweiligen Forschungsdatenzentren nur noch unter großen Anstrengungen zeitnah zu bearbeiten sind. Ein Grund hierfür ist, dass diese Zugangsform für die Mitarbeiter und Mitarbeiterinnen in den Forschungsdatenzentren sehr aufwendig ist. Zum einen müssen Programme wiederholt angepasst werden, da die derzeitigen Datenstrukturfiles aufgrund der sehr starken Anonymisierung kein gutes Abbild der Originaldaten darstellen. Zum anderen beansprucht die manuelle Ergebniskontrolle und Prüfung auf Geheimhaltung sehr viel Zeit.

An beiden Punkten soll das im Folgenden beschriebene Forschungsprojekt „Eine informationelle Infrastruktur für das ‚E-Science Age‘ (infinitE) auf dem Weg zum ‚Remote-Access‘ – Verbesserung der kontrollierten Datenfernverarbeitung bei wirtschaftsstatistischen Daten durch Datenstrukturfiles und automatisierte Ergebniskontrolle“ ansetzen. Es stellt einen Meilenstein auf dem Weg zum automatisierten Fernrechnen dar.

Das Forschungsdatenzentrum des Statistischen Bundesamtes führt dieses Projekt gemeinsam mit dem Institut für Angewandte Wirtschaftsforschung, dem Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit sowie mit den Forschungsdatenzentren der Statistischen Ämtern der Länder mit den Standorten Berlin und Hessen durch. Das Projekt wird vom Bundesministerium für Bildung und Forschung finanziell gefördert.

Begleitet wird das Projekt von der vom Rat für Sozial- und Wirtschaftsdaten (RatSWD) einberufenen Arbeitsgruppe „Future Data Access“, die vornehmlich eine mögliche Änderung der rechtlichen Rahmenbedingungen des Datenzugangs in Deutschland prüfen soll.

1 Aktuelle Situation in den Forschungsdatenzentren

Die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder ermöglichen der Wissenschaft einen geregelten Zugang zu amtlichen Mikrodaten. Dabei bieten sie unterschiedliche Zugangswege zur Nutzung der Mikrodaten der amtlichen Statistik an, die – ebenso wie das Datenangebot – laufend weiterentwickelt und ausgebaut werden.

1) Siehe Rolf-Engel, G./Wagner, G. G./Zwick, M.: „Fortschritte der informationellen Infrastruktur in Deutschland“, Festschrift für Johann Hahlen zum 65. Geburtstag und Hans-Jürgen Krupp zum 75. Geburtstag, Baden-Baden 2008.

Absolut anonymisierte Mikrodaten werden in Form sogenannter Public-Use-Files allen interessierten Personen zur Verfügung gestellt. Eine weitere Möglichkeit, Mikrodaten zu nutzen, stellen die sogenannten Scientific-Use-Files dar, das sind faktisch anonymisierte Mikrodaten auf CD oder DVD, die am eigenen Arbeitsplatz in der jeweiligen Institution ausgewertet werden können. Die Scientific-Use-Files finden im Bereich der personenbezogenen Erhebungen, wie zum Beispiel beim Mikrozensus, hohen Anklang. Bei wirtschaftsstatistischen Daten müssen die Anonymisierungsmaßnahmen aufgrund des höheren Reidentifikationsrisikos allerdings weiter gehen als bei personenbezogenen Daten. Hier sind datenverändernde Verfahren, wie Mikroaggregation oder stochastische Überlagerung, unvermeidbar, um die Unternehmen vor Deanonymisierung zu schützen.

Aus Vorbehalten gegenüber den datenverändernden Anonymisierungsverfahren und auch aufgrund der längeren Wartezeiten bis zur Bereitstellung eines Scientific-Use-Files, verschiebt sich die Nachfrage bei wirtschaftsstatistischen Daten derzeit von der Off-Site-Nutzung immer mehr in Richtung einer On-Site-Nutzung originaler Mikrodaten[2]), insbesondere in Form der kontrollierten Datenfernverarbeitung. Dieser Zugangsweg hat für Wissenschaftler den Vorteil, dass keine Reisen und Aufenthalte in einem oder mehreren Forschungsdatenzentren notwendig sind.

Für die Mitarbeiterinnen und Mitarbeiter in den Forschungsdatenzentren, aber auch für die Nutzer, gestaltet sich die On-Site-Nutzung von Mikrodaten derzeit sehr aufwendig, da die personellen Ressourcen in den Forschungsdatenzentren nicht ausreichen, um in größerem Umfang Fernrechenaufträge auszuführen.

Wenn ein solcher Antrag gestellt wird, erhält der Forscher bzw. die Forscherin ein sogenanntes Datenstrukturfile, das aus einem komplett anonymisierten Datensatz mit identischer Struktur wie die Originaldaten besteht und anhand dessen das Programm für die Datenanalyse geschrieben werden kann. Da die derzeitigen Datenstrukturfiles aufgrund der sehr starken Anonymisierung (Stichprobenziehung und Vertauschung) kein gutes Abbild der Originaldaten darstellen, müssen die Programme wiederholt angepasst werden. Besonders viel Zeit erfordern aber die manuelle Kontrolle der Ergebnisse und die Prüfung auf Geheimhaltung. Diese Überprüfung ist bei komplexen Tabellen und umfangreichen Schätzoutputs außerordentlich schwierig.

Mit den neuen Projekten „Amtliche Firmendaten für Deutschland“ und „Kombinierte Firmendaten für Deutschland“[3]) arbeiten die statistischen Ämter darüber hinaus gemeinsam mit weiteren Partnern bereits an neuen, sehr komplexen Datenbeständen. Die Aufgabe, die faktische Anonymität sicherzustellen und gleichzeitig das Analysepotenzial möglichst weitgehend zu erhalten, ist bereits bei wirtschaftsstatistischen Querschnitts- und Paneldaten komplex. In der Regel lassen sich größere Einschnitte in das Analysepotenzial nicht vermeiden. Es ist absehbar, dass die neuen im Querschnitt und über die Zeit verknüpften Datenbestände, die zurzeit bei den genannten Projekten entstehen, durch die notwendigen Anonymisierungsmaßnahmen für die Erstellung eines Scientific-Use-Files zur Off-Site-Nutzung ihr immenses Informationspotenzial wieder verlieren würden. Dies dürfte zu einer noch stärkeren Nachfrage im Bereich der aufwendigen On-Site-Nutzung führen. Die Forschungsdatenzentren der verschiedenen Datenproduzenten sind aber schon jetzt durch die intensive Nutzung, insbesondere im Bereich der kontrollierten Datenfernverarbeitung, stark belastet.

Hier soll das Projekt infinitE Abhilfe schaffen.

2 Projektarbeiten

Das im Folgenden beschriebene Forschungsvorhaben stellt einen Meilenstein auf dem Weg zum automatisierten Fernrechnen dar. Die Idealsituation hierbei wäre ein autorisierter Zugriff auf die Daten der amtlichen Statistik rund um die Uhr von jedem beliebigen Rechner aus, bei dem die Forscherinnen und Forscher das gewünschte Ergebnis nach sofortiger vollautomatischer Prüfung auf Geheimhaltung in Echtzeit erhielten.

Eine solche „Remote-Access“-Anwendung (oder auch echtes Fernrechnen), die vollkommen automatisiert ist und bei der keine manuellen Eingriffe mehr erfolgen müssen, ist jedoch eine Zukunftsvision. Vorher muss noch eine Reihe entscheidender methodischer, technischer und rechtlicher Fragen einer solchen automatisierten Datenfernverarbeitung geklärt werden. In Ländern mit vergleichbaren rechtlichen Rahmenbedingungen wie in Deutschland ist ein solcher Online-Zugriff auf geschützte Mikrodaten[4]) – das bedeutet im Grunde dieses Verfahren – bisher noch nicht realisiert worden.

Unter dem Begriff „Remote-Access“ existiert derzeit eine Vielzahl von Systemen. Dabei kann die konkrete technische Ausgestaltung der Systeme jeweils sehr unterschiedlich sein. Bisher sind die meisten Ansätze nur als sogenanntes „Remote Execution“ zu bezeichnen, da sie – gegenüber einem echten „Remote Access“ – immer noch manuelle Eingriffe an mehreren Stellen erfordern.

Mit LISSY in Luxemburg[5]) und den Verfahren des niederländischen[6]) und dänischen[7]) nationalen statistischen Amtes liegen zwar erste Anwendungen vor, aber all diese Verfahren stellen noch keine vollständig automatisierten Zugriffsrouti-

2) Siehe Bender, S./Rosemann, M./Zühlke, S./Zwick, M. (Hrsg.): „Betriebs- und Unternehmensdaten im Längsschnitt – Neue Datenangebote und ihre Forschungspotenziale“ in AStA – Wirtschafts- und Sozialstatistisches Archiv, Band 2, Heft 3, 2008, sowie Zwick, M.: „Forschungsdatenzentren – Nutzen und Kosten einer informationellen Infrastruktur für Wissenschaft, Politik und Datenproduzenten“ in WiSta 12/2006, S. 1233 ff.

3) *www.forschungsdatenzentrum.de/afid.asp*, sowie Konold, M./L'Assainato, S.: “Matching Business Data from Different Sources: The Case of the KombiFiD-Project in Germany”, New Techniques and Technologies for Statistics, Brüssel 2009.

4) *http://neon.vb.cbs.nl/casc/.. %5Ccasc %5Cglossary.htm#R* (Stand: 22. Juni 2009).

5) Siehe hierzu z. B. Coder, J./Cigrang, M.: “LISSY Remote Access System”, Working paper No. 7 of the Joint ECE/Eurostat work session on statistical data confidentiality, Luxemburg 2003.

6) Siehe Hundepool, A./de Wolf, P.-P.: “OnSite@Home: Remote Access at Statistics Netherlands”, UNECE/Eurostat Work Session on Statistical Data Confidentiality, Genf 2005.

7) Siehe Borchsenius, L.: “New Developments in the Danish system for access to microdata”, Monographs of Official Statistics, Luxemburg 2005.

nen zur Verfügung oder sind – wie zum Beispiel bei LISSY – auf nur einen Datensatz beschränkt. In Schweden existiert mit MONA[8]) eine praktikable Fernrechenlösung, die aber nur durch die besondere rechtliche Situation bei der Anwendung des Forschungsdatengeheimnisses bei der Datennutzung in Schweden möglich ist. In Deutschland liegt mit SAM[9]) eine erste technische Lösung vor. Auch JoSuA[10]) könnte zu einer solchen Anwendung ausgebaut werden.

Ein sinnvolles Zwischenziel für Deutschland wäre mittelfristig eine „Remote-Access"-Lösung wie in den Niederlanden, in Dänemark oder beim NORC[11]). Das Projekt „Eine informationelle Infrastruktur für das E-Science Age" leistet hier Grundlagenforschung, um erste Schritte in diese Richtung unternehmen zu können.

Mit infinitE werden folgende Ziele verfolgt:

1) Es sollen grundsätzliche Strategien für das Erstellen anonymisierter Datenstrukturfiles, die es erlauben, einen Programmablauf auf syntaktische wie semantische Fehler zu überprüfen, entwickelt werden. Die Datenstrukturfiles, die die Nutzer der kontrollierten Datenfernverarbeitung derzeit erhalten, erlauben nur eine syntaktische Überprüfung. Für das Erstellen solcher Datenstrukturfiles mit erweiterten Funktionen kommen insbesondere die datenverändernden Verfahren der multiplikativen stochastischen Überlagerung, der mehrdimensionalen Mikroaggregation und der multiplen Imputation in Betracht.[12])

2) Es sollen Verfahren der standardisierten und vollständig automatisierten Ergebniskontrolle entwickelt und bewertet werden. Ergebnisse der kontrollierten Datenfernverarbeitung und der Arbeiten am Gastwissenschaftlerarbeitsplatz müssen vor Freigabe auf Geheimhaltung überprüft werden. Diese Überprüfung gestaltet sich bei komplexen Tabellen und umfangreichen Schätzoutputs außerordentlich schwierig und daher zeit- und personalaufwendig. Automatisierte maschinelle Verfahren konnten bisher nur teilweise für standardisierte, regelmäßig vorliegende Ergebnisse entwickelt werden. Für die flexiblen Auswertungen in den Forschungsdatenzentren reichen die bisher (auch international) entwickelten Verfahren bei weitem nicht aus, sodass hier erheblicher Forschungsbedarf besteht. Im Projekt soll hierbei die Fragestellung noch erweitert und ein systematischer Vergleich von datenseitiger und ergebnisseitiger Sicherung des Schutzes der Merkmalsträger im Hinblick auf das Analysepotenzial vorgenommen werden. Ziel ist es, Verfahren zu entwickeln, die es dem Nutzer erlauben, vor dem Analyselauf zu entscheiden, ob eine Analyse mit anonymisierten Daten und ohne Einschränkung bei den Ergebnissen oder auf der Grundlage des originalen Datensatzes bei Einschränkung der Ergebnisfreigabe durchgeführt werden soll. Auch Mischformen könnten möglich sein.

2.1 Erstellen von Datenstrukturfiles

Das erste Ziel des Projektes ist es, die Projektdaten als sogenannte Datenstrukturfiles aufzubereiten. Diese anonymisierten Datensätze, die die Struktur des Originaldatensatzes aufweisen, werden dem Forscher bzw. der Forscherin zugesandt, nachdem sie einen Nutzungsantrag gestellt haben. Damit entwickeln diese ihren Programmcode zur Auswertung und senden ihn an das betreffende Forschungsdatenzentrum. Dieser Programmcode wird dann im Forschungsdatenzentrum zunächst auf kritische Inhalte/Befehle geprüft und auf die Originaldaten angewendet und der Output – nach einer Prüfung auf Datensicherheit und Geheimhaltung – an den bzw. die Forscher/-in zurückgesendet.

Bisher bestehen die Datenstrukturfiles aus einer kleinen Stichprobe des Originalmaterials, auf die zusätzliche Anonymisierungsmaßnahmen angewendet werden, oder aus zufällig generierten Werten im Wertebereich des Datensatzes. Bei beiden Vorgehensweisen bleiben die Merkmale zwar erhalten, ihre Ausprägungen und die Abhängigkeitsstruktur (Filter, Varianz-Kovarianz-Matrix) zu anderen Merkmalen werden dabei jedoch komplett zerstört. Somit kann ein Forscher bzw. eine Forscherin zwar prüfen, ob sein bzw. ihr Programm lauffähig ist, bekommt aber keine Hinweise, ob die inhaltliche Fragestellung adäquat umgesetzt wurde. Daher können die Auswertungsprogramme der Wissenschaftler/-innen oft nicht eins zu eins für die spätere Anwendung auf die Originaldaten übernommen werden. Meist sind weitere Anpassungsarbeiten durch die Wissenschaftler/-innen und die Mitarbeiter/-innen der Forschungsdatenzentren nötig.

Bei komplexeren Daten, wie den Linked-employer-employee-Datensätzen [Gehalts- und Lohnstrukturerhebung (1995, 2001), Verdienststrukturerhebung (2006)], sind die Datenstrukturfiles in ihrer derzeitigen Form wenig hilfreich, da sie in der Praxis zu großen Anpassungsproblemen und hohem Abstimmungsbedarf zwischen den externen Wissenschaftlern bzw. Wissenschaftlerinnen und den Forschungsdatenzentren führen. Mit diesen Datenstrukturfiles sind zum Beispiel keine Konsistenzprüfungen dahingehend möglich, ob das vom Forscher bzw. von der Forscherin entwickelte Auswertungsprogramm richtig ist und fehlerfrei auf die Originaldaten angewendet werden kann. Um die Programme für das Fernrechnen geeignet anpassen zu können, muss die empirisch arbeitende Wissenschaft neben univariaten Berechnungen auch vermehrt Programme testen können, die auf multivariate Zusammenhänge abstellen. Eine Auswertung mit multivariaten Analysen ist mit den bisherigen Datenstrukturfiles schwierig, da hier die Kovarianzen nicht erhalten bleiben und auch von Seiten der Forscher/-innen

8) MONA: Microdata ON-line Access, Statistics Sweden.

9) Siehe Heitzig, J.: „Wissenschaftsserver zur Auswertung von Mikrodaten", Wiesbaden 2006.

10) Das Datenzentrum des Forschungsinstituts zur Zukunft der Arbeit in Bonn (IdZA) hat eine Anwendung entwickelt, die es externen Forschern erlaubt, Auswertungen von Einzeldaten über das Internet zu starten. Diese Anwendung JoSuA ist zum einen nutzerfreundlich, da der Forscher den Status seines Auftrages von seinem Arbeitsplatz aus überwachen kann, zum anderen erleichtert sie die Arbeiten des IdZA, da die Programme nicht mehr von Hand gestartet werden müssen.

11) NORC: National Opinion Research Center at the University of Chicago.

12) Zu den Verfahren siehe Statistisches Bundesamt (Hrsg.): „Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten", Band 4 der Schriftenreihe „Statistik und Wissenschaft", Wiesbaden 2005.

immer wieder Modellanpassungen bezüglich der Originaldaten vorgenommen werden müssen. Das führt vor allem dann zu Problemen, wenn die volle Anzahl der Beobachtungen benötigt wird (z. B. bei Auswertungen mit mehreren Wellen). Hinzu kommt, dass nicht immer alle logischen Restriktionen korrekt abgebildet werden. Letztlich führt das dazu, dass die entwickelten Programmcodes mehrere Male zwischen den Wissenschaftlern und den Mitarbeitern des jeweils beteiligten Forschungsdatenzentrums hin und her geschickt werden müssen, bis das erwünschte Ergebnis vorliegt. Das bedeutet aber auch, dass alle durchgeführten Analysen jeweils wieder überprüft werden müssen, um sicherzustellen, dass bei der Bereitstellung der Ergebnisse der Datenschutz nicht verletzt wird.

Ein weiterer Vorteil von semantisch und syntaktisch fehlerfreien Datenstrukturfiles besteht darin, dass die Wissenschaftler/-innen die Anzahl und den Umfang ihrer Ergebnistabellen genauer bestimmen können und am eigenen Arbeitsplatz die Auswertungen so lange anpassen können, bis die gewünschten Ergebnistabellen produziert werden. Somit wird der Aufwand reduziert, Tabellen auf Geheimhaltung zu prüfen, die eventuell nicht in eine Publikation eingehen. Die Datenstrukturfiles sollen für die gängigen Statistikprogramme, wie STATA, SPSS oder SAS, erstellt werden.

Es wurden bereits datenverändernde Anonymisierungsverfahren entwickelt bzw. auf die Anforderungen der Wirtschaftsstatistiken der deutschen statistischen Ämter bzw. der Bundesagentur für Arbeit angepasst.[13]) Ob die jetzt zu entwickelnden Datenstrukturfiles absolut anonymisierte Public-Use-Files oder aber faktisch anonym sind, muss der Projektverlauf zeigen. Ein Public-Use-File hätte den erheblichen Vorteil, dass es unabhängig von der Antragstellung frei von jedermann von einer Webseite geladen werden kann und somit auch für die ausländische Wissenschaft verfügbar ist. Bei einem faktisch anonymen Datenstrukturfile muss erst über den Nutzungsantrag abgestimmt werden und ein Vertrag zur Datennutzung unterzeichnet worden sein, bevor das Datenmaterial an die Forscher versendet werden kann. Wenn ein Public-Use-File die Datenstruktur ausreichend widerspiegelte, wäre es gegenüber einem faktisch anonymen Datensatz zu bevorzugen, da das Auswertungsprogramm ohnehin auf die Originaldaten angewendet wird, unabhängig davon, ob das Datenstrukturfile absolut oder faktisch anonym ist. Ob die Datenstruktur bereits ausreichend mit einem absolut anonymen Datensatz oder letztendlich nur mit einem faktisch anonymen Datensatz abgebildet werden kann, ist nur im Rahmen der Projektarbeit zu klären.

Bis eine echte „Remote-Access"-Lösung in Form einer funktionierenden technischen Infrastruktur umgesetzt wird, ist es unumgänglich, Datenstrukturfiles zu verwenden, um dem Forscher/der Forscherin eine gewisse Flexibilität bei der Bearbeitung der Auswertungsprogramme zu ermöglichen und diese unabhängig von Zeit und Ort entwickeln zu können.

Auch für ein funktionsfähiges „Remote-Access" sind die entwickelten Datenstrukturfiles weiterhin notwendig. Da die Ansicht der Originaldaten am Bildschirm des Nutzers eine datenschutzrechtlich nicht erlaubte Übermittlung von originalen Daten darstellen würde, könnte es beim „Remote-Access" sinnvoll sein, dem Nutzer stattdessen die Ansicht der Einzeldaten aus den Datenstrukturfiles zu präsentieren. Bei der Berechnung der Analysen wird jedoch auf die Originaldaten zugegriffen. Die Daten selbst, das heißt sowohl die Datenstrukturfiles als auch die Originaldaten, verbleiben auf den Servern in den geschützten Räumen der amtlichen Statistik. Insofern stellen die Datenstrukturfiles sowohl eine Zwischenlösung bis zur Entwicklung des echten „Remote-Access" dar als auch ein Endprodukt, welches als wichtiger Teil in eine hardwaretechnische Zugangslösung implementiert werden kann.

2.2 Ergebnisseitige Geheimhaltung

Um der Wissenschaft einen Datenzugang zu vertraulichen Daten zu ermöglichen, stellt das Verfahren des „Remote-Access" eine gute Lösung dar. Dabei können externe Wissenschaftler an ihrem eigenen Computer Analysen über einen Remote Server durchführen und bekommen die Ergebnisse in Echtzeit angezeigt. Ein Problem ist hierbei die Frage, wie ein Wissenschaftler bei einer möglichen Datenschutzverletzung kontrolliert bzw. wie diese Datenschutzverletzung mit Sicherheit verhindert werden kann. Der sicherste Weg ist, die Ergebnisse bereits vor der Übermittlung in Echtzeit auf potenzielle Datenschutzrisiken zu überprüfen.

Bislang werden die Ergebnisse der kontrollierten Datenfernverarbeitung und der Arbeiten am Gastwissenschaftlerarbeitsplatz vor Freigabe manuell auf Geheimhaltung überprüft. Diese Überprüfung gestaltet sich bei komplexen Tabellen und umfangreichen Schätzoutputs außerordentlich schwierig und ist sehr zeit- und personalaufwendig.

In der amtlichen Statistik wird, insbesondere bei wirtschaftsstatistischen Daten, üblicherweise ein Teil der Felder unterdrückt, um die Tabellengeheimhaltung zu sichern. Da die von der amtlichen Statistik verbreiteten Tabellendaten durch Zwischen- und Randsummen linear miteinander verknüpft, also additiv sind, müssen zusätzliche Felder gesperrt werden (Sekundärsperrungen), um zu verhindern, dass die primär geheimen Felder sich durch Differenzbildung aufdecken lassen. Die Bestimmung geeigneter Zellsperrungsverfahren – die den durch die Sperrung entstehenden Informationsverlust minimieren – ist ein komplexes lineares Optimierungsproblem. Eine Übersicht über die gebräuchlichen Standardverfahren gibt zum Beispiel Gießing.[14])

Oft ist ein Teil der Felder aus einer Tabelle identisch mit Feldern einer anderen Tabelle. In solchen Fällen muss die Auswahl der Sekundärsperrungen tabellenübergreifend koordiniert werden, um zu verhindern, dass Nutzer durch einen Abgleich solcher überlappenden Tabellen gesperrte Felder aufdecken können.

13) Siehe Fußnote 12 und Bender, S., u. a., Fußnote 2.

14) Gießing, S.: „Statistische Geheimhaltung in Tabellen" in Statistisches Bundesamt (Hrsg.): „Methoden zur Sicherung der statistischen Geheimhaltung", Band 31 der Schriftenreihe „Forum der Bundesstatistik", Wiesbaden 1999, S. 6 ff.

Eine Ergebnisüberprüfung der am Gastwissenschaftlerarbeitsplatz bzw. im Rahmen der kontrollierten Datenfernverarbeitung erzeugten Tabellen ist u. a. erforderlich, um zu gewährleisten, dass die von den Nutzern abschließend publizierten Ergebnisse nicht die Sekundärsperrungen aus den eigenen Publikationen der statistischen Ämter aufdecken. Dazu müssen die Sekundärsperrungen zwischen der Standardpublikation und den Tabellendaten von Nutzern koordiniert werden. Geeignete maschinelle Verfahren hierzu stehen nicht zur Verfügung. Wegen der nachrangigen Behandlung der Nutzertabellen (gegenüber den Standardpublikationen) ist mit erheblicher Beeinträchtigung der Ergebnisqualität durch massive Sekundärsperrungen zu rechnen.

Unter anderem aus diesen Gründen wurden in den letzten Jahren in der Literatur zunehmend datenverändernde Verfahren als Alternative bzw. Ergänzung zu den Sperrverfahren vorgeschlagen. Die meisten der vorgeschlagenen Verfahren verändern die Daten auf der Ebene der Aggregate. Allerdings gibt es auch Verfahren, die die Einzeldaten verändern. Weil auch bei diesen auf Einzeldaten anzuwendenden Verfahren im Vordergrund steht, die Ergebnisqualität bestimmter Tabellen zu erhalten, gelten sie als Tabellengeheimhaltungsverfahren und nicht als Verfahren zur Anonymisierung von Mikrodaten.

Während der Datenschutz bei Tabellenoutputs schon lange Thema in den Statistischen Ämtern des Bundes und der Länder ist, existiert bisher keine systematische Untersuchung der Datenschutzproblematik bei Schätzoutputs und nichtlinearen Analysen. Hierfür können die Untersuchungen von Gomatam, Karr, Reiter und Sanil als Grundlage dienen.[15])

Ein Ansatz zur ergebnisseitigen Geheimhaltung stammt von Heitzig.[16]) Er überträgt die Idee des Jackknife-Ansatzes zur Schätzung von Standardfehlern auf die Geheimhaltung von Mikrodaten. Grundgedanke des Verfahrens ist es, jeweils einen Wert der zugrunde liegenden Originaldaten mit einem zufälligen Wert (aus einer Verteilung mit einer ausreichenden Streuung) zu ersetzen. Die Analysen werden dann mit allen veränderten Datensätzen nacheinander durchgeführt und es wird das Intervall der Ergebniswerte veröffentlicht.

Es sei an dieser Stelle darauf hingewiesen, dass es sich bei dem hier beschriebenen Ansatz nur um die theoretische Basis des Verfahrens handelt. Würde das Verfahren in dieser Form implementiert, wäre in der Praxis der Rechenaufwand häufig zu hoch. Im Hinblick auf eine praktikablere Lösung ist es gelungen, für eine ganze Reihe von Analyse„klassen“ mathematische Ansätze zu finden, die die Intervalle mit geringem Rechenaufwand berechnen bzw. approximieren. Der eigentliche Aufwand besteht darin, diese Methode weiterzuentwickeln, das heißt für weitere Analyse„klassen“ verfügbar zu machen. Dies ist jedoch nicht Gegenstand des Projektes. Mit dem beantragten Projekt soll lediglich herausgefunden werden, ob sich die Methode für die bereits verfügbaren Analyse„klassen“ in der Praxis bewährt. Nur dann ist es sinnvoll, diese Methode weiterzuentwickeln.

Zur ergebnisseitigen Geheimhaltung existieren weitere Ansätze. So verfolgt das US Census Bureau zum Beispiel einen Ansatz, bei dem die automatische Ergebniskontrolle nicht durch ein Vergröbern der Ergebnisse erreicht wird, sondern durch Nutzungsbeschränkungen, die durch das System kontrolliert werden.[17]) Diese Nutzungsbeschränkungen betreffen zum einen die den Nutzern für Analysen zur Verfügung gestellten Daten (z. B. Vergröberung der Einzeldaten, etwa indem Kategorien bestimmter Merkmale zusammengefasst werden), zum anderen die mithilfe des Systems durchführbaren Untersuchungen (auf einen begrenzten Katalog von Analysen).

Weiterhin sind aus der Literatur Ansätze bekannt, die sich speziell mit der Problematik der Geheimhaltung von Residuen von Regressionen beschäftigen.[18]) Während Reiter u. a. vorschlägt, den Nutzern anstelle der Original-Residuen synthetisch generierte Residuen zur Verfügung zu stellen, beruht der Ansatz von Sparks u. a. auf der Erzeugung von Box-Plots für die Residuen.

Zusätzlich muss untersucht werden, wie eine Kombination von Tabellen- und Schätzoutputs von multivariaten Analysen aus derselben Teilpopulation zu behandeln ist und ob hier erhöhte Reidentifikationsrisiken auftreten.

3 Zusammenfassung und Ausblick

Das hier beschriebene Forschungsvorhaben stellt eine wichtige Brücke dar zwischen den bisher erreichten Verbesserungen bei den Datenzugangswegen für die Wissenschaft in den vergangenen Jahren und den von den Forschungsdatenzentren für die Zukunft geplanten Entwicklungen. Es ist daher als ein wesentlicher Meilenstein auf dem Weg zum echten Fernrechnen anzusehen.

Angesichts der steigenden Nachfrage nach Mikrodaten in Deutschland, insbesondere in Form einer On-Site-Nutzung, wird eine manuelle Durchführung der Aufträge zum Fernrechnen aus Kapazitätsgründen immer schwieriger. Mit der Ausweitung der Nachfrage auf Ergebnisse weiterer Statistiken wird auch die zeitnahe Bereitstellung von Scientific-Use-Files zusätzlicher Erhebungen problematisch.

Der nächste (auch parallel mögliche) Schritt für Deutschland könnte eine „FDZ in FDZ“-Lösung sein, bei der über einen „Remote-Zugriff“ die Daten eines Forschungsdatenzentrums in einem anderen Forschungsdatenzentrum bearbeitet werden könnten. Dies könnte als Testimplementation für einen späteren echten „Remote-Access“ dienen und würde eine Aufgabenverlagerung in den Forschungsdatenzentren hin

15) Gomatam, S./Karr, A. F./Reiter, J. P./Sanil, A.: “Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access servers”, Statistical Science 20, S. 163, 2005.

16) Heitzig, J.: “The ‘Jackknife’ Method: Confidentiality Protection for Complex Statistical Analyses”, UNECE/Eurostat Work Session on Statistical Data Confidentiality, Genf 2005 (*www.unece.org/stats/documents/ece/ces/ge.46/2005/wp.39.e.pdf.*; Stand: 22. Juni 2009).

17) Siehe Zayatz, L.: “New Implementations of Noise for Tabular Magnitude Data, Synthetic Tabular Frequency and Microdata, and a Remote Microdata Analysis System”, Proceedings of the Joint UNECE/Eurostat work session on statistical data confidentiality, Manchester 2007.

18) Siehe Reiter, J. P.: “New Approaches to Data Dissemination: A Glimpse into the Future (?)”, Chance, Vol. 17, Nr. 3, 2004, S. 12 ff.

zu mehr Datenerschließung, Datendokumentation, Internationalisierung und weg von einer aufwendigen Gästebetreuung erlauben.

Langfristig erscheint das echte Fernrechnen sowohl national als auch international als einzig praktikable Lösung. Ein einmal entwickeltes Verfahren könnte zudem sehr schnell auf andere Erhebungen übertragen werden und somit eine „just in time"-Bereitstellung der Daten ermöglichen. Die Informationstechnik ist mittlerweile so weit entwickelt, dass ein Online-Zugang von überall aus möglich ist bzw. bald mit der entsprechenden Bandbreite möglich sein wird.

Das echte Fernrechnen erlaubt eine von Zeit und Ort unabhängige flexible Bearbeitung der Daten durch die Wissenschaftlerinnen und Wissenschaftler und hat den Vorteil, dass die Daten in den geschützten Räumen (und auf den geschützten Servern) der amtlichen Statistik verbleiben. Weiterhin wird diese Form des Datenzugangs die Vernetzung der Forscher untereinander und die wissenschaftliche Transparenz fördern, da alle Wissenschaftler jederzeit Zugang zu den Daten haben und Ergebnisse replizieren können.

Im Zeitalter der elektronisch basierten Wissenschaft (E-Science) sollte ferner darauf geachtet werden, dass die Entwicklung der informationellen Infrastruktur nicht hinter den technischen Möglichkeiten zurückbleibt, die bei Weitem noch nicht ausgeschöpft sind und auch in Zukunft noch weiteres Entwicklungspotenzial bereithalten. Dessen ungeachtet sind vor Einführung eines echten Fernrechnens noch erhebliche rechtliche Fragen zu klären. Bei anderen rechtlichen Rahmenbedingungen sind einfachere technische Lösungen eines „Remote-Access", wie zum Beispiel in Schweden, denkbar. Die rechtlichen Aspekte werden derzeit von der Arbeitsgruppe „Future Data Access" des Rates für Sozial- und Wirtschaftsdaten geprüft.

Mit dem Projekt infinitE wird zum einen die dringend notwendige methodische Grundlagenforschung für ein vollautomatisiertes Fernrechnen erfolgen. Zum anderen werden schon während des Projektverlaufs die Mitarbeiter und Mitarbeiterinnen in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder entlastet. Dies wird dadurch möglich, dass Datenstrukturfiles und Werkzeuge für die Erstellung von Datenstrukturfiles für beliebige Statistiken für die kontrollierte Datenfernverarbeitung und Richtlinien sowie Instrumente für eine Kategorisierung und automatisierte Geheimhaltungsprüfung erstellt werden. Dabei werden die in anderen Ländern bereits entwickelten Verfahren berücksichtigt und die in internationalen Arbeitsgruppen bereits vorhandenen Erfahrungen genutzt. Profitieren wird das Vorhaben darüber hinaus von den in den letzten Jahren durchgeführten methodisch orientierten Projekten im Bereich der Anonymisierung wirtschaftsstatistischer Einzel- und Paneldaten.

Auszug aus Wirtschaft und Statistik

Herausgeber: Statistisches Bundesamt, Wiesbaden

Schriftleitung: Roderich Egeler
Präsident des Statistischen Bundesamtes
Verantwortlich für den Inhalt:
Brigitte Reimann,
65180 Wiesbaden
- Telefon: +49 (0) 6 11/75 20 86
- E-Mail: wirtschaft-und-statistik@destatis.de

Vertriebspartner: SFG Servicecenter Fachverlage
Part of the Elsevier Group
Postfach 43 43
72774 Reutlingen
Telefon: +49 (0) 70 71/93 53 50
Telefax: +49 (0) 70 71/93 53 35
E-Mail: destatis@s-f-g.com

Erscheinungsfolge: monatlich

Allgemeine Informationen über das Statistische Bundesamt und sein Datenangebot erhalten Sie:
- im Internet: www.destatis.de

oder bei unserem Informationsservice
65180 Wiesbaden

- Telefon: +49 (0) 6 11/75 24 05
- Telefax: +49 (0) 6 11/75 33 30
- www.destatis.de/kontakt