Forschungsdatenzentrum
der Bundesagentur für Arbeit
im Institut für Arbeitsmarkt-
und Berufsforschung

FDZ-Methodenreport

06/2012
DE

Methodische Aspekte zu Arbeitsmarktdaten

Datenschutz am Forschungsdatenzentrum

Aktualisierte Fassung vom 21.12.2012

Daniela Hochfellner,
Dana Müller,
Alexandra Schmucker,
Elisabeth Roß

Bundesagentur für Arbeit

Datenschutz am Forschungsdatenzentrum

Daniela Hochfellner (IAB)
Dana Müller (IAB)
Alexandra Schmucker (IAB)
Elisabeth Roß (IAB)

Aktualisierte Fassung vom 21.12.2012

Die FDZ-Methodenreporte befassen sich mit den methodischen Aspekten der Daten des FDZ und helfen somit Nutzerinnen und Nutzern bei der Analyse der Daten. Nutzerinnen und Nutzer können hierzu in dieser Reihe zitationsfähig publizieren und stellen sich der öffentlichen Diskussion.

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Inhaltsverzeichnis

Zusammenfassung

Forschungsdaten, die aus dem Bereich der Bundesagentur für Arbeit (BA) bzw. aus den Befragungen des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) stammen, sind für Wissenschaft und Politikberatung von zunehmend hoher Bedeutung. Zahlreiche Forschungsfragen aus der Arbeitsmarkt- und Berufsforschung lassen sich mit diesen Daten beantworten. Es handelt sich um Sozialdaten, die den Datenschutzbestimmungen des Sozialgesetzbuches X (SGB X) bzw. den Regeln der statistischen Geheimhaltung unterliegen. Das SGB X und das SGB III räumen unter bestimmten Voraussetzungen Nutzungsrechte ein. Auch Forschungsvorhaben externer Forschungsinstitute, die diese aus eigenem Antrieb oder z.B. im Auftrag des Bundesministeriums für Arbeit und Soziales (BMAS) durchführen, profitieren davon. Um der Wissenschaft Sozialdaten leichter zugänglich zu machen, wurde das Forschungsdatenzentrum der BA im IAB (FDZ) geschaffen. Daten unterschiedlicher Anonymisierungsgrade stehen dort datenschutzgerecht über standardisierte und transparente Wege zur Verfügung. Ziel dieses Artikels ist es, das Spannungsverhältnis zwischen Forschungsinteressen einerseits und Datenschutz andererseits sowie die praktische Umsetzung der ausgleichenden Maßnahmen darzustellen.

Abstract

Research data of the Federal Employment Agency as well as surveys of the Institute for Employment Research are highly relevant for the scientific community and policy consulting. These data help to find answers to various research questions regarding employment and occupational research. The legal basis for data access is mainly Section 67 of the German Social Code Book X (SGB X). Since the establishment of the Research Data Centre (FDZ) of the Federal Employment Agency (BA) in the Institute of Employment Research (IAB) social data of the BA and the IAB are accessable for researchers using standardised and transparent principles. The remainder of this paper is to discuss the trade-off between the capability of research interests and data protection as well as the satisfaction of these demands.

Keywords: Datenschutzbestimmungen, Sozialdaten, Anonymisierung

Danksagung: Wir danken unseren Kolleginnen und Kollegen am FDZ für ihre hilfreichen Kommentare und Anregungen, sowie Johanna Eberle für die technische Unterstützung. Darüber hinaus danken wir Felix Ritchie (UK Office for National Statistics) für zahlreiche Anregungen und Informationen zum Thema.

1 Einleitung

Die Bedeutung von Forschungsdaten für die Wissenschaft und Politikberatung ist unbestreitbar. Die Nachfrage nach umfassenden Datensätzen, angereichert mit Zusatzinformationen aus anderen Datenquellen, wächst stetig. Insbesondere prozessproduzierte Daten werden für die Sozialforschung aufgrund ihrer Vorzüge immer attraktiver. Im Vergleich zu Befragungsdaten handelt es sich bei den administrativen Daten um Vollerhebungen, in denen hoch zuverlässige Informationen meist über lange Zeiträume gesammelt werden. Somit treten die üblichen Probleme von Befragungsdaten wie Antwortverweigerungen, Panelausfälle, Erinnerungslücken und -fehler nicht auf. Das Forschungsdatenzentrum (FDZ) der Bundesagentur für Arbeit (BA) im Institut für Arbeitsmarkt- und Berufsforschung (IAB) stellt umfangreiche administrative Sozialdaten, die für Analysen im Bereich der Arbeitsmarktforschung besonders geeignet sind, für Forschungszwecke zur Verfügung[1]. Es handelt sich um den Datenbestand aus den Verwaltungsprozessen der BA, die Daten der kommunalen Träger der Grundsicherung für erwerbsfähige Hilfebedürftige und die Daten aus den Meldeverfahren der Sozialversicherung. Alle diese Daten werden über die Verfahren der Statistik der BA zusammengeführt, konsolidiert und aufbereitet, im IAB weiter verarbeitet und als standardisierte Forschungsdatensätze im FDZ bereit gestellt.

Sozialdaten unterliegen jedoch dem besonderen Schutz des Sozialgeheimnisses (§35 Abs. 1 Satz 1 SGB I), da es es dabei um Pflichtangaben handelt, die für die Berechnung von Beitragshöhen und späteren Ansprüchen gegenüber den Sozialversicherungen (z.B. Rentenversicherung) erforderlich sind. Juristisch gesehen stehen sich dabei zwei widerstreitende Verfassungsgrundsätze gegenüber: auf der einen Seite das Recht auf informationelle Selbstbestimmung, auf der anderen Seite die im Grundgesetz verankerte Wissenschafts- und Forschungsfreiheit (Art. 5 GG). Um hier einen Ausgleich zu finden, wurden Rechtsgrundlagen geschaffen, die eine wissenschaftliche Nutzung der Daten bei gleichzeitiger Einhaltung des Datenschutzes ermöglichen (§75 SGB X, §282 Abs. 7 SGB III). In der Praxis ergibt sich daraus jedoch ein Zielkonflikt zwischen einem möglichst großen Analysepotenzial der Daten und maximalem Datenschutz: Je mehr Informationen vorhanden sind, desto größer ist das Analysepotential. Anderseits steigt mit dem Informationsgehalt das Risiko der Deanonymisierung. Damit ein standardisierter Datenzugang für die Forschung unter Einhaltung des Datenschutzes ermöglicht wird, wurde das FDZ auf Empfehlung der Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI) 2004 eingerichtet.

Im Folgenden wird dargestellt, wie das FDZ in der Praxis den beschriebenen Zielkonflikt löst. Hierzu werden vom FDZ von der Aufbereitung standardisierter Daten über die Zugangswege bis hin zur Ergebniskontrolle unterschiedliche Methoden angewandt.

[1] Neben den Verwaltungsdaten stehen in beträchtlichem Umfang Daten aus groß angelegten Befragungen zur Verfügung, die zum Teil mit Informationen aus den prozessproduzierten Daten angereichert sind. Für die Nutzung dieser Daten gelten die gleichen Bedingungen wie bei den Sozialdaten.

2 Warum Datenschutz?

Zunächst stellt sich die Frage, warum die Daten des FDZ überhaupt schützenswert sind. Das Gesetz räumt der Forschung Nutzungsrechte an Sozialdaten nur unter der Auflage von gesetzlich normierten Datenschutzvorgaben ein (§§ 67 ff SGB X). Nur durch Einhaltung dieser Vorgaben ist die Forschung mit den sensiblen Sozialdaten überhaupt möglich. Der Datenschutz ist somit untrennbar mit der Einräumung von Nutzungserlaubnissen verbunden. Im Folgenden wird auf die gesetzliche Grundlage und die damit verbundene Aufgabe des FDZ eingegangen.

2.1 Gesetzlicher Hintergrund

Seit der Entscheidung des Bundesverfassungsgerichts zum Volkszählungsurteil[2] ist die verfassungsrechtliche Grundlage des Datenschutzes im Recht auf informationelle Selbstbestimmung anerkannt. Danach darf grundsätzlich jeder selbst über die Verwendung und Preisgabe seiner persönlichen Daten bestimmen. Einschränkungen hierzu sind nur auf einer gesetzlichen Grundlage erlaubt. Die Regelungen zum Datenschutz sichern dieses Recht, indem sie bestimmte Eingriffe erlauben und gleichzeitig Grenzen setzen. Hervorzuheben ist der zentrale Begriff „personenbezogene Daten“, der jede Information über eine bestimmte oder bestimmbare natürliche Person (Betroffener) umfasst. Der Sozialdatenschutz des 2. Kapitels im SGB X ist enger gefasst als der Schutz personenbezogener Daten nach dem Bundesdatenschutzgesetz (BDSG). Grund hierfür ist der Umstand, dass Sozialdaten[3] bei den Betroffenen nicht auf freiwilliger Basis erhoben werden, sondern zwangsweise. Beispielsweise erhält derjenige, der Arbeitslosengeld beantragt, Leistungen nur, wenn er Angaben über seine Person macht. Die Betroffenen sind gesetzlich verpflichtet, die Verarbeitung ihrer personenbezogenen Daten hinzunehmen. Auch Betriebe sind gezwungen – im Rahmen des Meldeverfahrens zur Sozialversicherung – Informationen über sich preiszugeben.

Sozialdaten unterliegen gemäß § 35 SGB I dem Sozialgeheimnis[4]. Es gilt das Verbot mit Erlaubnisvorbehalt. Das bedeutet, dass alles, was im Umgang mit Sozialdaten nicht ausdrücklich gesetzlich erlaubt ist, verboten ist und alles Verbotene ausnahmslos den Tatbestand einer Ordnungswidrigkeit oder einer Straftat erfüllt. Jede Nutzung, jede Speicherung, jede Weitergabe und jedes Offenbaren bedarf also einer gesetzlichen Rechtfertigung. Im Sozialgesetzbuch gehören zu den schutzbedürftigen Daten nicht nur die personenbezo-

[2] Das Urteil des Bundesverfassungsgerichts 1983 zur Volkszählung lautet („Volkszählungsurteil“ BVerfGE 65, 1): „ (. . .) Freie Entfaltung der Persönlichkeit setzt unter den modernen Bedingungen der Datenverarbeitung den Schutz des Einzelnen gegen unbegrenzte Erhebung, Speicherung, Verwendung und Weitergabe seiner persönlichen Daten voraus. Dieser Schutz ist daher von dem Grundrecht des Art 2 Abs. 1 in Verbindung mit Art 1 Abs. 1 GG umfasst. Das Grundrecht gewährleistet insoweit die Befugnis des Einzelnen, grundsätzlich selbst über die Preisgabe und Verwendung seiner persönlichen Daten zu bestimmen.“

[3] „Sozialdaten sind Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person (Betroffener), die von einer in § 35 SGB I genannten Stelle im Hinblick auf ihre Aufgaben nach diesem Gesetzbuch erhoben, verarbeitet oder genutzt werden“ (§ 67 Abs. 1 SGB X).

[4] „Jeder hat Anspruch darauf, dass die ihn betreffenden Sozialdaten (§ 67 Abs. 1 SGB X) von den Leistungsträgern nicht unbefugt erhoben, verarbeitet oder genutzt werden (Sozialgeheimnis). (. . .)“ (§ 35 Abs. 1 SGB I).

genen Informationen, sondern – anders als im BDSG – auch Betriebs- und Geschäftsgeheimnisse[5].

Um einen Ausgleich zu schaffen zwischen der verfassungsrechtlich garantierten Freiheit wissenschaftlicher Forschung[6] einerseits und dem verfassungsrechtlich verbürgten Recht auf informationelle Selbstbestimmung andererseits, hat der Gesetzgeber in den §§ 282 Abs. 7 SGB III und 75 SGB X Bedingungen für den Zugang der Forschung zu Sozialdaten festgelegt.

- In § 282 Abs. 7 SGB III ist geregelt, dass die BA externen Forschungseinrichtungen (faktisch) anonymisierte[7] Daten für die Zwecke der Arbeitsmarkt- und Berufsforschung übermitteln darf. Diese sogenannten Scientific Use Files (SUF) enthalten Mikrodaten, die so vergröbert wurden, dass eine Deanonymisierung nur mit unverhältnismäßig großem Aufwand an Zeit, Kosten und Arbeitskraft möglich ist.[8] Mit den SUFs können eine Vielzahl von Forschungsfragen analysiert werden. Obwohl das Deanonymisierungsrisiko sehr gering ist, ist die Nutzung dieser Datensätze an bestimmte Bedingungen geknüpft (siehe Kap. 3).

- Da das Analysepotential der SUFs durch die Anonymisierungsmaßnahmen eingeschränkt ist, können bestimmte Fragestellungen anhand dieser Daten nicht mehr beantwortet werden. Deshalb hat das FDZ schwach anonymisierte Datensätze erstellt, die an besonderen technisch abgeschotteten Gastwissenschaftlerarbeitsplätzen[9] angeboten werden oder die per Datenfernverarbeitung[10] analysiert werden können. Die Einsichtnahme von Sozialdaten an diesen Gastwissenschaftlerarbeitsplätzen stellt rechtlich gesehen eine „Datenübermittlung“ dar (§ 67 Abs. 6 Nr. 3b SGB X) und bedarf deshalb nach § 75 SGB X der Genehmigung durch das Bundesministerium für Arbeit und Soziales (BMAS). Das FDZ hat hierfür ein standardisiertes Verfahren entwickelt (siehe Kap. 3).[11]

---

5 „Betriebs- und Geschäftsgeheimnisse stehen Sozialdaten gleich“ (§ 35 Abs. 4 SGB I).

6 „Kunst und Wissenschaft, Forschung und Lehre sind frei. Die Freiheit der Lehre entbindet nicht von der Treue zur Verfassung.“ (Art. 5 Abs. 3 GG)

7 „Anonymisieren ist das Verändern von Sozialdaten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können.“ (§ 67 Abs. 8 SGB X)

8 „Für die Durchführung wissenschaftlicher Vorhaben dürfen vom Statistischen Bundesamt und den statistischen Ämtern der Länder Einzelangaben an Hochschulen oder sonstige Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung übermittelt werden, wenn die Einzelangaben nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können und die Empfänger Amtsträger, für den öffentlichen Dienst besonders Verpflichtete oder Verpflichtete nach Absatz 7 sind.“ (§ 16 Abs. 6 BStatG)

9 Die Gastrechner des FDZ sind konfigurierte PCs, die weder einen Zugang zum Internet haben noch Datenübertragungen auf externe Speichermedien oder Drucker erlauben.

10 Hierbei erstellen Wissenschaftlerinnen und Wissenschaftler auf der Basis von Testdaten Auswertungsprogramme. Das FDZ führt die Auswertungen mit den Originaldaten durch und übermittelt die Ergebnisse dem Forscher bzw. der Forscherin nach einer Datenschutzprüfung.

11 § 75 SGB X regelt generell die Übermittlung von Sozialdaten an Dritte. Neben dem vom FDZ angebotenen standardisierten Datenzugang über Gastaufenthalt besteht weiterhin die Möglichkeit einer kostenpflichtigen individuellen projektbezogenen Datenübermittlung durch die BA.

2.2 Datenschutz als Aufgabe des FDZ

Neben der Erstellung und Dokumentation von Forschungsdaten besteht die Hauptaufgabe des FDZ in der praktischen Umsetzung der sogenannten „statistical disclosure control“ (vgl. Ritchie (2011)). Hierunter versteht man die Sicherstellung der Geheimhaltung von Informationen über statistische Einheiten z.B. Personen oder Betriebe. Um dies zu gewährleisten, muss vor Herausgabe einer Information deren Deanonymisierungspotential geprüft werden. Als hoch riskant gelten dabei nicht nur Informationen wie beispielsweise Name, Adresse oder Sozialversicherungsnummer, sondern auch Charakteristika oder Kombinationen von Charakteristika, die eine indirekte Identifikation ermöglichen. Obwohl davon auszugehen ist, dass auf Forscherseite keinerlei Interesse an der Identifikation von einzelnen Individuen oder Betrieben besteht, muss auch bei der Weitergabe von Forschungsdaten die „statistical disclosure control“ eingehalten werden. Ziel ist es, zu verhindern, dass Informationen veröffentlicht[12] werden, anhand derer Dritte bestimmte Personen oder Betriebe wiedererkennen oder Betroffene sich selbst identifizieren können. In diesem Zusammenhang muss berücksichtigt werden, dass die Betroffenen selbst und auch dritte Personen über zusätzliches Wissen verfügen könnten, durch das eine Deanonymisierung möglich wird.

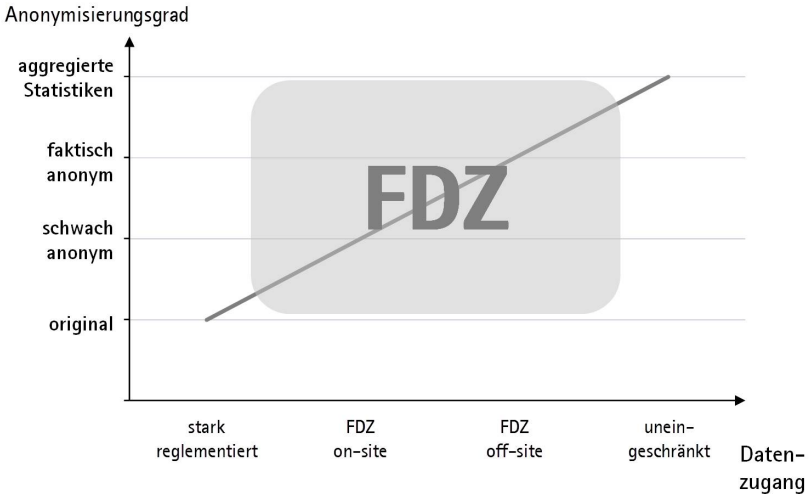
Ein Beispiel soll dieses Problem verdeutlichen: In einer wissenschaftlichen Publikation wird das Durchschnittseinkommen von angestellten Zahnärztinnen nach Landkreisen ausgewiesen. Auf Grund einer 2%-Stichprobe liegen für viele Kreise, insbesondere für Kreise mit geringer Einwohnerzahl, nur ein oder zwei Personen der Berechnung zu Grunde. Da der Beruf des Zahnarztes meist als selbständige Tätigkeit ausgeübt wird, kann es durchaus möglich sein, dass es in einem Kreis tatsächlich nur eine angestellte Zahnärztin gibt, die zufällig in der Stichprobe enthalten ist. Über dieses zusätzliche Wissen, dass es nur eine angestellte Zahnärztin im Kreis gibt, werden in der Regel weder die Forschenden noch die Mitarbeiterinnen und Mitarbeiter des FDZ verfügen. Aber die betroffene Zahnärztin und Einwohner des Kreises können Kenntnis davon haben. Somit kann sich die Zahnärztin selbst identifizieren und Dritte erfahren ihr Einkommen. Eine ähnliche Problematik liegt vor, wenn es nur zwei Zahnärzte gibt und beide in der Stichprobe enthalten sind. Hier können Dritte zwar das Einkommen des einzelnen Zahnarztes nicht erkennen, aber die beiden betroffenen Zahnärzte können mit dem Wissen um ihr eigenes Einkommen das Einkommen des jeweils anderen berechnen. Für Betriebe ist das Deanonymisierungsrisiko ungleich größer, da zusätzliches Wissen über Betriebe leicht zugänglich ist. Besonders Großbetriebe können über die Angaben von Branche und Standort leicht identifiziert werden.

[12] Veröffentlichen ist auch das Verfügbarmachen für unbefugte Dritte.

3 Das FDZ-Portfolio zum Datenschutz

Die Aufgabe des FDZ der BA im IAB ist es, die Anonymität der statistischen Einheiten zu sichern. Diese steht immer im Zusammenhang mit dem Aggregationsniveau der zu schützenden Informationen. Allgemein betrachtet reicht das Spektrum des Anonymisierungsgrads von den Originaldaten bis hin zu stark vergröberten aggregierten Statistiken. Je nachdem, wie stark die Daten anonymisiert wurden, gestaltet sich der Datenzugangsweg. So können aggregierte Statistiken teilweise frei über das Internet veröffentlicht werden, während die Originaldaten nur nach ausführlicher Prüfung und bei unabdingbarer Notwendigkeit weitergegeben werden. Dieser Zusammenhang wird deutlich in Abbilung 1.

Abbildung 1: Anonymisierungsgrad und Datenzugang



Quelle: eigene Darstellung

Aus der Abbildung wird ersichtlich, dass sowohl eine große Spannweite zwischen Originaldaten und aggregierten Statistiken als auch zwischen uneingeschränktem und stark reglementiertem Datenzugang besteht. Diese bestimmen den Handlungsspielraum der datenhaltenden Institutionen. Da das FDZ weder Originaldaten noch aggregierte Statistiken zur Verfügung stellt, besitzt es einen etwas geringeren Wirkungsbereich, innerhalb dessen es sich jedoch nach unterschiedlichen Bedürfnissen flexibel ausrichten kann. So bietet das FDZ seine Mikrodatensätze zum einen als faktisch anonymisierte Scientific Use Files (SUF) an, die die Nutzerinnen und Nutzer in den Räumen ihrer Forschungseinrichtung analysieren können. Zum anderen können auch schwach anonymisierte Daten, die detaillierte Informationen enthalten, im Rahmen von Gastaufenthalten oder über Datenfernverarbeitung genutzt werden. Im Allgemeinen gilt: Je höher das Anonymisierungsniveau, umso flexibler gestaltet sich der Datenzugang. Umso sensibler die Information, umso stärker reglementiert ist der Datenzugang. Für alle Datenzugangswege stellt das FDZ die Datensicherheit mit unterschiedlichen Verfahren, die in Zusammenarbeit mit dem Justiziariat des IAB entwickelt wurden, sicher. Um diese zu koordinieren, arbeitet das FDZ nach einem Portfolio-Ansatz in Anlehnung an Lane/Heus/Mulcahy (2008). Die vier Hauptfelder des FDZ-Portfolios sind in Abbildung 2 dargestellt.

Abbildung 2: Vier-Felder-Portfolio

Quelle: eigene Darstellung

Grundsätzlich unterscheiden wir Maßnahmen, die im Vorfeld der Datennutzung umgesetzt werden, und jene, die im Anschluss an die Datennutzung erfolgen. Im Vorfeld der Datennutzung werden zum einen Zugangsvoraussetzungen geprüft und die Nutzungsbedingungen vertraglich vereinbart. Zum anderen werden die Mikrodaten so anonymisiert, dass der Datenschutz gewährleistet ist. Zusätzlich werden die Ergebnisse nach Analysen mit den schwach anonymisierten Daten einer Datenschutzprüfung unterzogen. Die Detaills der einzelnen Aspekte werden nachfolgend dargestellt sowie in Tabelle 11 im Anhang zusammengefasst.

3.1 Prüfung der Zugangsvoraussetzungen

Gemäß den gesetzlichen Vorgaben (Kap. 2) ist die Nutzung der Daten des FDZ an bestimmte Bedingungen geknüpft. Um zu klären, ob diese Voraussetzungen erfüllt sind, muss vorab ein Antrag gestellt werden. Folgende formelle Bedingungen werden vom FDZ geprüft: Zunächst muss es sich um ein wissenschaftliches Projekt im Bereich der Arbeitsmarkt- und Berufsforschung handeln. Allem voran steht zudem die Prüfung der Erforderlichkeit der Datennutzung. Es muss nachgewiesen werden, dass das Forschungsziel mit keinen anderen Daten, die leichter verfügbar sind (z.B. aggregierte Daten) erreicht werden kann. Neben der formellen Prüfung dient die zusätzlich geforderte Beschreibung des Forschungsprojekts dazu, die inhaltliche Machbarkeit des Forschungsvorhabens in Bezug auf die beantragten Daten zu prüfen. Hierbei beraten die Mitarbeiterinnen und Mitarbeiter des FDZ die Antragssteller hinsichtlich der Analysepotenziale und Qualitätsaspekte der Daten.

Je nach Art des Datenzugangs müssen weitere Voraussetzungen erfüllt sein: Bei der Beantragung eines SUFs ist nachzuweisen, dass es sich bei der durchführenden Stelle um eine unabhängige wissenschaftliche Forschungseinrichtung handelt. Darüber hinaus legt der Antragssteller in einem Datensicherheitskonzept dar, dass ausreichende technische und organisatorische Maßnahmen zur sicheren Speicherung und Verarbeitung der Daten in der Einrichtung gegeben sind. Bei der Beantragung eines Gastaufenthalts zur Nutzung der

schwach anonymisierten Daten muss das Forschungsprojekt auch im öffentlichen Interesse stehen. Nach der inhaltlichen Prüfung des Antrags vom FDZ wird dieser dem BMAS zur Genehmigung vorgelegt. Da bei der ausschließlichen Datenfernverarbeitung die Nutzerinnen und Nutzer keinen direkten Zugriff auf die Mikrodaten haben, sind die Zugangsbedingungen weniger streng. Dennoch hat es sich auch hier um ein wissenschaftliches Forschungsprojekt zu handeln und der Bezug zur Arbeitsmarkt- und Berufsforschung bzw. zum Sozialleistungssystem muss vorliegen.[13]

3.2 Regelungen zu Datenzugang und -nutzung

Nach erfolgreicher Prüfung des Antrags werden Verträge geschlossen, in denen die Bedingungen der Datennutzung geregelt sind. In allen Verträgen sind die zentralen Prinzipien der Zweckbindung, der zeitlichen Befristung und der Begrenzung der Daten und des Personenkreises verankert. Folglich ist die Datennutzung nur für ein definiertes inhaltliches Projekt mit den beantragten Daten innerhalb des vertraglich festgelegten Zeitraums erlaubt. Zudem sind die Personen, die Zugang zu den Daten erhalten, festgelegt. Dieser Personenkreis sollte so klein wie möglich gehalten werden. Darüber hinaus beinhalten die Verträge Verbote zur Weitergabe der Daten an Dritte, zur Zuspielung anderer Mikrodaten sowie zur Deanonymisierung.

Die Verträge variieren je nach Art des Datenzugangs. So ist für SUFs das Datensicherheitskonzept der Forschungseinrichtung zusätzlich Bestandteil des Nutzungsvertrags. Darüber hinaus ist das Forschungsinstitut nach Ablauf der Vertragsdauer verpflichtet, alle Mikrodaten zu löschen und ggf. Datenträger, auf denen die Daten übermittelt wurden, an das FDZ zurück zu schicken. Im Nutzungsvertrag für Gastaufenthalte sind Verhaltensvorgaben für den Gastaufenthalt vorgegeben. Da bei der Nutzung von schwach anonymisierten Daten der Datenschutz zusätzlich durch die Kontrolle der Ergebnisse gewährleistet ist (siehe Kap. 3.4), werden die Nutzerinnen und Nutzer vertraglich verpflichtet, nicht rückzurechnen. D.h. sie dürfen bei der Datenschutzprüfung gelöschte Werte einer Tabelle nicht anhand von Differenzenbildung bereits geprüfter Werte aus früheren Ergebnissen oder ähnlichen Verfahren wieder erschließen.[14] In allen Verträgen finden sich neben den erläuterten Nutzungsbedingungen auch der Hinweis auf Sanktionen bei Missbrauch.

3.3 Anonymisierung

Neben den Beschränkungen beim Datenzugang wird der Datenschutz bereits in der Datenaufbereitung berücksichtigt. Ein wichtiger Schritt zur Sicherung der Daten ist das Ziehen einer Stichprobe. Damit allein verringert sich das Deanonymisierungsrisiko deutlich. Wenn wie beispielsweise bei der Stichprobe der Integrierten Arbeitsmarktbiografien (SIAB) eine Ziehungswahrscheinlichkeit von 1:50 vorliegt und man glaubt, man hätte eine Person

[13] Detaillierte Informationen, welche Angaben in den Anträgen zur Datennutzung abgefragt werden, sowie Bearbeitungshilfen finden sich auf der Homepage des FDZ unter `http://fdz.iab.de/de/FDZ_Data_Access.aspx`.

[14] Schwach anonymisierten Daten können entweder im Rahmen von Gastaufenthalten oder über den Weg der Datenfernverarbeitung analysiert werden. Ein Gastaufenthaltsvertrag schließt auch immer die Datennutzung über Datenfernverarbeitung mit ein.

identifiziert, dann besteht die Möglichkeit, dass es in der Grundgesamtheit noch 49 andere Personen mit den gleichen Charakteristika gibt. Somit ist zur eindeutigen Identifikation einer Person in der Stichprobe das zusätzliche Wissen erforderlich, dass entweder die Charakteristika der Person in der Grundgesamtheit einzigartig sind oder dass die vermutliche identifizierte Person in der Stichprobe enthalten ist. Neben der Ziehung von Stichproben setzt das FDZ weitere Methoden der Anonymisierung ein. Dazu werden drei Anonymisierungsstufen unterschieden:

- Schwach anonymisiert,
- Faktisch anonymisiert oder
- Absolut anonymisiert.

Bei den schwach anonymisierten Daten werden Identifikatoren wie Namen, Adresse, Sozialversicherungsnummer oder Betriebsnummer gelöscht und die Ausprägungen einiger besonders sensibler Merkmale wie z.B. die Staatsangehörigkeit zusammengefasst.[15] Da das Risiko der Deanonymisierung von Großbetrieben oder Branchenführern durch detaillierte Angaben zum Wirtschaftszweig und Standort des Betriebs erheblich erhöht ist, werden standardmäßig nur die Wirtzschaftsgruppen (3-Stellerebene der Wirtschaftszweigklassifikation) und das Bundesland zur Verfügung gestellt. In begründeten Fällen können jedoch auch Informationen auf 5-Stellerebene (Wirtschaftsunterklasse) und zum Kreis genutzt werden. Diese sensiblen Merkmale dienen in der regel entweder dazu, aggregierte Statistiken auf dieser Ebene zuzuspielen (z.B. Arbeitslosigkeitsquoten nach Kreisen) oder dazu, eigene regionale oder brachenspezifische Gruppierungen zu bilden, die in den vorgegebenen Klassifikationen nicht vorgesehen sind. Sehr spezifische Analysen zu bestimmten Wirtschaftsunterklassen und/oder kleinen regionalen Einheiten werden häufig aufgrund des sehr hohen Deanonymisierungsrisikos nicht genehmigt.

Scientific Use Files sind faktisch anonymisierte Mikrodaten, deren Informationsgehalt soweit reduziert wurde, dass eine Deanonymisierung nur mit unverhältnismäßig großem Aufwand an Zeit, Kosten und Arbeitskraft möglich ist.[16] Dabei muss häufig zwischen der Vergröberung unterschiedlicher Merkmale abgewogen werden. Will man z.B. detaillierte Regionalangaben in den Beschäftigtendaten erhalten, müssen dafür andere Variablen (wie Betriebsinformationen) stark zusammengefasst oder sogar gelöscht werden. Generell werden die Empfehlungen von Müller et al. (1991) bei der SUF-Erstellung berücksichtigt.

Als absolut anonymisiert gelten aggregierte Daten, bei denen weder auf einzelne Personen noch auf Betriebe - auch nicht auf Großbetriebe oder Branchenführer - geschlossen werden kann. Eine Ergebnistabelle aus einem Mikrodatensatz muss nicht automatisch absolut

[15] Damit geht die schwache Anonymisierung noch einen Schritt weiter als die Pseudonymisierung von Daten. „Pseudonymisieren ist das Ersetzen des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren.“ (§ 67 Abs. 8a SGB X)

[16] „Anonymisieren ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können.“ (§ 67 Abs. 8 SGB X)

anonym sein. Wenn beispielsweise einzelne Zellen nur mit einer Person besetzt sind, ist auch eine aggregierte Tabelle nicht anonym. Da Probleme dieser Art sehr häufig bei der Analyse von schwach anonymisierten Daten entstehen, müssen diese Ergebnisse einer nachträglichen Kontrolle unterzogen werden (siehe Kap. 3.4). Darüber hinaus gelten sogenannte Campus Files als absolut anonym. Campus Files sind Mikrodatensätze, die durch Informationsreduktion und datenverändernde Verfahren keine Identifikation der enthaltenen Personen oder Betriebe zulassen. Durch diese massiven Eingriffe sind diese jedoch nicht mehr für inhaltliche Analysen geeignet, sondern dienen lediglich zur Vermittlung von Erhebungs-, Datenmanagement- und Analysetechniken in der Lehre an Hochschulen und Forschungsinstituten (vgl. Kirchner/Gschwind, 2011).

3.4 Ergebniskontrolle

Da es sich bei den schwach anonymisierten Daten immer noch um Sozialdaten handelt und daher ein Restrisiko der Deanonymisierung anhand von Ergebnistabellen besteht, ist trotz Nutzungsvertrag eine Output-Kontrolle notwendig. Um diese Prüfung möglichst schnell und effizient durchführen zu können, müssen die Analyseprogramme bestimmten Vorgaben des FDZ entsprechen. Allerdings sollen damit die Forscherinnen und Forscher nicht in der Verwendung von Analysemethoden eingeschränkt werden. Dadurch lassen sich die Prüfungen aber nicht vollständig in ein standardisiertes und automatisiertes Verfahren einbinden. Vielmehr müssen sich die Datenschutzprüfungen immer auf den vorliegenden Einzelfall beziehen. Dies erfordert, dass die Mitarbeiterinnen und Mitarbeiter des FDZ alle Ergebnisse vor der Herausgabe sichten. Die Prüfung erfolgt nach bestimmten Kriterien und Richtlinien. Es existieren im Expertenkreis allgemeingültige Regeln, die grundsätzlich Anwendung finden sollten.[17] Wie diese Datenschutzprüfung in der Praxis umgesetzt wird, ist in den nächsten Kapiteln dargestellt.

[17] Ein Beispiel hierfür wäre ESSnet. Dies ist ein internationales Projekt des European Statistical System, das sich um alle Bereiche, die mit Datenschutz in Verbindung gebracht werden können, kümmert, wie z.B. die Herausgabe von Standards, die bei Datenschutzprüfungen eingehalten werden sollten. (`http://neon.vb.cbs.nl/casc/handbook.htm`)

4 Datenschutzprüfungen am FDZ

Alle Ergebnisse, die auf Basis der schwach anonymisierten Daten erzeugt wurden, werden vor der Übermittlung an die Nutzerinnen und Nutzer einer Datenschutzprüfung unterzogen. Der Prüfungsaufwand hängt hierbei von den zu prüfenden Ergebnisdateien ab.

4.1 Theoretische Unterscheidung von Analyseergebnissen

Die produzierten Ergebnisse können ihrem Inhalt nach als „sicher“ oder „unsicher“ klassifiziert werden. Mittels dieser Unterscheidung wird auf die Höhe des Risikos der Deanonymisierung des Datenmaterials rückgeschlossen. Man geht davon aus, dass bei Resultaten, die als sicher eingestuft werden, keine Gefahr bezüglich der Deanonymisierung des Datenmaterials besteht. Im Gegensatz dazu besteht bei unsicher bewerteten Auswertungsausgaben ein Restrisiko der Deanonymisierung. Aus diesem Grund werden in diesen Auswertungsergebnissen Löschungen vorgenommen, bis die Ergebnistabellen absolut anonym sind. Der Sinn der Datenschutzprüfung besteht folglich darin, als „unsicher“ klassifizierte Analyseausgaben in unbedenkliche „sichere“ Ausgaben zu transformieren. Die Einordung des Outputs erfolgt anhand folgender Gesichtspunkte nach Brandt et al. (2010):

- Verwendetes Datenmaterial
- Art der durchgeführten Analysen
- Einschränkungen des Datenmaterials auf oder Hinzunahme von bestimmten Merkmalen
- Angewandte Datentransformationen

Wenn anhand dieser Klassifizierung der Output bezüglich der Datenschutzmaßnahmen des FDZ als sicher eingestuft wurde, werden bei diesen Ergebnissen keine Löschungen vorgenommen. Generell unbedenklich sind z.B. Koeffizienten multivariater Schätzungen für große Populationen. In den meisten Fällen handelt es sich bei Auswertungen allerdings um Ergebnisse, die nicht direkt als „sicher“ eingestuft werden können, was vor allem bei deskriptiven Auswertungen der Fall ist. Als nicht sicherer Output gelten grundsätzlich die

- Ausgabe von statistischen Kennzahlen, wie z.B. Mittelwerten,
- Ausgabe einzelner Datenpunkte z.B. in einem Scatterplot (diese lassen eventuelle Rückschlüsse auf ein Individuum zu),
- Ausgabe der Perzentile sowie
- Ausgabe der Beobachtungsanzahlen.

Die Klassifizierung in „sicher“ und „unsicher“ heißt nicht, dass sichere Outputs ohne vorherige Durchsicht übermittelt werden, sondern lediglich, dass in „sicheren“ Outputs keine Löschungen vorgenommen werden. Es existieren Richtlinien zur Programmgestaltung, an die sich die Gastwissenschaftlerinnen und Gastwissenschaftler halten müssen, um die Klassifikation von Analyseergebnissen zu erleichtern.

4.2 Voraussetzungen der Durchführbarkeit von Datenschutzprüfungen

Bei der Datenschutzprüfung werden nicht nur die einzelnen Ergebnistabellen unabhängig voneinander auf ausreichende Fallzahlen kontrolliert. Vielmehr wird das ganze Vorgehen der Datenaufbereitung und Analyse betrachtet. Der Ouput wird nicht losgelöst, sondern in Bezug auf das Forschungsvorgehen kontrolliert. Dafür ist es erforderlich, dass die Programme nachvollziehbar und ausführlich dokumentiert sind. Folgende Kriterien sollen bei der Programmierung erfüllt sein, damit ein Programm den Richtlinien des FDZ[18] entspricht:

1. Ausführliche Dokumentation der Auswertungsschritte:
   Um die Ergebniskontrollen durchführen zu können, müssen sich die FDZ-Mitarbeiterinnen und -Mitarbeiter in den Programmen zurecht finden können und die Programmierung zumindest im Gesamten nachvollziehen können. Die Dokumentation der Programme gewährleistet dies.

2. Syntaxgestützte Programmierung:
   Die Auswertungen müssen über eine Programmdatei mittels Eingabe des entsprechendes Programmcodes erfolgen. Die Programmierung muss derart aufgebaut sein, dass auch in den Ergebnisdateien die Programmcodes enthalten sind, da so die Abfolge der einzelnen Programmierungsschritte ersichtlich ist. Die Datenaufbereitung muss in der Regel in Stata erfolgen. Bei weiterführenden Analysen kann mit anderen Softwarepaketen gearbeitet werden.

3. Speichern von Ergebnisdateien:
   Damit ersichtlich wird, welche Programmdatei die jeweilige Outputdatei generiert, muss für jede Programmdatei eine entsprechend benannte Ausgabedatei angelegt werden.

4. Anlegen einer Master-Datei:
   Die Master-Datei wird angelegt, um aus dieser alle Programme, die zu einem Auswertungsprogramm gehören, auf einmal sowohl bei der Datenfernverarbeitung als auch nach einem Gastaufenthalt zu starten. Inhalt der Master-Datei ist somit der Aufruf der einzelnen Programme in der richtigen Reihenfolge. Ebenso sollte auch in der Master-Datei kurz der Inhalt der aufgerufenen Programme erläutert werden.

Da bei Nichteinhaltung der Richtlinien die Auswertungen meist schwer nachvollzogen werden können, behält sich das FDZ vor, allein bei Zweifeln bezüglich der Gefährdung des Datenschutzes, die Auswertungsergebnisse komplett zu löschen.

4.3 FDZ-Richtlinien zur Prüfung der Analyseergebnisse

Ziel der angewandten Routinen ist die Veränderung der Ergebnisse durch Löschungen, sodass kein Deanonymisierungspotential mehr vorhanden ist. Am FDZ existieren allgemeine Kriterien für die Ergebnisprüfung. Bei der Löschung wird grundsätzlich unterschieden zwischen:

[18] Weiterführende Informationen finden Sie unter: `http://doku.iab.de/fdz/access/Vorgaben_DAFE.PDF`.

- der Primärsperrung, die eine Identifikation von Informationen in einer Zelle einer Tabelle unterdrückt,
- der Sekundärsperrung, die eine Identifikation von Informationen durch Zwischen- und oder Randsummen verhindert und
- der Dominanzsperrung[19], die eine Identifikation von Dominanzbetrieben unterbindet.

Eine vollständige automatische Datenschutzprüfung gibt es nicht. Jedoch behilft sich das FDZ mit einem eigens entwickelten Programmskript, das die Ergebnisdateien bei ausgewählten Stata-Befehlen auf kleine Fallzahlen durchsucht und entsprechend Löschungen vornimmt. Da die automatische Datenschutzprüfung nur Standardausgaben prüft, werden alle zusätzlichen Auswertungen manuell kontrolliert und ggf. Löschungen vorgenommen. Das Skript wird kontinuierlich angepasst und weiter entwickelt. Im Folgenden wird kurz auf die standardisierten Löschungen eingegangen. Kritisch eingestuft und damit gelöscht werden Ergebnisse, die auf weniger als 20 Beobachtungen beruhen. Diese Mindestanforderung gilt sowohl für die Betriebs- als auch die Personendaten. Diese Grenze wurde aus folgendem Grund gewählt: Das FDZ prüft jeden Output eigenständig und vergleicht die Ergebnisse nicht mit denen der vorherigen Auswertungen. Darüber hinaus verpflichten sich die Wissenschaftlerinnen und Wissenschaftler vertraglich, keine Rückrechnungen aus den einzelnen zugeschickten Auswertungsergebnissen vorzunehmen.

4.3.1 Statistische Kennzahlen

Statistische Kennzahlen lassen auf den ersten Blick keinen Rückschluss auf die zugrundegelegten Fallzahlen zu. Das heißt aber nicht, dass die Ausgabe statistischer Kennzahlen, wie z.B. Mittelwerte, nicht als problematisch zu sehen ist. Auch hier gilt das Prinzip, dass die angesprochenen Kennzahlen nur dann als sicher eingestuft werden, wenn die Berechnungsgrundlage mindestens 20 Beobachtungen umfasst. Als Spezialfall ist die Ausgabe statistischer Kennzahlen bei Dummys zu sehen. Bei binär kodierten Merkmalen verteilen sich die Ausprägungen dieser auf lediglich zwei Kategorien. Auch wenn die Gesamtanzahl der Beobachtungen einer Dummy-Variable größer als 20 ist, kann es sein, dass durch eine schiefe Verteilung nur drei Personen auf die eine Ausprägung fallen. In diesem Fall wird die Auswertung als unsicher angesehen, auch wenn nicht direkt über die Gesamtanzahl der Beobachtungen auf eine geringe Besetzung einer Kategorie geschlossen werden kann. Denn dies kann mit Hilfe des Mittelwerts einfach herausgefunden werden. Um diese Fälle identifizieren und prüfen zu können, ist es erforderlich, dass bei Mittelwerten neben

[19] Das Risiko der Indentifikation eines Betriebes in der Auswertung steigt, z.B. durch die Verwendung von detaillierten Branchen oder kleinräumigen Regionalangaben. Die Prüfung auf Dominanzfälle wird durch folgende Maßnahmen durchgeführt: Die Mindesfallzahlgrenze von 20 Einheiten wird auch auf die Anzahl der Betriebe angewandt, wenn Angaben zur Anzahl der Beschäftigten aus Betriebsdaten ausgewiesen werden. Die ausschließliche Verwendung von Stichproben anstelle von Grundgesamtheiten sowie die Prüfung der Programmschritte stellt ebenfalls sicher, dass keine Dominazfälle identifiziert werden können. Darüber hinaus werden detaillierte Branchen- und Regionalangaben nur mit besonderer Begründung zur Verfügung gestellt. Standardmäßig werden nur Angaben auf Ebende der Bundesländer und der 3-Steller der Wirtschaftszweigklassifikationen angeboten (siehe Kap. 3.3). Wir orientieren uns bei der Dominanzsperrung an den Vorgaben der Bundesagentur für Arbeit (2012).

der Fallzahl immer auch Minimum, Maximum und Standardabweichung ausgegeben werden. Da das Programmskript des FDZ nur die Standardausgaben von Kennzahlen erkennt, müssen in speziellen Fällen die Besetzungen der beiden Kategorien bei Dummyvariablen nachträglich berechnet und ggf. die statistischen Kennzahlen gelöscht werden.

4.3.2 Perzentile

Bei der Ausgabe von Perzentilen muss darauf geachtet werden, dass mindestens 20 Beobachtungen im jeweiligen Perzentil enthalten sind. Bei einer detaillierten Ausgabe (1% Perzentile) müssen folglich insgesamt mindestens 2000 Beobachtungen in die Ausgabe mit einfließen, um die Datenschutzvorgaben des FDZ zu gewährleisten. Grunsätzlich gilt, je mehr Information man erhalten will, umso mehr Beobachtungen müssen für die gesamte Verteilung vorhanden sein:

- Mind. 20 Beobachtungen für die Ausgabe von Mittelwerten (Ausnahme Dummys, siehe Kap. 4.3.2)
- Mind. 40 Beobachtungen für die Ausgabe von 50%-Perzentilen
- Mind. 80 Beobachtungen für die Ausgabe von 25%- oder 75%-Perzentilen
- Mind. 200 Beobachtungen für die Ausgabe von 10%- oder 90%-Perzentilen
- Mind. 400 Beobachtungen für die Ausgabe von 5%- oder 95%-Perzentilen
- Mind. 2000 Beobachtungen für die Ausgabe von 1%- oder 99%-Perzentilen

4.3.3 Gewichtungen

Im Falle von gewichteten Ausgaben wird die Datenschutzprüfung stets anhand der ungewichteten Werte vorgenommen. Die Ergebnisdateien müssen eine eindeutige Zuordnung der gewichteten zur dazugehörigen ungewichteten Ausgabe erlauben. Löschungen in den ungewichteteten Tabellen werden in die entsprechenden gewichteten übertragen. Fehlt die ungewichtete Ausgabe, wird die gewichtete Tabelle komplett gelöscht.

4.3.4 Grafiken

Die Prüfung und Weitergabe von Grafiken ist ein zusätzlicher Service des FDZ.[20] Prinzipiell können nur Grafiken herausgegeben werden, die mit Hilfe der Auswertungsprogramme erstellt wurden. Aufgrund der zeitintensiven Datenschutzprüfungen von Grafiken sollten diese nur dann erstellt werden, wenn sie nicht nachträglich aus den Werten der Ergebnisdateien erstellt werden können. Zu jeder Grafik ist nachzuweisen, auf wie vielen Beobachtungen

[20] Derzeit werden die grafischen Abbildungen aufgrund des hohen Prüfungsaufwands nicht übermittelt. Es erfolgt die Prüfung der Grafiken anhand von Ergebnisstabellen, aus denen die Nutzerinnen und Nutzer ihre Grafiken wieder erzeugen können. Dafür bietet das FDZ Hilfsmittel an. (`http://doku.iab.de/fdz/access/Vorgaben_DAFE.PDF`)

die einzelnen dargestellten Werte basieren. Auch für Grafiken gilt die Fallzahlgrenze von mindestens 20 Beobachtungen. Somit sind z.B. Scatterplots durch ein hohes Deanonymisierungspotenzial gekennzeichnet, da hinter den einzelnen ausgegebenen Datenpunkten mit hoher Wahrscheinlichkeit weniger als 20 Beobachtungen stehen.

4.3.5 Dateiformate

Statistikprogramme bieten ihren Nutzerinnen und Nutzern in der Regel die Möglichkeit, die Auswertungsergebnisse in unterschiedlichen Formaten (z.B. LaTeX oder ASCII) als gesonderte Dateien zu speichern. Da eine Prüfung dieser Dateien immer unter Berücksichtigung der dazugehörigen Programm- und Ergebnisdateien erfolgen müsste, würde sich die Komplexität und der zeitliche Aufwand dafür im FDZ enorm erhöhen. Damit würde eine zügige Übermittling der Ergebnisse behindert werden. Deshalb müssen die so erzeugten Ergebnisse wieder in die urspüngliche Ausgabedatei direkt unter den dazugehörigen Analyseergebnissen eingebunden werden.[21] Zum Beispiel können mit Stata-Ado-Dateien Ergebnisse in LaTeX-Codes ausgegeben werden. Diese werden nur dann weitergegeben, wenn Sie im log-file direkt unter den dazugehörigen Stata-Tabellen zu finden sind. Bei deskriptiven Tabellen werden sie komplett gelöscht, sobald eine Zelle zu geringe Fallzahlen aufweist. Da die Durchsicht der LaTeX-Codes zusätzlichen Aufwand bedeutet, sind die Nutzerinnen und Nutzer dazu angehalten, sich lediglich die Ergbnisse, die sie für ihre Publikation benötigen, mit LaTeX-Codes ausgeben zu lassen.

4.3.6 Weitergabe von aggregierten Datensätzen

Es besteht die Möglichkeit, aus den schwach anonymisierten Mikrodaten aggregierte Datensätze zu erstellen und übermittelt zu bekommen. Da die Prüfung aggregierter Daten zeitaufwändig ist, existieren auch für die Übermittlung aggregierter Daten bestimmte Regeln. Vor der Generierung von aggregierten Datensätzen erfolgt ein Gespräch mit dem FDZ. Darin muss geklärt werden, auf welche Ebene aggregiert wird und welche Variablen in welchem Aggregationszustand (Summe, Mittelwert, etc.) im Datensatz enthalten sind. Damit die Daten geprüft werden können, muss zu jeder aggregierten Variable ein zusätzliches Merkmal erzeugt werden, das die Fallzahl enthält, auf derer der aggregierte Wert basiert. Da pro Projekt nur einmal ein aggregierter Datensatz übermittelt wird, empfielt das FDZ einen Gastaufenthalt zur Erstellung dieses Datensatzes.

Natürlich sind neben den angesprochenen Datenabfragen noch weitere Analysemethoden möglich. Da in diesem Beitrag die grundsätzliche Vorgehensweise des FDZ bei der Outputkontrolle vorgestellt werden soll, wird an dieser Stelle auf Erläuterungen zu anderen Auswertungsmöglichkeiten verzichtet. Die nachfolgenden Beispiele sollen dieses Vorgehen veranschaulichen.

[21] In Stata kann dies z.B. mit dem Befehl „type PFADANGABE“ erfolgen.

4.4 Beispiele

Abschließend wird mit Hilfe einiger Beispiele gezeigt, welche Informationen bei der Datenschutzprüfung entfernt bzw. nicht frei gegeben werden. Die Beispiele wurden mit den Testdaten[22] des IAB-Betriebspanels erzeugt. Zunächst wird die Primärsperrung, die eine Löschung von Informationen innerhalb einer Tabelle umfasst, veranschaulicht. Tabelle 1 ist die Originaltabelle und Tabelle 2 umfasst die Datenschutzprüfung. Beispiel 1 weist die Anzahl der Betriebe mit und ohne Betriebsrat für unterschiedliche Betriebsgrößenklassen in Ostdeutschland aus.

Wie bereits erwähnt liegt die Grenze der zu löschenden Werte bei kleiner 20. Im Beispiel betrifft dies den Wert 16. Um soviel Informationen wie möglich beizubehalten, werden in der Regel die Randsummen belassen und der dazugehörige Wert, hier 142, anonymisiert. Dies reicht aber noch nicht aus, um eine Reidentifkation auszuschließen, da über die Randsummen und die verbleibenden Werte in der Tabelle auf die gelöschten Werte geschlossen werden kann. Deshalb müssen zwei weitere Werte entfernt werden. Es ist nicht notwendig die kompletten Werte bei mehrstelligen Zahlen zu entfernen, es reicht die Löschung der letzten ganzen Stelle. Es gibt keine Regel, welche zwei Werte als nächstes gelöscht werden. Dies obliegt der prüfenden Person am FDZ. Im Allgemeinen wird versucht, die nächst kleineren Werte zu entfernen. Dies ist im vorliegenden Beispiel der Wert 39 und dazugehörig der Wert 547.

Beispiel 1: Ostdeutschland

	Betriebsrat		
Anz. svpfl. Besch.	Ja	Nein	Total
1 1-4 svb	43	1,380	1,423
2 5-9 svb	39	547	586
3 10-19 svb	89	487	576
4 20-49 svb	250	590	840
5 50-99 svb	255	245	500
6 100-199 svb	290	110	400
7 200-499 svb	283	65	348
8 500-999 svb	142	16	158
Total	1,391	3,440	4,831

Tabelle 1: vorher

	Betriebsrat		
Anz. svpfl. Besch.	Ja	Nein	Gesamt
1 1-4 svb	43	1,380	1,423
2 5-9 svb	3*	54*	586
3 10-19 svb	89	487	576
4 20-49 svb	250	590	840
5 50-99 svb	255	245	500
6 100-199 svb	290	110	400
7 200-499 svb	283	65	348
8 500-999 svb	14*	/	158
Total	1,391	3,440	4,831

Tabelle 2: nachher

Die Löschung von Werten kann sich auf weitere Tabellen erstrecken, wenn Informationen für bestimmte Merkmale differenziert abgebildet werden. So handelt es sich im Beispiel 1 um Informationen, die nur Ostdeutschland betreffen. Wenn die gleichen Informationen für West- und Gesamtdeutschland abgebildet werden, so dürfen diese Tabellen nicht unabhängig voneinander betrachtet werden. Eine Reidentifikation bereits gelöschter Werte

[22] Die Testdaten des IAB-Betriebspanels dienen der Erstellung und Prüfung von Auswertungsprogrammen im Vorfeld der Datenfernverarbeitung. Die Testdaten wurden durch Ziehung einer Substichprobe und Verfremdung der Daten bei gleichzeitigem Erhalt wichtiger Datenstrukturen erzeugt. Dadurch sind keine inhaltlichen Auswertungen mit den Testdaten möglich.

ist ansonsten durch Differenzenbildung (Gesamt-West=Ost) möglich. Im vorliegenden Beispiel 2 wurden daher in der Tabelle 4 die Werte, die an gleicher Stelle stehen wie im Beispiel 1, gelöscht. Die Tabelle 5 beinhaltet die Werte für Gesamtdeutschland und bleibt unverändert, weil ein Rückrechnen nicht mehr möglich ist.

Beispiel 2

Westdeutschland

	Betriebsrat		
Anz. svpfl. Besch.	Ja	Nein	Total
1 1-4 svb	64	2,461	2,525
2 5-9 svb	54	847	901
3 10-19 svb	130	762	892
4 20-49 svb	364	853	1,217
5 50-99 svb	365	370	735
6 100-199 svb	391	165	556
7 200-499 svb	402	90	492
8 500-999 svb	198	22	220
Total	1,968	5,570	7,538

Tabelle 3: vorher

	Betriebsrat		
Anz. svpfl. Besch.	Ja	Nein	Gesamt
1 1-4 svb	64	2,461	2,525
2 5-9 svb	5*	84*	901
3 10-19 svb	130	762	892
4 20-49 svb	364	853	1,217
5 50-99 svb	365	370	735
6 100-199 svb	391	165	556
7 200-499 svb	402	90	492
8 500-999 svb	19*	2*	220
Total	1,968	5,570	7,538

Tabelle 4: nachher

Deutschland

	Betriebsrat		
Anz. svpfl. Besch.	Ja	Nein	Total
1 1-4 svb	107	3,841	3,948
2 5-9 svb	93	1,394	1,487
3 10-19 svb	219	1,249	1,468
4 20-49 svb	614	1,443	2,057
5 50-99 svb	620	615	1,235
6 100-199 svb	681	275	956
7 200-499 svb	685	155	840
8 500-999 svb	340	38	378
Total	3,359	9,010	12,369

Tabelle 5: vorher

	Betriebsrat		
Anz. svpfl. Besch.	Ja	Nein	Gesamt
1 1-4 svb	107	3,841	3,948
2 5-9 svb	93	1,394	1,487
3 10-19 svb	219	1,249	1,468
4 20-49 svb	614	1,443	2,057
5 50-99 svb	620	615	1,235
6 100-199 svb	681	275	956
7 200-499 svb	685	155	840
8 500-999 svb	340	38	378
Total	3,359	9,010	12,369

Tabelle 6: nachher

Bei der Ausgabe von Tabellen müssen die Forscherinnen und Forscher immer die Anzahl der Betriebe ausweisen. Denn wie viele Betriebe hinter dem Ergebnis stehen, ist für die Datenschutzprüfung von entscheidender Bedeutung. Das Beispiel 3 veranschaulicht die Problematik. Die Spalte 2 (sum) beinhaltet die Anzahl der übernommenen Auszubildenden in ausgewählten Wirtschaftszweigen, die Spalte 3 (N) die Anzahl der entsprechenden Betriebe. Hinter einer ausreichend großen Fallzahl an Auszubildenden können wenige Betriebe stecken. Ist dies der Fall, müssen beide Werte gelöscht werden. So wird auch verhindert, dass Großbetriebe identifiziert werden könnten.

Beispiel 3

r90b	sum	N
1 land-/forstwir	13	10
2 bergbau/gewinn	1	1
3 energie-/wasse	118	21
4 nahrung/genuss	130	32
5 textilien/lede	5	4
6 holzwaren/papi	35	13
7 chemische/phar	78	14
8 gummi/kunststo	164	23
9 glas, steine,	20	12
10 metallerzeugu	138	21
Total	702	151

Tabelle 7: vorher

r90b	sum	N
1 land-/forstwir	/	/
2 bergbau/gewinn	/	/
3 energie-/wasse	118	21
4 nahrung/genuss	130	32
5 textilien/lede	/	/
6 holzwaren/papi	/	/
7 chemische/phar	/	/
8 gummi/kunststo	164	23
9 glas, steine,	/	/
10 metallerzeugu	138	21
Total	702	151

Tabelle 8: nachher

Bei den Kennzahlen für ausgewählte Merkmale wird zum Beispiel bei Dummyvariablen der Mittelwert geprüft, da hier die Möglichkeit der Reidentifikation von kleinen Werten besteht. Im vorliegenden Beispiel 4 wird für die Variable r61 (Ausbildungsplätze angeboten: ja/nein) ein Mittelwert von .085 ausgewiesen. Dies entspricht für die Ausprägung 1 die prozentuale Verteilung von 8,57 %. Mittels Multiplikation der Fallzahl mit dem Mittelwert (140 x 0,857143) lässt sich berechnen, dass 12 Betriebe den Wert 1 haben.

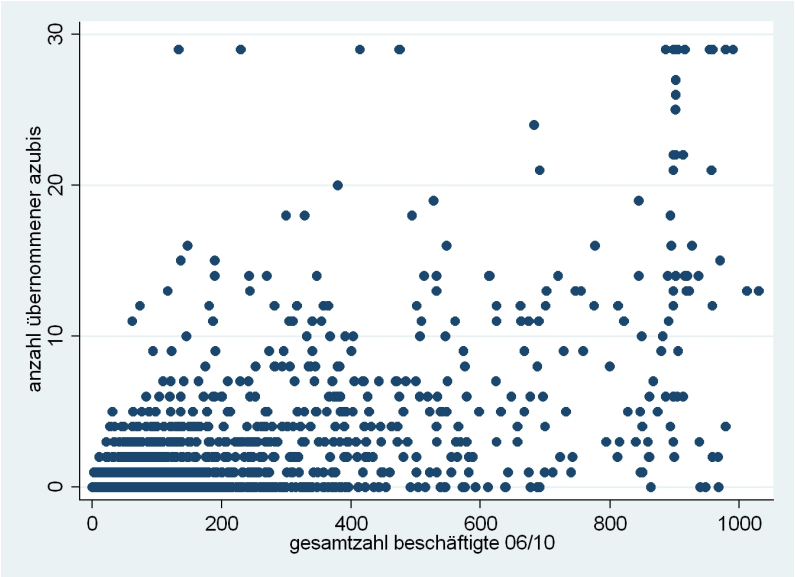
Variable	Obs	Mean	Std. Dev.	Min	Max
r60	201	2.373134	.9192794	1	3
r61	140	.0857143	.2809469	0	1
r62a	73	2.219178	2.340742	1	15

Tabelle 9: vorher

Variable	Obs	Mean	Std. Dev.	Min	Max
r60	201	2.373134	.9192794	1	3
r61	140	/	/	/	/
r62a	73	2.219178	2.340742	1	15

Tabelle 10: nachher

Das letzte Beispiel widmet sich der Weitergabe von Grafiken. Grafiken können nur weitergegeben werden, wenn einzelne Datenpunkte weniger als 20 Betrieben zugeordnet werden können. Die Prüfung der Grafiken erfolgt analog zu der Prüfung von Tabellen. Im vorliegenden Beispiel würde die Grafik nicht nach außen gegeben, da sich hinter jedem Datenpunkt ein Betrieb verbirgt.



5 Ausblick

Das FDZ arbeitet kontinuierlich an der Verbesserung der Bereitstellung von Daten unter Einhaltung des Datenschutzes für die Forscherinnen und Forscher im In- und Ausland. Die Daten des FDZ können zum Beispiel im Rahmen des Drittmmittelprojektes FDZ-in-FDZ (PFiFF) neben Nürnberg auch in den deutschen Städten Berlin, Bremen, Dresden und Düsseldorf an den FDZ der Statistischen Landesämter sowie am Michigan Center on the Demography of Aging (MICDA) im Institute for Social Research (ISR) an der University of Michigan im Rahmen eines Gastaufenthaltes analysiert werden. Darüber hinaus wird im EU-Projekt Data without Boundaries (DwB) daran gearbeitet, den Zugang zu Mikrodaten für die Forscherinnen und Forscher EU-weit zu erleichtern. Im FDZ sind wir dabei die Datenfernverarbeitung zu automatisieren und arbeiten am Einsatz der Software JoSuA[23] des International Data Service Center (IDSC) des des Instituts zur Zukunft der Arbeit (IZA).

[23] Weiterführende Informationen zu JoSuA finden sich unter: `http://idsc.iza.org/josua`

Literatur

Brandt, Maurice/Franconi, Luisa/Guerke, Christopher/Hundepool, Anco/Lucarelli, Maurizio/Mol, Jan/Ritchie, Felix/Seri, Giovanni/Welpton, Richard. Guidelines for the checking of output based on microdata research. Final report of ESSnet sub-group on output SDC 2010

Bundesagentur für Arbeit. Statistische Geheimhaltung: Rechtliche Grundlagen und fachliche Regelungen der Statistik der Bundesagentur für Arbeit. März 2012, abgerufen am 18.05.2012 ⟨URL: http://statistik.arbeitsagentur.de/Statischer-Content/Grundlagen/Statistische-Geheimhaltung/Generische-Publikationen/Statistische-Geheimhaltung.pdf⟩

Bundesstatistikgesetz (BStatG) – Gesetz über die Statistik für Bundeszwecke vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 3 des Gesetzes vom 7. September 2007 (BGBl. I S. 2246).

BVerfG. Urteil v. 15.12.1983, Az. 1 BvR 209, 269, 362, 420, 440, 484/83.

Kirchner, Antje/Gschwind, Lutz. Panel Arbeitsmarkt und soziale Sicherung - Die PASS Campus Files. Datensätze für den Einsatz in der wissenschaftlichen Lehre. FDZ-Methodenreport 06/2011 2011

Lane, Julia/Heus, Pascal/Mulcahy, Tim. Data Acess in a Cyber World: Making Use of Cyberinfrastructure. Transactions on Data Privacy 2008

Müller, Walter/Blien, Uwe/Knoche, Peter/Wirth, Heike. Die faktische Anonymität von Mikrodaten. Stuttgart: Metzler-Poeschel. 1991

Ritchie, Felix. Statistical disclosure detection and control in a research environment. WISERD DATA RESOURCES 006 2011

SGB X. Zehntes Buch Sozialgesetzbuch – Sozialverwaltungsverfahren und Sozialdatenschutz – (SGB X), in der Fassung der Bekanntmachung vom 18. Januar 2001 (BGBl. I S. 130), zuletzt geändert durch Entscheidung des Bundesverfassungsgerichts vom 23. November 2010 (BGBl. I S. 1718).

Sozialgesetzbuch (SGB) Erstes Buch (I) – Allgemeiner Teil (SGB I) vom 11. Dezember 1975 (BGBl. I S. 3015), zuletzt geändert durch Artikel 110 Absatz 5 des Gesetzes über die weitere Bereinigung von Bundesrecht vom 8. Dezember 2010 (BGBl. I S. 1864.

Sozialgesetzbuch (SGB) Drittes Buch (III) – Arbeitsförderung (Artikel 1 des Gesetzes vom 24. März 1997, BGBl. I S. 594), zuletzt geändert durch Artikel 12 Absatz 8 des Gesetzes vom 24. März 2011 (BGBl. I S. 453).

Tabelle 11: FDZ Portfolio

	Allgemein	faktisch anonymisierte Daten	schwach anonymisierte Daten
Voraussetzungen	Wissenschaftliche Forschung und Erforderlichkeit der Daten	Arbeitsmarkt- und Berufsforschung	Forschung im Sozialleistungsbereich
		Unabhängige wissenschaftliche Forschungseinrichtung	-
		Datensicherheitskonzept	-
		-	Öffentliches Interesse
		-	Genehmigung des BMAS
Zugang und Nutzung	Zweckbindung Zeitliche Befristung	Gewährleistung der Datensicherheit	-
		-	Richtlinien für Gastaufenthalte
	Verbote zur Weitergabe an Dritte, zur Zuspielung anderer Mikrodaten sowie zur Deanonymisierung	Löschung der Mikrodaten nach Projektende	-
	Begrenzung des Personenkreises	-	Verbot des Rückrechnens gelöschter Werte
Anonymisierung	Stichprobenziehung Löschen von Originalidentifikatoren Löschen oder Vergröbern sensibler Merkmale	Löschen oder Vergröbern weiterer Merkmale	-
Ergebniskontrolle	-	-	Absolute Anonymisierung der Ergebnisse

Impressum

FDZ-Methodenreport 06/2012

Herausgeber
Forschungsdatenzentrum (FDZ)
der Bundesagentur für Arbeit
im Institut für Arbeitsmarkt- und Berufsforschung
Regensburger Str. 104
90478 Nürnberg

Redaktion
Stefan Bender, Iris Dieterich

Technische Herstellung
Iris Dieterich

Bezugsmöglichkeit
http://doku.iab.de/fdz/reporte/2012/MR_06-12.pdf

Internet
http://fdz.iab.de/

Rückfragen zum Inhalt an:
Alexandra Schmucker
Tel: 0911 / 179-1762
E-Mail: alexandra.schmucker@iab.de

Dana Müller
Tel: 0911 / 179-2409
E-Mail: dana.mueller@iab.de

Forschungsdatenzentrum (FDZ)
Regensburger Str. 104, 90478 Nürnberg