Sonderdruck aus:

Mitteilungen
aus der
Arbeitsmarkt- und
Berufsforschung

Stefan Bender und Jürgen Hilzendegen

Die IAB-Beschäftigtenstichprobe als scientific use file

28. Jg./1995

1

Mitteilungen aus der Arbeitsmarkt- und Berufsforschung (MittAB)

Die MittAB verstehen sich als Forum der Arbeitsmarkt- und Berufsforschung. Es werden Arbeiten aus all den Wissenschaftsdisziplinen veröffentlicht, die sich mit den Themen Arbeit, Arbeitsmarkt, Beruf und Qualifikation befassen. Die Veröffentlichungen in dieser Zeitschrift sollen methodisch, theoretisch und insbesondere auch empirisch zum Erkenntnisgewinn sowie zur Beratung von Öffentlichkeit und Politik beitragen. Etwa einmal jährlich erscheint ein „Schwerpunktheft", bei dem Herausgeber und Redaktion zu einem ausgewählten Themenbereich gezielt Beiträge akquirieren.

Hinweise für Autorinnen und Autoren

Das Manuskript ist in dreifacher Ausfertigung an die federführende Herausgeberin
Frau Prof. Jutta Allmendinger, Ph. D.
Institut für Arbeitsmarkt- und Berufsforschung
90478 Nürnberg, Regensburger Straße 104
zu senden.

Die Manuskripte können in deutscher oder englischer Sprache eingereicht werden, sie werden durch mindestens zwei Referees begutachtet und dürfen nicht bereits an anderer Stelle veröffentlicht oder zur Veröffentlichung vorgesehen sein.

Autorenhinweise und Angaben zur formalen Gestaltung der Manuskripte können im Internet abgerufen werden unter http://doku.iab.de/mittab/hinweise_mittab.pdf. Im IAB kann ein entsprechendes Merkblatt angefordert werden (Tel.: 09 11/1 79 30 23, Fax: 09 11/1 79 59 99; E-Mail: ursula.wagner@iab.de).

Redaktion

Ulrike Kress, Gerd Peters, Ursula Wagner, in: Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit (IAB), 90478 Nürnberg, Regensburger Str. 104, Telefon (09 11) 1 79 30 19, E-Mail: ulrike.kress@iab.de: (09 11) 1 79 30 16, E-Mail: gerd.peters@iab.de: (09 11) 1 79 30 23, E-Mail: ursula.wagner@iab.de: Telefax (09 11) 1 79 59 99.

Herstellung

Satz und Druck: Tümmels Buchdruckerei und Verlag GmbH, Gundelfinger Straße 20, 90451 Nürnberg

Verlag

W. Kohlhammer GmbH, Postanschrift: 70549 Stuttgart: Lieferanschrift: Heßbrühlstraße 69, 70565 Stuttgart: Telefon 07 11/78 63-0; Telefax 07 11/78 63-84 30: E-Mail: waltraud.metzger@kohlhammer.de, Postscheckkonto Stuttgart 163 30.
Girokonto Städtische Girokasse Stuttgart 2 022 309.
ISSN 0340-3254

Bezugsbedingungen

Die „Mitteilungen aus der Arbeitsmarkt- und Berufsforschung" erscheinen viermal jährlich. Bezugspreis: Jahresabonnement 52,- € inklusive Versandkosten: Einzelheft 14,- € zuzüglich Versandkosten. Für Studenten, Wehr- und Ersatzdienstleistende wird der Preis um 20 % ermäßigt. Bestellungen durch den Buchhandel oder direkt beim Verlag. Abbestellungen sind nur bis 3 Monate vor Jahresende möglich.

Zitierweise:

MittAB = „Mitteilungen aus der Arbeitsmarkt- und Berufsforschung" (ab 1970)
Mitt(IAB) = „Mitteilungen" (1968 und 1969)
In den Jahren 1968 und 1969 erschienen die „Mitteilungen aus der Arbeitsmarkt- und Berufsforschung" unter dem Titel „Mitteilungen", herausgegeben vom Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit.

Internet: http://www.iab.de

Die IAB-Beschäftigtenstichprobe als scientific use file

*Stefan Bender und Jürgen Hilzendegen**

Die seit 1973 aufgebaute Beschäftigtenstatistik ist neben der Volks- und Berufszählung, dem Mikrozensus und der Einkommens- und Verbrauchsstichprobe (EVS) eine der für die Sozialforschung wichtigsten Datenquellen der deutschen Sozialstatistik. Angesichts der Bedeutung, die die Beschäftigtenstatistik als Datenquelle für die Arbeitsmarktforschung hat, haben das Wissenschaftszentrum für Sozialforschung in Berlin (WZB), das Zentrum für Umfragen, Methoden und Analysen in Mannheim (ZUMA) und das IAB ein gemeinsames Projekt durchgeführt, mit dem Ziel, diese Daten nach dem Konzept der faktischen Anonymität und soweit datenschutzrechtlich möglich, allen interessierten Forschern über das Zentralarchiv in Köln (ZA) zugänglich zu machen (scientific use file). Die Basis für die Erstellung der anonymisierten Datei bildet eine l-Prozent-Stichprobe aus der Historikdatei (IAB-Beschäftigtenstichprobe), die für den Zeitraum 1975-1990 tagesgenaue Verlaufsinformationen von etwa 430.000 Sozialversicherungspflichtig Beschäftigten enthält. Die IAB-Beschäftigtenstichprobe wird um zusätzliche Merkmale für Betriebe, sowie über den Bezug von Lohnersatzleistungen (z.B. Arbeitslosengeld) ergänzt.

Bei den Anonymisierungsmaßnahmen orientierte man sich eng an den Ergebnissen des Anonymisierungsprojektes zur Umsetzung der faktischen Anonymität beim Mikrozensus und der EVS. Die statistischen Ämter haben im Anschluß an dieses Projekt Regeln für die Weitergabe dieser Mikrodaten an die Wissenschaft beschlossen, die dem im Bundesstatistikgesetz von 1987 definierten Konzept der faktischen Anonymität entsprechen. Faktische Anonymität bedeutet, daß die Einzelangaben nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft den Befragten einer Erhebung zugeordnet werden können. Dieses Konzept garantiert weiterhin einen ausreichenden Schutz der Befragten, ohne daß das Analysepotential der Daten durch zu starke Anonymisierungsmaßnahmen eingeschränkt werden muß. Damit besteht für die empirische Sozialforschung die Möglichkeit, sehr informations- und umfangreiche Daten der amtlichen Statistik für Sekundäranalysen zu verwenden.

Die für den Mikrozensus und die EVS geltenden Anonymisierungsmaßnahmen konnten für die Querschnittsangaben der Personen weitestgehend übernommen werden. Da die IAB-Beschäftigtenstichprobe aber auch Informationen über Betriebe als weitere schützenswerte Einheit, sowie Verlaufsangaben über Beschäftigte und Betriebe enthält, waren hier neue Wege zu beschreiten.

Gliederung

* Dipl.-Soz. Stefan Bender und Dipl.-Sozialwirt Jürgen Hilzendegen sind Wiss. Mitarbeiter im IAB. Wir danken Herrn John für umfassende Beratungen im Datenmanagement (SIMAT), Herrn Majer für Hilfen bei der Hardware (BS2000) und Herrn Gommlich für den Datentransfer von BS2000 in DOS. Desweiteren bedanken wir uns für die vielen hilfreichen Anmerkungen der Mitglieder des Projektbeirats, insbesondere bei Herrn Schimpl-Neimanns (ZUMA). Vielfältige Hilfen, Kommentare und Anregungen bekamen wir im Laufe des Projekts von Kollegen aus dem IAB, namentlich Petra Beckmann, Uwe Blien, Hans Dietrich, Frido Dietz, Werner Karr, Udo Lehmann, Susanne Kohaut und ganz besonders bei Helmut Rudolph. Inhaltliche Fehler und Unzulänglichkeiten liegen in der Verantwortung der Autoren.

0 Einleitung

Das IAB wird häufig von Wissenschaftlern unterschiedlicher Fachrichtungen um Überlassung von Mikrodaten aus der Beschäftigtenstatstik gebeten, kann aber diesen Wünschen nicht entsprechen, u.a. aus Gründen des Datenschutzes. Angesichts der Bedeutung, die die Beschäftigtenstatistik als Datenquelle für die Arbeitsmarktforschung hat (vgl. Alba/Müller/ Schimpl-Neimanns 1994), führten das IAB, das Zentrum für Umfragen, Methoden und Analysen in Mannheim (ZUMA) und das Wissenschaftszentrum Berlin für Sozialforschung (WZB) ein gemeinsames Projekt durch[1], aus diesen Daten einen scientific use herzustellen, der über das Zentralarchiv in Köln (ZA) der Wissenschaft zugänglich ist.

Nach §16 Abs.6 BStatG müssen Daten, die an die Wissenschaft übermittelt werden, faktisch anonym sein. Die faktische Anonymität ist dann gegeben, wenn ein Datenangreifer unverhältnismäßig viel Zeit, Kosten und Arbeitskraft aufbringen muß, um einen Datensatz zu deanonymisieren. Auch bei anonymisierten Daten kann daher eine spätere Deanonymisierung nicht mit Sicherheit ausgeschlossen werden. Die Definition der faktischen Anonymität impliziert, daß für jeden einzelnen Datensatz – oder zumindest für jede Klasse von Datensätzen – das „Unverhältnismäßigkeitskriterium“ neu bestimmt werden muß. So müssen praktisch für jeden zu anonymisierenden Datensatz spezifische Anonymisierungsmaßnahmen abgeleitet und vorgenommen werden.

Daher wurden die Arbeiten des Projekts von einem wissenschaftlichen Beirat begleitet[2], der die Konzeption der Anonymisierung und den verbleibenden Nutzen der Daten für Forschungszwecke begutachtete, sowie den erreichten Grad der faktischen Anonymisierung beurteilte. Er hat den Verlauf des Projektes durch vielfältige Anregungen geprägt.

Gegenstand des Projektes ist die zur Verfügung stehende l %-Stichprobe aus der Historikdatei der Bundesanstalt für Arbeit (IAB-Beschäftigtenstichprobe), die in Anlehnung an die faktische Anonymisierung des Mikrozensus und der Einkommens- und Verbrauchsstichprobe (EVS)[3] anonymisiert wird. Die IAB-Beschäftigtenstichprobe deckt einen auswertbaren Zeitraum von 16 Jahren (01.01.1975 – 31.12.1990) ab und umfaßt in den Jahresquerschnitten jeweils etwa 200.000 und im Längsschnitt etwa 430.000 Sozialversicherungspflichtig Beschäftigte. Von diesem Personenkreis liegen tagesgenaue Verlaufsinformationen ihrer Sozialversicherungspflichtigen Beschäftigungen und dem Bezug von Arbeitslosengeld, -hilfe und Unterhaltshaltgeld vor. Für eine allgemeine Nutzung wird die IAB-Beschäftigtenstichprobe aufbereitet und um zusätzliche Merkmale für Betriebe sowie über den Bezug von Lohnersatzleistungen ergänzt.

Die IAB-Beschäftigtenstichprobe ist ein prozeßproduzierter Datensatz und unterscheidet sich daher von den meisten sozialwissenschaftlichen Umfragen. Daher erscheint eine ausführliche Darstellung der Grundlagen der IAB-Beschäftigtenstichprobe – die Beschäftigtenstatistik (Kap. l) und die Historikdatei (Kap.2) mit ihren Besonderheiten – notwendig. Bei der Beschreibung der Stichprobe steht die Verbindung der Angaben der Sozialversicherungspflichtig Beschäftigten mit zusätzlichen Informationen über Betriebscharakteristika aus der Beschäftigtenstatistik sowie mit Daten der Leistungsempfängerdatei im Mittelpunkt (Kap. 3).

Den größten Raum nimmt die Darstellung der verschiedenen Anonymisierungsmaßnahmen der IAB-Beschäftigtenstichprobe ein (Kap. 4 und 5). Die Anonymisierungsmaßnahmen für die IAB-Beschäftigtenstichprobe beziehen sich auf drei unterschiedliche Ebenen, nämlich die der Personen, der Betriebe und der Längsschnittinformationen. Bei der Anonymisierung der Personenangaben (Kap. 4.2) kann weitestgehend auf die Ergebnisse zur faktischen Anonymität von Mikrodaten (vgl. Müller et al. 1991) zurückgegriffen werden. Bei der Anonymisierung der Betriebsangaben (Kap. 4.3) müssen neue Wege beschritten werden, da zum einen alle Großbetriebe in der IAB-Beschäftigtenstichprobe enthalten sind, zum anderen einige dieser Großbetriebe in spezifischen Wirtschaftszweigen durch einfache Größenvergleiche reidentifiziert werden könnten. Bei der Längsschnittanonymisierung (Kap. 5) wird darauf geachtet, daß die spezifischen Charakteristika der IAB-Beschäftigtenstichprobe erhalten bleiben. Hierzu wird der gesamte Erwerbsverlauf eines sozialversicherungspflichtig Beschäftigten um einen konstanten Betrag verschoben. Dadurch werden die jeweiligen (Verbleibs-)dauern in spezifischen Zuständen (z.B Betriebszugehörigkeit) von der Anonymisierung nicht berührt.

l Die Beschäftigtenstatistik

1.1 Das Meldeverfahren

Grundlage der Beschäftigtenstatistik[4] ist das mit Wirkung vom 1. Januar 1973 eingeführte integrierte Meldeverfahren zur Kranken-, Renten- und Arbeitslosenversicherung (DEVO/DÜVO[5]). Dieses Meldeverfahren verlangt, daß die Arbeitgeber für alle Sozialversicherungspflichtig beschäftigten Arbeitnehmer innerhalb bestimmter Fristen Meldungen in einheitlicher und datenverarbeitungsgerechter Form[6] an die Sozialversicherungsträger abgeben (vgl. Statistisches Bundesamt 1992:6). Meldungen sind vorgeschrieben bei Beginn einer Beschäftigung (Meldefrist:2 Wochen), Ende einer Beschäftigung (Meldefrist:6 Wochen) und aus Anlässen, die bestimmte Änderungen des Beschäftigungsverhältnisses bewirken[7] (z.B. Unterbrechungen[8] für mindestens einen Monat). Neben diesen drei Arten von „Veränderungs“meldungen ist zum Ende des Kalenderjahres für jeden Sozialversicherungspflichtigen Arbeitnehmer, der zum 31. Dezember eines Jahres beschäftigt ist (vgl. Hoffmann/Wermter 1976:33), eine Jahresmeldung[9] vorgeschrieben (Meldefrist: drei Monate). Den Jahresmeldungen kommt hierbei eine besondere Rolle zu, „denn das Meldeverfahren ist so konzipiert, daß ein Be-

---

[1] Projekt 6-466: Faktische Anonymisierung der Beschäftigtenstichprobe.

[2] Dem Beirat gehörten folgende Personen an: Knut Gerlach (Universität Hannover), Roland Habich (WZB, Berlin), Werner Karr (IAB, Nürnberg), Peter Mohler (ZUMA, Mannheim), Walter Müller (Universität Mannheim), Erwin Rose (ZA, Köln), Werner Schmidt (Bundesbeauftragter für den Datenschutz, Bonn) und Rolf Ziegler (Universität München).

[3] Aus Gründen der besseren Lesbarkeit wird im folgenden nur noch von der Anonymisierung des Mikrozensus gesprochen.

[4] Der gesetzliche Auftrag für die Durchführung der Beschäftigtenstatistik ist im Arbeitsförderungsgesetz (AFG) vom 25.6.1969 verankert.

[5] DEVO (Datenerfassungsverordnung) – Verordnung über die Erfassung von Daten für die Träger der Sozialversicherung und für die Bundesanstalt für Arbeit vom 24.11.1972 (BGB1.1:2159ff.).
DÜVO (Datenübertragungsverordnung) – Verordnung über die Datenübermittlung auf maschinell verwertbaren Datenträgern im Bereich der Sozialversicherung und der Bundesanstalt für Arbeit vom 18.12.1972 (BGB1. 1:2482 ff.).

[6] 2. DEVO vom 29.5.1980 (BGB1.1:593ff). 2. DÜVO vom 29.5.1980 (BGB1. 1:616 ff).

[7] Anmeldung (§ 3, 2. DEVO); Abmeldung (§ 4, 2. DEVO); Änderungen (§ 6, 2. DEVO)

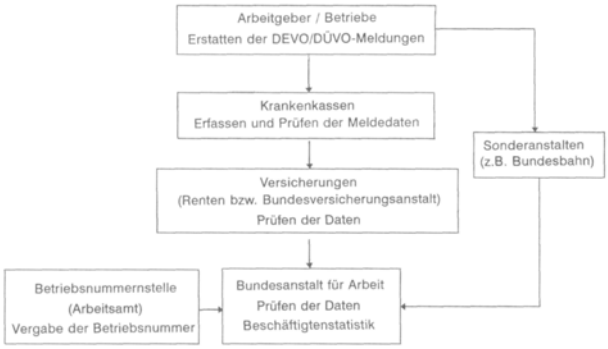
[8] „Unterbrechungsperioden sind Zeitspannen, in denen ein Beschäftigungsverhältnis gegen Entgelt faktisch nicht besteht, wohl aber ein Weiterbeschäftigungsanspruch nach Ableistung des Wehr- oder Ersatzdienstes, nach dem Mutterschaftsurlaub, nach längerer Krankheit oder – unter bestimmten Bedingungen – nach Schlechtwetterperioden im Baubereich“ (Cramer/Majer 1991:82)

[9] § 5, 2. DEVO

schäftigungsverhältnis im Zeitablauf lückenlos durch Jahresmeldungen bestätigt oder durch eine Abmeldung beendet werden muß.“ (Wermter/Cramer 1988:479)

Der Meldeweg aller o.g. Meldungen läßt sich wie folgt skizzieren: Alle Meldungen werden von den Arbeitgebern an die zuständigen Krankenkassen geleitet, die dann die Daten an die Datenstelle der Rentenversicherung in Würzburg (für die Arbeiter) oder an die Bundesversicherungsanstalt für Angestellte in Berlin (für die Angestellten) weitergeben. Diese wiederum führen bestimmte Datenteile der Bundesanstalt für Arbeit zu, die den einzelnen Datensatz über eine vergebene Nummer des beschäftigten Betriebs (Betriebsnummer) um zusätzliche Merkmale erweitert (vgl. Abbildung l)[10].

Abbildung 1: Schematische Darstellung des Meldeverfahrens



1.2 Die Quartalsdatei und das Jahreszeitraummaterial

Die Bundesanstalt legt für jeden Versicherten unter seiner Sozialversicherungsnummer ein sogenanntes Versichertenkonto an. In diesem sind alle Meldungen zu der jeweiligen Versicherungsnummer in chronologischer Reihenfolge nach dem Wirksamkeitsdatum gespeichert. Grundsätzlich gilt, daß keine Daten ohne Versicherungsnummer übernommen und weitergeleitet werden. Aus den Meldungen eines Jahres baut die Bundesanstalt vierteljährlich (Termine: 31.3, 30.6, 30.9, 31.12) eine Stichtagsdatei auf, sowie eine Jahresdatei aus den Meldungen eines Jahres.

Sechs Monate“ nach einem Berichtsstichtag werden Auswertungen zur Ermittlung des Beschäftigtenbestandes durchgeführt. Dabei wird jedes Versichertenkonto daraufhin geprüft, ob der betreffende Versicherte am Stichtag in einem Beschäftigungsverhältnis steht (zum Verfahren vgl. Wermter/Cramer 1988:478). Diese herausgefilterten Personen stellen die Grundlage für die Erstellung der Bestandsergebnisse dar. Die sechsmonatige Wartefrist bis zur Auszählung stellt einen Kompromiß zwischen Aktualität und Vollständigkeit des Datenmaterials dar und führt zur Anwendung des sogenannten Abschneideverfahrens (vgl. Kap. 1.5.2).

Beim Jahreszeitraummaterial wird ausschließlich auf Jahresmeldungen, Abmeldungen und Unterbrechungsmeldungen mit den zusätzlichen Informationen über Beginn und Beendigung der Beschäftigung, sowie über das seit Jahresbeginn bis zum Meldedatum gezahlte beitragspflichtige Bruttoentgelt eines Kalenderjahres zurückgegriffen. Die Beobachtungseinheit ist hier nicht die beschäftigte Person, sondern der Beschäftigungsfall[12]. Es ist daher möglich, daß eine Person zu einem bestimmten Stichtag, die mehrere Beschäftigungsphasen oder -Verhältnisse hatte, mehrfach im Datensatz enthalten ist (vgl. Mayer/Becker 1984:997). Dies ist bei der Stichtagsdatei explizit ausgeschlossen (vgl. Wermter/Cramer 1988:478).

Mit der Erstellung des Jahreszeitraummaterials müßte theoretisch so lange gewartet werden, bis alle Meldungen des betreffenden Jahres bei der Bundesanstalt für Arbeit eingegangen sind. Praktisch wird allerdings das Jahreszeitraummaterial etwa zwei Jahre nach Ende des Auswertungsjahres erstellt. In diesem Zeitraum sind ca. 98% bis 100% der für das Berichtsjahr relevanten Meldungen bei der Bundesanstalt für Arbeit eingegangen. Das Jahreszeitraummaterial liegt seit 1975 vor (vgl. Statistisches Bundesamt 1992:8f).

1.3 Definitions- und Abgrenzungsprobleme

Die Aussagekraft der Beschäftigtenstatistik hängt, soweit es um die Beschäftigungsverhältnisse geht, wesentlich von der jeweiligen Definition einer Sozialversicherungspflichtigen Beschäftigung ab (vgl. § 2 Abs. l, 2. DEVO). Die Grundgesamtheit[13] umfaßt dadurch Arbeiter und Angestellte und alle sich in einem betrieblichen Ausbildungsverhältnis befindenden Personen, soweit sie nicht aufgrund eines beamtenrechtlichen Dienstverhältnisses von der Sozialversicherungspflicht befreit sind. Nicht in der Beschäftigtenstatistik enthalten sind daher u.a. geringfügig Beschäftigte (vgl. § 8 SGB IV)[14], ordentlich Studierende und mithelfende Familienangehörige. Sozialversicherungspflichtig, jedoch nicht in der Beschäftigtenstatistik enthalten, sind u.a. Rehabilitanden, krankenversicherungspflichtige Selbständige, Empfänger von Vorruhestandsgeld, rentenversicherungspflichtige Selbständige und Wehr- oder Ersatzdienstleistende[15] (vgl. hierzu Cramer 1985:59f.).

Nach Herberger und Becker (1983) umfaßt die Beschäftigtenstatistik 1980 79,0%[16] aller Erwerbstätigen, hierbei ist der Deckungsgrad der Beschäftigtenstatstik in den einzelnen Berufen und Wirtschaftszweigen sehr unterschiedlich. So sind für 1980 gerade 56,9% aller Beschäftigten in der Wirtschaftsabteilung „Gebietskörperschaften und Sozialversicherung“ Sozialversicherungspflichtig, dagegen aber 98,2% in der „Energie- und Wasserversorgung“ (vgl. Herberger/Becker 1983:299-300). Neuere Zahlen, die Aufschlüsse über den Deckungsgrad der tatsächlichen Anzahl sozialversicherungspflichtig Beschäftigter geben, können durch Angaben aus dem IAB-Betriebspanel[17] (vgl. Projektgruppe Betriebspanel 1994) abgeschätzt werden[18] (vgl. Infratest Sozialforschung 1994). So schwankt hier der Deckungsgrad zwischen 58,9% im Wirtschaftsbereich „Landwirtschaft“ und 100% in „Bergbau/Energie/Wasser“.

Eine Sonderrolle nehmen Unterbrechungen des Beschäftigtenverhältnisses ein, bei denen dieses rechtlich bestehen

---

[10] Eine ausführliche Darstellung des Meldeverfahrens (2.DEVO/DÜVO) findet sich in Wermter/Cramer (l 988:471 ff.).

[11] Zur Problematik des Abschneidens vgl. Kap. 1.5.2

[12] Eine Definition für den Beschäftigungsfall findet sich bei Mayer/Becker (1984:997).

[13] vgl. auch die Definition im § 7 SGB IV für Beschäftigte.

[14] Für eine Definition vgl. Herberger/Becker 1983:293.

[15] Wehr- und Ersatzdienstleistende sind nur dann sozialversicherungspflichtig, wenn sie ihren Dienst aus einem weiterhin bestehenden Beschäftigtenverhältnis heraus angetreten haben.

[16] Es wird in der Literatur oft auch von 75% gesprochen (vgl. Hoffmann/Wermter 1976:33, Mayer/Becker 1984:995).

[17] Für die Informationen über das Betriebspanel bedanken sich die Autoren bei Lutz Bellmann und Susanne Kohaut (beide IAB/VH-5).

[18] Eine weitere denkbare Möglichkeit wäre ein Vergleich mit der Arbeitsstättenzählung 1987. Allerdings besitzen die beiden Datenbestände unterschiedliche Wirtschaftszweigklassifikationen.

bleibt, faktisch aber die Arbeit ruht und auch kein Entgelt bezahlt wird. Unterbrechungstatbestände sind Krankheiten nach Ende der Lohnfortzahlung, Mutterschaftsurlaub, Erziehungsurlaub, Wehr- und Zivildienst und Streik. Die gesetzlichen Regelungen, von denen die Dauer der Unterbrechungszeiten abhängen, wurden mehrfach geändert.

Der Sachverhalt, daß ruhende Beschäftigungsverhältnisse in der Beschäftigtenstatistik enthalten sind, führt zu Doppelzählungen desselben Arbeitsverhältnisses, wenn die Stelle in dieser Zeit etwa durch eine Aushilfskraft besetzt wird. Diese Problematik hat einiges Gewicht, da zu einem Stichtag mehrere Hunderttausend Erwerbsunterbrecher immer mitgezählt werden. Außerdem steht für eine Unterbrechung die Dauer des ruhenden Beschäftigungsverhältnisses nicht immer eindeutig fest (vgl. Cramer 1985:61).

Tabelle 1: Anteil der sozialversicherungspflichtig Beschäftigten an allen Beschäftigten nach Angaben im IAB-Betriebspanel (in %)

Wirtschaftsklasse		30.06.92	31.12.1992
01	Landwirtschaft	58,9	56,8
02	Bergbau/Energie/Wasser	100,0	99,3
03-06	Grundstoffverarbeitung	95,3	94,8
07-12	Investitionsgüter	94,1	94,0
13-16	Verbrauchsgüter	85,2	84,0
17-18	Baugewerbe	86,4	86,7
19-20	Handel/Verkehr	75,9	75,0
21-22	Versicherung	94,4	94,2
23-25	Gaststätten	60,2	59,1
26-27	Verlage	72,3	71,3
28	Gesundheitswesen	81,2	81,9
29-35	Rechtsbereich	71,5	71,9
36-38	Organisationen o. Erwerbscharakter	79,8	80,7
39-41	Gebietskörperschaften	65,0	61,9
Durchschnitt		80,5	79,8

1.4 Verfügbare Merkmale aus dem Meldeverfahren

Jeder Versicherte besitzt eine Sozialversicherungsnummer, die den Rententräger, das Geburtsdatum, das Geschlecht und den Anfangsbuchstaben des Nachnamens des Versicherten enthält. Die Arbeitgeber müssen Beginn- und Enddatum des Beschäftigungsverhältnisses angeben, das von den Krankenkassen auf Plausibilität geprüft wird. Von den Arbeitsämtern wird jedem Betrieb eine Betriebsnummer zugeteilt. In der Bundesanstalt für Arbeit wird anhand der Betriebsnummer die Wirtschaftsklasse[19] (dreistellig), eine Regionalziffer und die fünfstellige Dienststellennummer der Arbeitsverwaltung zugespielt.

Die Definitionen der einzelnen Merkmale sind institutionell, meist gesetzlich, festgelegt. „Diesen Legaldefinitionen liegen in der Regel keine volkswirtschaftlichen Begriffsbestimmungen zugrunde“ (Schmähl/Fachinger 1994:188).

Die verfügbaren Merkmale kann man in zwei Kategorien unterteilen. Sie basieren auf Angaben, die vorwiegend versicherungsrechtlichen Zwecken dienen und auf Angaben, die nur statistische Informationen enthalten. Die versicherungsrechtlichen Merkmale – die *Versicherungsnummer,* die *Beschäftigungszeit* und das *Versicherungspflichtige Entgelt* - werden von den Arbeitgebern und den Versicherten geprüft, so daß diese Angaben eine hohe Genauigkeit besitzen. Bei den statistischen Angaben muß unterschieden werden zwischen den Merkmalen, die dem beschäftigten Betrieb des Versicherten zugeordnet werden (Wirtschaftszweig und regionale Gliederung) und den Merkmalen, die sich unmittelbar auf den Versicherten beziehen. Letztere werden von den Arbeitgebern gemeldet (vgl. Cramer 1985:62).

Insgesamt sind folgende Merkmale für jeden Versicherten in der Beschäftigtenstatistik verfügbar (vgl. Statistisches Bundesamt 1992:9-12):

- *Geschlecht und Geburtsjahr*

Beide Merkmale sind Bestandteil der Versicherungsnummer.

- *Staatsangehörigkeit*

Deutsche sind nach Art. 116 Abs. l des Grundgesetzes definiert.

- *Ausbildung*

Diese umfaßt die Angaben zur erreichten Schul- und abgeschlossenen Berufsausbildung. Als abgeschlossene Berufsausbildung gilt ein Abschluß in einem anerkannten Lehr- oder Anlernberuf, einer Berufsfach- oder Fachschule, einer Fachhochschule oder wissenschaftlichen Hochschule. Die Kategorien der Schulbildung sind „Haupt-/Realschule“, „Fachhochschule“, „Hochschule“ und „Ausbildung unbekannt“.

- *Beruf (ausgeübte Tätigkeit)*

Die Berufsbezeichnung bezieht sich unmittelbar auf die ausgeübte Tätigkeit. Die Zuordnung des sozialversicherungspflichtig Beschäftigten zu dem Merkmal ‘Berufsordnung’erfolgt nach dem Schlüsselverzeichnis der Bundesanstalt für Arbeit (Klassifizierung der Berufe – Systematisches und alphabetisches Verzeichnis der Berufsbenennungen, Ausgabe 1975). Die Berufsordnung besteht aus einer dreistelligen Kennziffer und umfaßt 328 Berufe[20]. Die in einer Kennziffer zusammengefaßten Berufe sind dem Wesen ihrer Berufsaufgabe und Tätigkeit nach gleichartig. Die ersten zwei Ziffern der Kennziffer fassen die fachlich (nach Berufsaufgabe und Tätigkeit) näher zueinander gehörenden Berufe zusammen; diese zweistellige Kennziffer wird als Berufsgruppe bezeichnet. Insgesamt umfaßt die Systematik 86 Berufsgruppen.

- *Stellung im Beruf (darunter Voll-/Teilzeitbeschäftigung)*

Die Differenzierung nach Arbeitern und Angestellten erfolgt nach der Zugehörigkeit zu dem jeweiligen Träger der Rentenversicherung. Bei den Facharbeitern sind auch Meister und Poliere enthalten, wenn sie in der Arbeiterrentenversicherung pflichtversichert sind.

Zu der Kategorie „Beschäftigte in beruflicher Ausbildung“ zählen neben den Auszubildenden[21] auch Anlernlinge, Prak

[19] Der Wirtschaftszweig wird in Zusammenarbeit mit dem Betrieb vergeben.

[20] Für die Bundesanstalt für Arbeit 334 Berufe.

[21] s. Berufsbildungsgesetz vom 14. August 1969 (BBiG)

Tabelle 2: Beitragsbemessungsgrenzen (DM im Monat)

Jahr	Rentenversicherung der Arbeiter und Angestellten	Knappschaftliche Rentenversicherung	Krankenversicherung	Geringfügigkeitsgrenze
1975	2.800	3.400	2.100	350
1976	3.100	3.800	2.325	387,5
1977	3.400	4.200	2.550	425 (370*)
1978	3.700	4.600	2.775	390
1979	4.000	4.800	3.000	390
1980	4.200	5.100	3.150	390
1981	4.400	5.400	3.300	390
1982	4.700	5.800	3.525	390
1983	5.000	6.100	3.750	390
1984	5.200	6.400	3.900	390
1985	5.400	6.700	4.050	400
1986	5.600	6.900	4.200	410
1987	5.700	7.100	4.275	430
1988	6.000	7.300	4.500	440
1989	6.100	7.500	4.575	450
1990	6.300	7.800	4.725	470

* Ab 1.7.1977

tikanten, Volontäre, Schüler an Schulen des Gesundheitswesens und Teilnehmer an geförderten Maßnahmen zur beruflichen Fortbildung, Umschulung und betrieblichen Einarbeitung.

Die Kategorisierung „Voll-/Teilzeitbeschäftigung“ richtet sich nach dem Verhältnis zwischen vertraglich vereinbarter und betriebsüblicher Arbeitszeit. Hierbei werden die Teilzeitbeschäftigten in zwei Gruppen unterteilt (die Hälfte der üblichen Arbeitsstunden eines Vollbeschäftigten[22]).

- *Sozialversichungspflichtiges Bruttoentgelt (bis zur Beitragsbemessungsgrenze)*

Als Entgelt wird das Sozialversicherungspflichtige Bruttoarbeitsentgelt (§ 1385 RVO, § 112 AVG und § 130 RKG) verstanden, für das Sozialversicherungsbeiträge abzuführen sind (vgl. Mayer/Becker 1984:996).

Das gemeldete Entgelt ist daher nach oben durch die Beitragsbemessungsgrenze (§ 1385 II RVO bzw. § 122 II AVG) begrenzt. Diese Grenzen (Rentenversicherung der Arbeiter und Angestellten bzw. Knappschaftliche Rentenversicherung) werden jährlich an die Entwicklung der Löhne und Gehälter angepaßt (vgl. Tab. 2).

Die untere Grenze stellt die Entgeltgrenze für geringfügig Beschäftigte (§ 8 des vierten SGB) dar. Ist bei der oberen Grenze der Entgeltbetrag einfach abgeschnitten (Rentenversicherungsgrenze), so sind die geringfügig Beschäftigten im Datensatz überhaupt nicht enthalten (s.o). Für Einkommensanalysen ist es demnach wichtig, daß die Verteilung des Entgelts durch die Geringfügigkeits- und die Beitragsbemessungsgrenze nach unten und oben abgeschnitten ist.

Über die Zeit betrachtet, sind aufgrund der Einkommensentwicklung geringfügig Beschäftigte in die Sozialversicherungspflicht, und somit in die Datei hineingewachsen. „So weist dann auch die Beschäftigtenstatistik für die Teilzeitbeschäftigten unter 20 Wochenstunden von 1977 bis 1983 eine Zunahme um ca. 50.000 Personen oder knapp 16% aus (...)“ (Cramer 1985:60f.).

Durch obige Darstellung wird deutlich, daß der Entgeltbegriff sich langfristig gesehen verändert hat. Das Entgelt wurde allerdings auch „sukzessive auf Bezüge, die neben dem Lohn gezahlt wurden, ausgedehnt, so daß im Zeitverlauf immer mehr Lohnzuschläge dem beitragspflichtigen Entgelt zugerechnet wurden“ (Schmähl/Fachinger 1994:188). Dieser Sachverhalt ist bei Analysen zu berücksichtigen, da sonst Änderungen in der Legaldefinition fälschlicherweise als Strukturbrüche interpretiert werden könnten (vgl. Schmähl/Fachinger 1994:188).

- *Betriebsnummer*

Die Betriebsnummer wird den Arbeitgebern von den Arbeitsämtern zugeteilt. „Der Betrieb in der Beschäftigtenstatistik ist eine örtliche Einheit, die in den Fällen, in denen ein Arbeitgeber lediglich eine einzige Niederlassung oder Arbeitsstätte unterhält, mit dieser identisch ist.“ (Hoffmann/Wermter 1976:33) Ziel der Betriebsnummernvergabe war, die Einheit „Arbeitsstätte“ aus der Arbeitsstättenzählung 1970 abzubilden.

Dies ist allerdings nicht möglich, da Niederlassungen in derselben Gemeinde, die demselben Wirtschaftszweig angehören, zusammengefaßt werden dürfen, wenn die Meldungen zur Sozialversicherung von einer Stelle abgegeben werden (vgl. Hoffmann/Wermter 1976:33)[23]. Andererseits kommt es auch vor, daß eine Niederlassung mehrere Betriebsnummern erhält. Auch bei einem Inhaberwechsel bzw. Rechtsformwechsel mit Änderung der Besitzverhältnisse wird teilweise die alte Betriebsnummer übernommen, teilweise eine neue Betriebsnummer vergeben (vgl. Cramer 1985:63f.), da die Betriebsnummer persönlich an den betreffenden Arbeitgeber gebunden ist. Legt ein Arbeitgeber einen Betrieb still und eröffnet später einen neuen Betrieb, so wird für diesen neuen Betrieb die alte Betriebsnummer verwendet. Bei einer räumlichen Verlagerung des Betriebes wird der Betrieb ebenfalls unter derselben Nummer weitergeführt (vgl. Fritsch et al. 1994:69). Daneben wurden für eine Vielzahl von Betrieben Sondervereinbarungen getroffen.

- *Wirtschaftszweig*

Die Zuordnung des Wirtschaftszweiges erfolgt über den wirtschaftlichen Schwerpunkt des Betriebes (Betriebsnummer), der nach der Wertschöpfung bestimmt wird. Ist die Wertschöpfung nicht ermittelbar, so wird der wirtschaftliche Schwerpunkt mit Hilfe der Beschäftigtenzahl festgelegt. Die Verschlüsselung des Wirtschaftszweiges erfolgt über einen dreistelligen Code, der nach dem „Verzeichnis der Wirtschaftszweige für die Statistik der Bundesanstalt für Arbeit“ (Ausgabe 1973) erstellt wird.

[22] Vor dem 1.1.1988 liegt die Grenze bei 20 Wochenarbeitsstunden, danach bei 18 Stunden.

[23] Seit dem 30.11.1987 ist es mit Hilfe einer sogenannten Z-Betriebsnummer möglich, Betriebe mit mehreren Betriebsnummern zusammenzuführen. Das Merkmal ist allerdings in der Beschäftigtenstatistik nicht verfügbar.

- *Regionale Gliederung (Bundesgebiet bis Gemeindekennziffer)*

Die regionale Gliederung wird über die Betriebsnummer zugespielt (Arbeitsort). Seit 1989 wird intern für die Stichtage auch der Wohnort des Beschäftigten dem Datensatz zugespielt.

- *Beginn und Beendigung der Beschäftigung*

Aus Beginn und Beendigung von Beschäftigungen kann die Beschäftigungsdauer berechnet werden. Dabei ist zu beachten, daß diese Dauer nicht mit den tatsächlich geleisteten Arbeitstagen übereinstimmt, sondern vielmehr die Dauer des sozialversicherungspflichtigen Beschäftigungsverhältnisses darstellt. „Die Beschäftigungsdauer erstreckt sich damit auch auf Tage, an denen in der Regel nicht gearbeitet wird, wie z.B. Samstage, Sonntage und Feiertage. Das Beschäftigungsverhältnis gilt auch weiterhin als bestehend, wenn der Beschäftigte z.B. krank oder im Urlaub ist." (Mayer/Becker 1984:997)

1.5 Fehlerquellen in der Beschäftigtenstatistik

1.5.1 Fehlerquellen bezüglich der erfaßten Merkmale

Bei der Beschäftigtenstatistik sind – wie bei jedem anderen Datensatz auch – Fehler (Dateninkompatibilitäten) möglich. Neben den allgemeinen Fehlerquellen, wie mögliche Fehlcodierungen, sind auch spezifische Fehlerquellen der Beschäftigtenstatistik zu nennen. Diese umfassen die Vergabe von doppelten Versicherungsnummern und das Problem von Mehrfachbeschäftigung bzw. die Anmeldung von Beschäftigungsverhältnissen.

Wie bereits erwähnt, besitzt jeder Versicherte eine eindeutige Versicherungsnummer. Es besteht allerdings die geringe Möglichkeit, daß ein und dieselbe Person im Laufe ihres Erwerbslebens mehr als eine Versicherungsnummer zugeteilt bekommt. Dies kann bei sehr langen Unterbrechungen, insbesondere bei Ausländern, vorkommen. Allerdings sind solche Fehler seit der 2. DEVO/DÜVO (1981) weitestgehend ausgeschlossen.

Ein Problem stellen Mehrfachbeschäftigungen dar, da sich die Meldungen grundsätzlich auf Beschäftigungsverhältnisse beziehen. Mehrfachbeschäftigungen derselben Person können nur durch Prüfung von Zeitüberschneidungen im Versichertenkonto – also erst in der Beschäftigtenstatistik – und nicht an der Meldung festgestellt werden. Ein weiterer Problemkreis ist, daß Anmeldungen von Beschäftigungsverhältnissen nicht unbedingt mit einem Beschäftigungsverhältnis verknüpft sein müssen, da eine Kontrolle über das tatsächliche Bestehen erst durch eine Abmeldung bzw. Jahresmeldung möglich ist.

Es sind eine Reihe von weiteren möglichen Fehlerquellen (z.B. unzulässige Werte) beim Datengenerierungsprozeß denkbar. Um diese Fehler auf ein Minimum zu reduzieren, ist im Rahmen der DEVO/DÜVO ein komplexes Verfahren aus Kontrollen und Gegenkontrollen implementiert (vgl. Wermter/Cramer 1988:471ff.). Dieses beschränkt sich dabei nicht nur auf die bloße Kontrolle von Merkmalen, sondern beinhaltet auch einige Plausibilitätstests.

Der Tätigkeitsschlüssel auf den Meldungen enthält vollständige Angaben über die Stellung im Beruf, die Arbeitszeit, die Ausbildung und eine dreistellige Berufskennziffer. Er wird bei den Krankenkassen maschinell auf zulässige Schlüsselzahlen überprüft und gegebenenfalls verbessert. Ihre Zuverlässigkeit hängt von der Sorgfalt der Personen ab, die von den Arbeitgebern mit der Datenerfassung betraut werden. Dies gilt auch für die Angaben zur Staatsangehörigkeit. Beginn- und Enddatum eines Entgeltzeitraumes werden bei den Krankenkassen auf Vollständigkeit und Zulässigkeit zunächst einzeln und dann auch am Mitgliederbestand auf Einhaltung der Meldefolge und Überschneidungen geprüft. Die Versicherten können selbst diese Angaben kontrollieren und gegebenenfalls Korrekturen bewirken.

In der Bundesanstalt für Arbeit wird der Dateneingang bzgl. der Quantität und der Qualität der einzelnen Merkmale sowie deren Laufzeit geprüft. Sind z.B. in der Höhe des eingegangenen Datenvolumens bereits Unregelmäßigkeiten erkennbar, erfolgen Rückfragen bei den betreffenden Stellen. Weiterhin werden Fehlerstatistiken durch maschinelle Fehlerprüfungen (z.B. unzulässige Schlüsselzahlen) erstellt. Die wichtigste Plausibilitätsprüfung ist die Beobachtung der Strukturergebnisse im Zeitvergleich. Hauptsächlich werden hierbei die Beschäftigungsentwicklung in den Wirtschaftszweigen und die Veränderungen von Eckzahlen für Arbeitsamtsbezirke betrachtet. Darüber hinaus nehmen die Statistischen Landesämter regional detailliertere Plausibilitätsprüfungen vor. Unplausible Ergebnisse werden dem Statistischen Bundesamt mitgeteilt, das diese Hinweise an die Bundesanstalt für Arbeit weiterleitet (vgl. Wermter 1981:432-433).

1.5.2 Fehler bei der Bestandsermittlung durch den verzögerten Dateneingang

Neben Fehlerquellen bei den erfaßten Merkmalen können auch Fehler durch fehlenden oder verzögerten Dateneingang entstehen. Anmeldungen und Jahresmeldungen gelten solange als Beschäftigung, bis die Beendigung durch eine Abmeldung belegt ist. Für fehlende Abmeldungen bietet das Meldeverfahren keine Ersatzinformation, die zu einer Berichtigung der Daten verwendet werden kann. Wie bei jeder Fortschreibungsstatistik besteht jedoch auch bei der Beschäftigtenstatistik die Gefahr der Bestandsüberhöhung. Daher soll ein spezielles „Abschneideverfahren" Fehler aus dem verzögerten und fehlenden Dateneingang in Grenzen halten. Dazu wird der Bestand durch Abfragen der Konten nach einer Wartefrist von sechs Monaten ausgezählt. Dieser Verarbeitungsschritt wird drei Monate später wiederholt, indem die in der Zwischenzeit eingegangenen Meldungen berücksichtigt werden. In der Höhe der Differenzen zwischen diesen beiden ermittelten Beständen werden die Konten „abgeschnitten", die zeitlich weit zurückliegende Jahresmeldungen und Anmeldungen als letzte Meldung enthalten, d.h. sie werden für die nächste Quartalsauszählung nicht mehr als Beschäftigung gewertet. Zu Unrecht abgeschnittene Fälle werden nach und nach bei späteren Zählungen wieder in den Beschäftigtenbestand übernommen (zur Problematik vgl. Wermter/Cramer 1988:479, Cramer et al. 1990:19f., oder auch Cramer/Majer 1991).

2 Die Historikdatei

Die Historikdatei besteht aus sämtlichen Meldungen der Sozialversicherungsträger an die Bundesanstalt für Arbeit[24], die seit Bestehen des Meldeverfahrens eingegangen sind. Die

[24] Der Entstehungsprozeß der Historikdatei ist in Cramer et al. 1989:3f. beschrieben.

Merkmale sind nach bestimmten Vorgaben sortiert und bereinigt. Daher hat die Historikdatei gegenüber den Stichtags- und Jahreszeitraumdateien wesentliche Vorzüge. Die Historikdatei enthält den bis auf den aktuellen Rand vollständigen Informationsgehalt, insbesondere auch verspätet eingegangene Meldungen, die in aktuellen Auswertungen nicht berücksichtigt werden können.

Die Historikdatei ist ein „prozeßproduzierter" Längsschnittdatensatz[25], da sie Informationen beinhaltet, die für die Aufgabenerfüllung der öffentlichen Verwaltung benötigt werden (vgl. Schmähl 1985:277). Es handelt sich somit um sekundärstatistisches Datenmaterial.

Im Zuge der Aufbereitung der Historikdatei werden einige Bereinigungsprozeduren durchgeführt[26], die unmittelbare Probleme der Beschäftigtenstatistik ausräumen. So werden alle Korrekturmeldungen (inkl. der Stornierungen), die sich meist auf Merkmale vorangegangener Beschäftigungsverhältnisse beziehen, eingearbeitet. Da die Angaben zu Anmeldungen vollständig in den Jahresmeldungen enthalten sind, macht es keinen Sinn, Anmeldungen gesondert in der Datei zu speichern. Vollständig oder teilweise identische Datensätze von Beschäftigtenverhältnissen werden ebenfalls gelöscht (vgl. Cramer et al. 1989:5f.).

Fehlt im Kontoverlauf für ein bestimmtes Jahr die Jahresmeldung, liegt aber für das vorangegangene Jahr eine Jahresmeldung und für das nachfolgende Jahr unter derselben Betriebsnummer eine Beschäftigung ab dem Jahresbeginn vor, so wird die offenbar fehlende Meldung ergänzt. Dies wurde im Jahreszeitraum von 1976 bis 1987 bei ca 1,4 Millionen Fällen (ca 3,5% des Gesamtbestandes) durchgeführt. Dabei ist von keinem zufälligen Prozeß auszugehen, da eine Häufung von Jahresmeldungen in einigen Jahren (1976, 1978, 1980, 1983 und 1985) in bestimmten Wirtschaftszweigen festzustellen ist (vgl. Cramer et al. 1990:12). Die Ersetzungen erscheinen richtig und notwendig zu sein, denn in „ ... allen aufgelisteten Wirtschaftszweigen ergeben sich erst durch die ersetzten Jahresmeldungen plausible Beschäftigungsverläufe." (Cramer et al. 1990:13).

In der Datei sind außerdem Beschäftigungsverhältnisse zu finden, die mit einer Jahresmeldung bzw. Unterbrechungsmeldung, nicht jedoch mit einer regulären Abmeldung „enden". Mögliche Gründe sind die Beendigung des Beschäftigungsverhältnisses ohne Abmeldung, die Beantragung einer zweiten Versicherungsnummer oder die Verzögerung der Abgabe bzw. Weiterleitung von Jahresmeldungen (vgl. Wermter/Cramer 1988:479). Für diese Problemfälle werden folgende Setzungen durchgeführt:

- Ist die letzte Meldung eine Jahresmeldung, dann wird danach auf Beschäftigungsende geschlossen.

- Bei einer Unterbrechungsmeldung wird – vom Datum der Unterbrechung an – noch ein Jahr Beschäftigung unterstellt und dann das Beschäftigungsverhältnis beendet. Diese Zeitspanne wird gesondert ausgewiesen.

- Wenn nach Unterbrechungszeiten im Kontoverlauf keine Abmeldungen folgen, so wird ein unterbrochenes Beschäftigungsverhältnis vom Beginn der Unterbrechung bis zum Beginn des nächsten Beschäftigungsverhältnisses gezählt. Erreicht die Dauer der Unterbrechung die Einjahresgrenze, so wird eine Abmeldung vorgenommen (vgl. Cramer/Majer 1991:82). Die Einjahresgrenze ist eine suboptimale Setzung, die auf einfache Weise die Fortzahlung von unterbrochenen Beschäftigungsverhältnissen in möglichen Querschnittsauswertungen ermöglichen soll. Da Unterbrechungen für Wehr- und Zivildienst und bei Bezug von Erziehungsgeld länger dauern können, müßten sie spezifisch für einzelne Personengruppen definiert werden. In der Historikdatei kann mittels einer Variablen zwischen aktiven und ruhenden Beschäftigungen unterschieden werden (vgl. Cramer/Majer 1991:82).

Nach Durchführung dieser Bereinigungsprozeduren umfaßt die Historikdatei über 600 Millionen Datensätze (vorher 790 Millionen). Dabei handelt es sich um alle Meldungen, die bis Ende Januar 1990 bei der Bundesanstalt für Arbeit eingegangen sind. Die Vollständigkeit des Datenmaterials wurde zunächst über die Verteilung dieser Datensätze auf Meldejahre und die Endziffer der Geburtstage kontrolliert. Hierbei ergibt sich folgendes Bild:

- Die Jahre 1973 und 1974 weisen einen sehr unregelmäßigen Datenbestand auf. Für diese Jahre ist die Historikdatei nicht vollständig genug (vgl. Cramer et al. 1990:2).

- Das Jahr 1975 enthält für die Geburtstagsendziffern 2 und 5 im Vergleich zu den folgenden Jahren deutlich weniger Datensätze als für die übrigen Endziffern (vermuteter Fehlbetrag: 787.000 Datensätze). Das Jahr 1975 ist demnach für Auswertungen problematisch (vgl. Cramer/Majer/John 1990:2f).

- Die Jahre 1981 bis 1988 (speziell 1985) sind für die Geburtsendziffer 0 aufgrund einer Untererfassung des Ausgangsmaterials nicht vollständig. Der Gesamtdatenbestand ist im Jahre 1985 um etwa 0,22% zu niedrig, in den übrigen Jahren (1981-1989) zwischen 0% und 0,07% (vgl. Cramer/Majer 1991:83, Cramer/Majer/John 1990:3f).

- Es gibt einige unplausible Beschäftigungsverläufe in einzelnen Wirtschaftszweigen, die allerdings kein großes Gewicht besitzen[27]. In den meisten Fällen dürfte es sich um Änderungen der Wirtschaftszweigzuordnung aufgrund von Änderungen des wirtschaftlichen Schwerpunktes oder der regionalen Zuordnung bestimmter Betriebe handeln (vgl. Cramer/Majer/John 1990:14).

- Für das Jahr 1980 wird vermutet, daß Angestellte zu einem beträchtlichen Teil falsch, nämlich als Arbeiter, kodiert sind.

- Fragwürdig erscheint die Entwicklung bei Vollzeit- und Teilzeitbeschäftigung für die Jahre 1982 (zu viele Teilzeitmeldungen), 1983 und 1984 (zu wenige Teilzeitmeldungen) (vgl. Cramer/Majer/John 1990:15).

Eine wichtige Rolle spielt auch der verspätete Dateneingang, der zum Großteil aus neuen Meldungen aus „regulärer Beschäftigung gegen Entgelt" besteht. Nimmt man den verspäteten Dateneingang im Laufe des Jahres 1989 als Maßstab für

---

[25] Schmähl weist zurecht daraufhin, daß der Begriff „prozeßproduziert" nicht sehr zweckmäßig gewählt ist. Die Daten sind vielmehr „prozeß-notwendig" (Schmähl 1985:277).

[26] Die Programme liegen dokumentiert vor (vgl. Rudolph 1993).

[27] Zu nennen sind:
In der Erdöl-, Erdgasgewinnung (Wirtschaftszweig: 07) steigt 1984/85 die Beschäftigung um ca. 16% an. Im Kali- und Salzbergbau fehlen 1981 9.000-10.000 Beschäftigte. Bei Ziehereien und Kaltwalzwerke (20) geht ab 1986/87 die Beschäftigung um etwa 5% zurück. Gleiches gilt für den Wagenbau (24) ab 1987. Zum Jahresanfang 1980 fehlen bei der Bundespost (64) ca. 15.000 Beschäftigte, die allerdings im Laufe des Jahres ausgeglichen werden (Anstieg der Beschäftigung). Verlags- und Pressewesen (77) zeigt ab 1982 einen ungewöhnlichen Verlauf (vgl. Cramer/Majer/John 1990:14).

Tabelle 3: Erwartete Zunahme an Datensätzen bis Januar 1995

Jahr	Prozentualer Anteil
1982	0,3%
1983	0,4%
1984	0,5%
1985	0,6%
1986	1,1%
1987	1,9%

die Jahre 1982-1987 (vgl. Tab.3), dann ist mit zusätzlichen Meldungen von 0,3% (1982) bis 1,9 (1987) zu rechnen (vgl. Cramer/Majer 1991:83).

Die verspätet eingegangenen Meldungen konzentrieren sich auf bestimmte Personengruppen und bestimmte Wirtschaftszweige, zu denen Ausländer, Teilzeitkräfte, Frauen und Arbeiter (vgl. Cramer/Majer 1991:86) gehören[28]. Ein relativ unpünktliches Meldeverhalten zeigen eine Reihe von Betrieben in den Wirtschaftszweigen Landwirtschaft, Garten- und Weinbau, Kunststoffverarbeitung, Ledererzeugung, Bekleidungsgewerbe, Nahrungsmittelherstellung, Fleischverarbeitung, Handel, Hotels, Altenheime, Wäschereien, Frisöre, Bildungsstätten, Kunst/Theater/Film, sonstige Dienstleistungen, private Haushalte und Vertretungen fremder Staaten (vgl. Cramer/Majer/John 1990:10). Hauptsächlich kommen diese Verzögerungen bei kleineren Betrieben und Branchen mit höherer Personalfluktuation vor.

3 Die IAB-Beschäftigtenstichprobe 3.1

Stichprobenziehung

Aus der Historikdatei wird in einem ersten Schritt eine 5%ige systematische Stichprobe[29] (jede 20. Versicherungsnummer) gezogen. Die Ziehung erfolgt für Ausländer und Deutsche getrennt, da die Historikdatei nach Geburtskohorten sortiert ist und mit Klumpungen speziell bei den Ausländern zu rechnen ist (vgl. Cramer 1985:63, Karr et al. 1986:31)[30]. Aus dieser 5%-Stichprobe wird dann in einem zweiten Schritt jede 5. Versicherungsnummer für die l %-Stichprobe gezogen.

Die IAB-Beschäftigtenstichprobe deckt einen auswertbaren Zeitraum von 16 Jahren (01.01.1975 – 31.12.1990) ab und umfaßt in den Jahresquerschnitten jeweils etwa 200.000 Personen und im Längsschnitt 426.363 Personen. Dies entspricht in etwa 4,8 Mio. Datensätzen, die sich nur auf die sozialversicherungspflichtig Beschäftigten beziehen.

Die Auswahlgesamtheit besteht aus allen Beschäftigungsfällen, die im Zeitraum zwischen 1975 und 1990 mindestens einmal Sozialversicherungspflichtig beschäftigt waren. Damit wird aber auch gleichzeitig deutlich, daß die vorhandene Datei keine Rechteckdatei ist, da nur für bestimmte Zeiträume zwischen 1975 und 1990 Informationen vorliegen müssen. Da die Stichprobe nicht nach der Betriebsnummer geschichtet wurde, liegt für die Betriebsebene keinerlei Repräsentativität vor.

Die Qualität der Stichprobe ist nach verschiedenen Kriterien geprüft worden. So wurde eine Reihe von Auszählungen der Sozialversicherungspflichtig Beschäftigten in der Stichprobe mit den Angaben in den Quartalsdateien verglichen. Die vorliegenden Ergebnisse deuten auf eine gute Qualität der IAB-Beschäftigtenstichprobe hin. Die Darstellung der Ergebnisse würden den Rahmen dieses Beitrags sprengen, so daß sie zu einem späteren Zeitpunkt veröffentlicht werden.

3.2 Zusätzliche Datenbestände

32.1 Die Leistungsempfängerdatei der Bundesanstalt für Arbeit

Neben der Beschäftigtenstatistik verfügt die Bundesanstalt für Arbeit über eine weitere wichtige Datenquelle, die sogenannte Leistungsempfängerdatei: In ihr sind diejenigen Zeiträume erfaßt, in denen Personen Lohnersatzleistungen von der Bundesanstalt für Arbeit bezogen haben, wie Arbeitslosengeld, Arbeitslosenhilfe oder Unterhaltsgeld für die berufliche Fortbildung bzw. die Umschulung.

Damit sind allerdings nicht alle möglichen Arbeitslosigkeitsperioden erfaßt, denn für den Bezug von Arbeitslosengeld (bzw. die sogenannte „orginäre“ Arbeitslosenhilfe) müssen bestimmte Voraussetzungen erfüllt sein, und die danach mögliche Arbeitslosenhilfe greift nur bei persönlicher Bedürftigkeit. Desweiteren werden nicht alle Leistungen der Bundesanstalt für Arbeit an Personen erfaßt.

Insgesamt sind von 1975 bis 1990 ca. 60 verschiedene Leistungsarten in der Leistungsempfängerdatei vorhanden, die sich folgendermaßen zusammenfassen lassen:

1. Arbeitslosengeld und vergleichbare Leistungen (z.B. Arbeitslosenbeihilfe, Eingliederungsgeld bei Arbeitslosigkeit, Altersübergangsgeld)

2. Arbeitslosenhilfe und vergleichbare Leistungen (z.B. Arbeitslosenhilfe für ehemalige Entwicklungshelfer)

3. Unterhaltsgeld und vergleichbare Leistungen (z.B. Eingliederungsgeld bei beruflicher Fortbildung/Umschulung oder bei Deutsch-Sprachlehrgängen).

Nicht erfaßt werden folgende Leistungen der Bundesanstalt für Arbeit:

- Leistungen nach § 45 AFG (z.B. Lehrgangskosten, Lernmittelkosten, usw.)
- Berufsausbildungsbeihilfe (vgl. § 40 AFG)
- Leistungen zur Förderung der Arbeitsaufnahme (vgl. § 53 AFG)
- Leistungen zur Aufnahme einer selbständigen Tätigkeit (vgl. §55aAFG)

---

[28] Dieser Sachverhalt schlägt sich unmittelbar in den Korrekturläufen nieder. „Frauen-, Ausländer- und Teilzeitbeschäftigung wird durch die Aktualisierung überproportional angehoben, wobei die absoluten Zuwächse für Deutsche und Vollzeitbeschäftigte und ab 1982 auch für Männer größer sind.“ (Cramer/Majer/John 1990:8)

[29] Die systematische Auswahl (systematisches Stichprobenverfahren) wird oft als Hilfsverfahren für die uneingeschränkte Zufallsauswahl benutzt, da es (vielfach) einfacher ist eine solche Stichprobe zu ziehen. Es gilt im allgemeinen, daß eine systematische Stichprobe dann einer Zufallsauswahl gleichzusetzen ist, wenn die Anordnung der Gesamtheit in bezug auf die Untersuchungsmerkmale zufällig ist. Eventuelle Anordnungseffekte in der Gesamtheit können bei einer systematischen Auswahl zur Verbesserung der Schätzgenauigkeit (gegenüber einer uneingeschränkten Auswahl) führen (vgl. z.B. Statistisches Bundesamt 1960: 83ff). Da die Historikdatei nach Geburtskohorten sortiert ist, liegt solch ein Anordungseffekt vor. In diesem Fall dürfte die systematische Stichprobe mindestens so genau sein, wie eine geschichtete Stichprobe, wenn man die Geburtskohorten als Schichten definiert (vgl. z.B. Cochran 1972: 245ff).

[30] Die Historikdatei liegt sortiert auf 400 Magnetbändern nach 40 (etwa gleichmäßig besetzten) Geburtskohorten vor. Dabei sind jeweils folgende Geburtsjahre zu einer Kohorte zusammengefaßt: 1890-1919, 1920-1923, 1924-1926, 1927-1928, 1929-1930, 1931-1932, ..., 1965-1966 und 1967-1973 (vgl. Cramer et al. 1989:1-2).

- Leistungen nach § 91ff AFG (Arbeitsbeschaffungsmaßnahmen)
- und das Übergangsgeld (berufsfördernde Leistungen zur Rehabilitation, § 56ff AFG).

Die dokumentierten Leistungen sind in dem Merkmal „Leistungsart" abgelegt. Die Zuordnung bestimmter Leistungen durch die Bundesanstalt für Arbeit zu bestimmten Merkmalsausprägungen unterliegt einer Reihe von verfahrenstechnischen Änderungen; die Ursache hierfür liegt in den vielen Änderungen im AFG. So wurden bestimmte Leistungsarten neu eingeführt (z.B. Unterhaltsgeld als Darlehen) oder die Anspruchsvoraussetzungen für bestimmte Leistungen geändert. Daher erfordert die Interpretation von Leistungsbezügen (insbesondere beim Unterhaltsgeld) genaue Kenntnisse der AFG-rechtlichen Situation und der entsprechenden Ausführungserlasse. Daher werden in der IAB-Beschäftigtenstichprobe die Leistungsarten nach Arbeitslosengeld, Arbeitslosenhilfe und Unterhaltsgeld zusammengefaßt.

Die IAB-Beschäftigtenstichprobe wird anhand der Versicherungsnummer der Beschäftigten mit der Leistungsempfängerdatei der Bundesanstalt für Arbeit abgeglichen und um die Zeiträume ergänzt, in denen die in die Auswahl gelangten Beschäftigten Lohnersatzleistungen beziehen. Neben dem Merkmal „Leistungsart" sind in der IAB-Beschäftigtenstichprobe noch drei weitere Merkmale enthalten:

- Das Merkmal „Beendigungsgrund" gibt an, aus welchem Grund der Bezug von Leistungen durch die Bundesanstalt für Arbeit vorläufig (z.B. Beendigungsgrund: Sperrzeit 8 Wochen), endgültig (z.B. Arbeitsaufnahme) beendet wird oder der weitere Leistungsbezug durch den Empfänger neu beantragt werden muß (Beendigungsgrund: Bewilligungsabschnitt abgelaufen).

- Die Merkmale „Leistungsbeginn und Leistungsende" kennzeichnen den Zeitraum, in dem eine Person Leistungen durch die Bundesanstalt für Arbeit erhält.

Ähnlich wie bei der Historikdatei erfaßt ein einzelner Datensatz im Regelfall längstens einen Zeitraum von einem Jahr. Das heißt, erhält eine Person zwei Jahre lang fortlaufend Leistungen von der Bundesanstalt für Arbeit, so existieren im Regelfall zwei oder mehr Datensätze, die jeweils nur den Leistungsbezug eines Teilzeitraums anzeigen. In einigen Fällen kann aber auch ein einzelner Datensatz einen längeren Zeitraum des Leistungsbezugs kennzeichnen.

Da nicht für alle Datensätze in der Leistungsempfängerdatei eine Versicherungsnummer vorliegt, wird die Anzahl der Bezugszeiten von Arbeitslosengeld, Arbeitslosenhilfe und Unterhaltsgeld etwas unterschätzt. Der Anteil der Datensätze ohne Versicherungsnummer schwankt zwischen 1,4% (1983) und 8% (1975 und 1989). Dagegen weisen in den Jahren 1978 und 1979 aufgrund einer Untererfassung 50% der Datensätze in der Leistungsempfängerdatei keine Versicherungsnummer auf.

3.2.2 Betriebsinformationen aus der Beschäftigtenstatistik

Die Grundgesamtheit der Beschäftigtenstatistik besteht aus Arbeitnehmern, allerdings ermöglicht die Betriebsnummer (vgl. auch Kap. 1.4 und Kap. 3.2.2) eine Aggregation aller sozialversicherungspflichtig Beschäftigten auf Betriebsebene. Die Angaben über die Betriebe werden aus den Bestandsdaten der Beschäftigtenstatistik jeweils zum 30.6 eines Jahres (ab 1977) gewonnen.

Neben dem Wirtschaftszweig werden noch folgende Betriebsmerkmale für die IAB-Beschäftigtenstichprobe bereitgestellt:

- Betriebsgrößenklassen (nur Sozialversicherungspflichtig beschäftigte Personen)
- Anteil an Sozialversicherungspflichtig Beschäftigten mit FHS/Uni-Abschluß
- Anteil an Sozialversicherungspflichtig Beschäftigten mit abgeschlossener Berufsausbildung
- Anteil an Sozialversicherungspflichtig Beschäftigten ohne abgeschlossene Berufsausbildung (vgl. Cramer 1987:16).

Diese Merkmale sind allerdings in bestimmten Wirtschaftszweigen extrem verzerrt, da sich die aggregierten Merkmale immer auf die Sozialversicherungspflichtig Beschäftigten beziehen (vgl. die Bemerkungen zum Merkmal „Wirtschaftszweig" in Kap 1.3). Daher werden in einigen Untersuchungen, die sich auf Betriebsinformationen aus der Beschäftigtenstatistik beziehen, einzelne Wirtschaftsbereiche ausgeklammert (vgl. z.B. Cramer 1987:16 oder Boeri/Cramer 1991:71).

Da der Unternehmer selbst nicht Sozialversicherungspflichtig ist, sind Informationen über Betriebe, die keinen Sozialversicherungspflichtigen beschäftigten, nicht vorhanden. So sind Ein-Personen-Unternehmen (eventuell mit mithelfenden Familienangehörigen) nicht erfaßt.

Darüberhinaus ergeben sich einige problematische Typen von Betriebsnummernverläufen (vgl. Fritsch et al. 1994:71-74):

1. Perforierte Betriebsnummernverläufe[31]
2. Willkürlicher Wechsel der Betriebsnummer bei Wechsel des Inhabers oder Änderung der Rechtsform
3. Integration mehrerer Betriebe unter einer Nummer
4. Ausgliederung von Unternehmensteilen unter einer neuen Nummer
5. Veränderte Zuordnung der Beschäftigten zu mehreren Nummern[32]

Für die Betrachtung von Mobilitätsprozessen auf der Personenebene bedeuten diese ungeklärten Betriebsnummernverläufe – neben den Problemen der Betriebsnummer an sich (vgl. Kap. 1.4) -, daß nicht jede Betriebsnummernänderung einer Person einem Stellenwechsel entspricht (vgl. Karr et al. 1986:6).

Ein weiteres Problem ist, daß die Betriebsangaben reine Stichtagsinformationen sind, aber für die Personen Längsschnittinformationen vorliegen. So gelten die Betriebsinformationen für ein Beschäftigungsverhältnis immer nur für den 30.06 des jeweiligen Jahres, auch wenn das Beschäftigungverhältnis erst danach beginnt bzw. davor endet.

Trotzdem bietet die alleinige Information der Betriebsnummer einige Vorteile für die Analyse von Erwerbsverläufen. Durch die Bereitstellung der Betriebsnummer ist der betriebliche Kontext verfügbar und für Betrachtungen z.B. von innerbetrieblicher und außerbetriebliche Mobilität, von Segmentationsansätzen und von Lohnbetrachtungen verwendbar. Es ist möglich, auf der Personenebene einzelne Betriebsnummern – aufgrund der hohen Fallzahlen – in Analysen einzubeziehen.

[31] „Hierunter sind solche Fälle zu verstehen, in denen eine Betriebsnummer über einen gewissen Zeitraum nicht erscheint, dann aber wieder auftaucht." (Fritsch et al. 1994:71).

[32] Wie mit diesen problematischen Betriebsverläufen umgegangen werden kann, findet sich in Fritsch et al. (1992:7-24).

4 Die Anonymisierung der IAB-Beschäftigtenstichprobe

4.1 Grundlagen

Grundlage der hier dargestellten Anonymisierungsmaßnahmen sind die Ergebnisse des Projekts zur faktischen Anonymität des Mikrozensus und der Einkommens- und Verbrauchsstichprobe, das unter Leitung von Prof. Walter Müller in Kooperation mit dem Statistischen Bundesamt, dem Zentrum für Umfragen, Methoden und Analysen (ZUMA) und dem Lehrstuhl für Methoden der Empirischen Sozialforschung und Angewandte Soziologie der Universität Mannheim durchgeführt wurde (vgl. Müller et al. 1991). Eine Anlehnung an dieses Anonymisierungsprojekt bietet sich an, da hier zum ersten Mal umfangreiche empirische und systematische Tests der Anonymität von statistischen Mikrodaten durchgeführt wurden[33]. Im Rahmen dieses Projektes wurden Richtlinien erarbeitet[34], die sich die Statistischen Ämter inzwischen als Regeln für die Weitergabe zu eigen gemacht haben. Sie sind auch unter juristischen Gesichtspunkten geprüft. Aus diesem Grund bietet es sich an, diese Empfehlungen als Leitlinien für die Anonymisierung der IAB-Beschäftigtenstichprobe zu übernehmen.

Nach §16 Abs.6 BStatG müssen personenbezogene Daten, die an die Wissenschaft übermittelt werden, faktisch anonym sein. Neben der Anonymisierung von Personendaten ist das IAB auch verpflichtet, Betriebsdaten faktisch zu anonymisieren (vgl. § 35 Abs. l SGB). Die faktische Anonymität ist dann gegeben, wenn ein Datenangreifer unverhältnismäßig viel Zeit, Kosten und Arbeitskraft aufbringen muß, um einen Datensatz zu deanonymisieren. Mit dem Konzept der faktischen Anonymität geht man von einem rational kalkulierenden Datenangreifer aus, der die Kosten und den Nutzen seines Reidentifikationsversuchs abwägt. Bei faktisch anonymisierten Daten kann eine spätere Deanonymisierung nicht mit Sicherheit ausgeschlossen werden.

Unter einer Deanonymisierung (Reidentifikation) wird die nachträgliche Rekonstruktion eines Personen- bzw. Betriebsbezugs verstanden. Diesen Bezug kann ein Datenangreifer nur herstellen, wenn er über sogenanntes Zusatzwissen (Identifikationsfile) verfügt, das sich durch direkte Indikatoren (z.B. Namen und Anschrift) der Personen oder Betriebe auszeichnet (personenbezogene bzw. betriebsbezogene Daten). Gleichzeitig muß das Zusatzwissen auch gemeinsame Merkmale (Überschneidungsmerkmale) mit dem anonymen Mikrodatenfile aufweisen, da sonst der „Schlüssel" für eine Reidentifikation fehlt. Der Mikrodatenfile selbst muß über zusätzliche Merkmale verfügen, die nicht im Zusatzwissen des Angreifers enthalten sind. Diese stellen den Nutzen einer Deanonymisierung dar. Eine Deanonymisierung beinhaltet demnach zwei Komponenten, nämlich die Identifikation einer Person bzw. eines Betriebs und gleichzeitig das Gewinnen von zusätzlicher Information (vgl. Skinner et al. 1994:33, Keller-McNully/Unger 1993:477).

„Ein Angreifer wird dann versuchen, durch einen Abgleich der Überschneidungsmerkmale auf Identität oder sehr große Ähnlichkeit jene Datensätze im Mikrodaten- und Identifikationsfile zu ermitteln, die von ein und derselben Person stammen. Eine Reidentifikation wäre dann gegeben, wenn es anhand der Überschneidungsmerkmale möglich ist, für einen Datensatz des Mikrodatenfiles eine eins-zu-eins Relation zu einem Datensatz des Zusatzwissens herzustellen und wenn sichergestellt ist, daß sich diese Datensätze auf ein und dieselbe Person beziehen." (Wirth 1992:11) Diese Grundlogik einer Reidentifikation trifft auch für die Betriebsangaben zu.

Die anonymisierte IAB-Beschäftigtenstichprobe soll ausschließlich der Wissenschaft zu Analysezwecken zur Verfügung gestellt werden (scientific use file). Eine Beschränkung auf die Wissenschaft erfolgt, weil keine plausiblen Motive einer Deanonymisierung aufgrund der beruflichen Interessenlage von empirisch arbeitenden Wissenschaftlern rekonstruierbar sind (vgl. Wirth 1992:15, Müller et al. 1991:132ff., Müller/Blien/Wirth 1995).

Bei einer gezielten Suche nach einer bestimmten Person bzw. einem bestimmten Betrieb hängt das Reidentifikationsrisiko von vier Bedingungen ab, die erfüllt sein müssen, um eine erfolgreiche Deanonymisierung eines gesuchten Datensatzes zu erhalten[35]:

A.) Der Datensatz des zu identifizierenden Falles muß im Mikrodatenfile enthalten sein *(Repräsentations- oder Selektivitätsproblem).*

Wenn die gesuchte Person bzw. der gesuchte Betrieb nicht im Mikrodatenfile enthalten ist, dann ist auch ein Reidentifikationsversuch sinnlos. Da die meisten Mikrodatenfiles als Stichproben vorliegen, ist für eine beliebige in der Grundgesamtheit enthaltene Person die maximale Erfolgswahrscheinlichkeit einer Reidentifikation gleich dem Auswahlsatz dieser Stichprobe.

„Die von Stichproben ausgehende Schutzwirkung besteht dann nicht mehr, wenn ein Angreifer weiß, welche in seinem Zusatzwissen enthaltenen Personen an der Mikrodatenerhebung teilgenommen haben." (Wirth 1992:11) In diesem Fall liegt response knowledge vor und das Repräsentationsproblem ist einer Repräsentationssicherheit gewichen. Für die Personen in der IAB-Beschäftigtenstichprobe kann das response knowledge leicht über den Stichprobenauswahlsatz bestimmt werden (vgl. Kap.3.1). Es beträgt ein Prozent. Dagegen läßt sich das response knowledge für Betriebe nicht so leicht bestimmen, da es unmittelbar von der Betriebsgröße abhängt (vgl. Kap. 4.3.1).

B.) Die Überschneidungsmerkmale müssen im Identifikationsfile und im Mikrodatenfile in identischer Weise abgebildet sein *(Kompatibilitätsproblem).*

Eine Reidentifikation setzt voraus, daß die gesuchten Datensätze in Mikrodatenfile und Zusatzwissen identisch abgebildet werden. Dateninkompatibilitäten können durch unterschiedliche Operationalisierungen von Merkmalen, unterschiedliche zeitliche Erhebungszeitpunkte und verschiedene Meßfehler (z.B. falsche Angaben der Befragten, Kodierfehler, Übertragungsfehler) entstehen. Da die IAB-Beschäftigtenstichprobe durch einen gänzlich anderen Generierungsprozeß (Erhebungszeitpunkt, -ziel oder -kontext, sowie Meßinstrument) als die meisten Identifikationsfiles gewonnen wird (vgl. Kap. 1.3), wird es zu vielfältigen Abweichungen in den Überschneidungsmerkmalen kommen.

---

[33] Darüberhinaus gibt es zahlreiche, ebenfalls wichtige Analysen zu Teilproblemen, bei denen überwiegend analytisch-statistische Verfahren verwendet werden (vgl. Bethlehem et al. 1990, Paaß/Wauschkuhn 1985, Skinner etal. 1990).

[34] In der Anonymisierung des Mikrozensus wird zwischen allgemeinen Schutzvorkehrungen und datenfilespezifischen Anonymisierungsmaßnahmen unterschieden. Erstes meint insbesondere die vertragliche Versicherung des Datenempfängers, jeden Reidentifikationsversuch zu unterlassen und durch geeignete technische und organisatorische Maßnahmen sicherzustellen, daß externe Zugriffe verhindert und interne Deanonymisierungsversuche rechtzeitig erkannt werden (Müller et al. 1991). Diese allgemeinen Schutzvorkehrungen werden auch für die Anonymisierung der IAB-Beschäftigtenstichprobe verwendet (vgl. Kap. 6).

[35] Die Darstellung folgt Skinner et al. 1994, Wirth 1992 und Müller et al. 1991, im Anschluß an Marsh et al. 1991 und Paaß/Wauschkuhn 1985.

C.) Der gesuchte zu identifizierende Fall muß im Hinblick auf die Ausprägungen der Überschneidungsmerkmale in der Population einmalig sein *(Problem der Populationseinzigartigkeit, Uniqueness)*.

Hierbei kommt dem Informationsgehalt der Überschneidungsmerkmale eine zentrale Bedeutung zu. Eine Reidentifikation setzt die Einzigartigkeit von Ausprägungskombinationen der Überschneidungsmerkmale voraus. Auf der anderen Seite gilt allerdings der Sachverhalt, daß mit steigender Anzahl von Überschneidungsmerkmalen gleichzeitig das Kompatibilitätsproblem anwächst. Je mehr Überschneidungsmerkmale für eine Reidentifikation verwendet werden, desto höher ist also das Risiko, daß alle Merkmale identisch abgebildet sind.

D.) Die Reidentifikation muß vom Angreifer auf Richtigkeit überprüft werden *(Sicherheitsproblem)*.

Dieser Sachverhalt wird auch als Verification of Population Uniqueness verstanden, d.h der Angreifer muß sicher sein, daß die Ausprägungskombination in der Population einmalig ist (vgl. Skinner et al. 1994:37ff., bzw. Müller et al. 1991:88f.).

Es reicht nicht aus, daß der Datenangreifer nur Informationen von den Personen bzw. Betrieben hat, die im Identifikationsfile vorliegen. Er muß auch über Informationen über Personen bzw. Betriebe verfügen, die nicht im Mikrodatenfile enthalten sind, um sich seiner Reidentifikation sicher zu sein. Durch Personen bzw. Betriebe mit identischen Ausprägungskombinationen in den Überschneidungsmerkmalen (statistischer Doppelgänger) wächst das Problem, daß die vorgefundene Ausprägungskombination der Überschneidungsmerkmale in der Population einmalig sein soll. Da der Datenangreifer keine Informationen über die gesamte Population besitzt, läßt sich diese Unsicherheit auch nicht beseitigen. Hierbei spielt auch das Kompatibilitätsproblem eine Rolle, da durch Abweichungen in den Überschneidungsmerkmalen das Sicherheitsproblem anwächst.

Zusammenfassend läßt sich sagen, daß eine Reidentifikation nicht mit Sicherheit, sondern nur mit einer bestimmten Wahrscheinlichkeit möglich ist (vgl. Marsh et al. 1991; Bethlehem et al. 1990). Durch die o.g. vier Faktoren (Repräsentations-, Kompatibilitäts-, Uniqueness- und Sicherheitsproblem) läßt sich die Wahrscheinlichkeit des Reidentifikationsrisikos (W_R) für eine bestimmte Person abschätzen (vgl. Skinner et al. 1994:42)[36]:

$$W_R = W_A W_{B/A} W_{C/A,B} W_{D/A,B,C}$$

mit

W_A: Wahrscheinlichkeit für Repräsentationsproblem
W_B: Wahrscheinlichkeit für Kompatibilitätsproblem
W_C: Wahrscheinlichkeit für Uniqueness
W_D: Wahrscheinlichkeit für Sicherheit

Mit einigen Annahmen (vgl. Skinner et al. 1994:41-45) ergibt sich:

$$W_R = W_A W_B W_{C/B} W_{D/A,B,C}$$

Diese Gleichung läßt sich für verschiedene Angriffsszenarien modifizieren (vgl. Müller et al. 1991:93ff oder auch Müller/Blien/Wirth 1995). Für den Spezialfall, daß Kenntnis über das Vorhandensein einer Person bzw. eines Betriebs im Datensatz (response knowledge) vorliegt, gilt:

Es muß unterstellt werden, daß nicht nur Repräsentationssicherheit (W_A= l) vorliegt, sondern daß die Wahrscheinlichkeit von Uniqueness in der Population durch die Wahrscheinlichkeit einer einzigartigen Ausprägungskombination im Mikrodatenfile ersetzt wird. Die Schutzfunktion der Bedingung C wird somit wesentlich gesenkt (W_c erhöht sich). Zusätzlich entfällt auch das Sicherheitsproblem, da eine einzigartige Ausprägung im Mikrodatenfile der gesuchte Fall sein muß (W_D=1). Somit bleibt bei response knowledge, neben einer erhöhten Wahrscheinlichkeit für Uniqueness, nur das Kompatibilitätsproblem (W_B) zum Schutz gegen einen Datenangriff (vgl. Müller et al. 1991: 94f.). Zusätzlich kann auch angenommen werden, daß sich die Suchheuristik des Datenangreifers ändern kann. So kann z.B. bei Nicht-Vorhandensein eines identischen Falles die Entscheidungsregel‚Übereinstimmung' die Entscheidungsregel ‚nächster Nachbar' ersetzen.

Ein gravierender Unterschied zwischen der IAB-Beschäftigtenstichprobe und dem Mikrozenus ist, daß die IAB-Beschäftigtenstichprobe ein Verlaufsdatensatz ist, der einen Zeitraum von 16 Jahren abdeckt. Die Längsschnitte enthalten Informationen (z.B. Dauern), die potentiell für einen Datenangriff nutzbar sind. Daher muß neben einer Querschnittsanonymisierung auch eine Längsschnittanonymisierung vorgenommen werden.

Unter einer Querschnittsanonymisierung wird die faktische Anonymisierung einzelner Merkmale zu einem bestimmten Zeitpunkt verstanden, d.h. es werden Informationen für einen bestimmten Stichtag (z.B. 30.06.1987) anonymisiert (vgl. Kap. 1.2 und Kap.4.3). Eine Verlaufsdatei enthält gegenüber einer Stichtagsdatei zusätzliche Informationen über einen Zeitraum (z.B. Dauern). Daher umfaßt die Längsschnittanonymisierung eine Anonymisierung der zusätzlichen Informationen, die eine Verlaufsdatei gegenüber einer Stichtagsdatei enthält.

Der Erfolg eines Deanonymisierungsversuchs hängt unmittelbar von dem verfügbaren Zusatzwissen ab. Die Verfügbarkeit und Nutzbarkeit des Zusatzwissens wurde im o.g. Mikrozensus-Projekt nur für personenbezogene Daten eruiert. Damit stand man bei der Anonymisierung von Betriebsangaben in der IAB-Beschäftigtenstichprobe vor einem Problem, da über das Reidentifikationsrisiko und dafür vorhandenes Zusatzwissen von Betrieben keine systematischen Kenntnisse vorhanden sind. Es muß davon ausgangen werden, daß ein Teil der Betriebe durch ihre spezifischen Merkmalskombinationen (Größe, räumliche Verteilung, Branchenzugehörigkeit) per se leichter zu identifizieren sind als Personen. Insbesondere durch das Hinzufügen von originären Betriebsinformationen ist eine Reidentifikation von einzelnen Betrieben vorstellbar. Bei der Diskussion über die Identifizierbarkeit von Betrieben sollte man allerdings nicht vergessen, daß in der IAB-Beschäftigtenstichprobe eine mögliche Deanonymisierung von Betrieben nur über die Betriebsnummer möglich ist. Die im Datensatz abgespeicherten Betriebsnummern stimmen allerdings nicht unbedingt mit den tatsächlichen Betrieben oder verfügbaren Informationen über Betriebseinheiten überein (vgl. Kap. 1.4 und Kap. 3.2.2).

Aufgrund des Meldeverfahrens sind die Merkmale Wirtschaftszweig, Region und Betriebsnummer gesondert von den übrigen Merkmalen der IAB-Beschäftigtenstichprobe zu betrachten, da sie sowohl ein Personen- als auch ein Be-

[36] Einen Bayesschen Ansatz zur Berechnung des Reidentifikationsrisikos findet sich in den Arbeiten von Duncan/Lambert (vgl. Duncan/Lambert 1986, Duncan/Lambert 1989, Lambert 1993).

triebsmerkmal darstellen. Der Wirtschaftszweig und die regionalen Kennziffern werden über die Betriebsnummer den Sozialversicherungspflichtig Beschäftigten zugespielt (vgl. Kap. 1.4), so daß alle unter einer Betriebsnummer arbeitenden Personen dem gleichen Wirtschaftszweig angehören und am gleichen Ort arbeiten. Die Anonymisierung dieser Merkmale muß daher auf der Betriebsebene erfolgen.

Zusammenfassend kann gesagt werden, daß sich bzgl. der Querschnittsanonymisierung die IAB-Beschäftigtenstichprobe und der Mikrozensus nur im Datengenerierungsprozeß unterscheiden. Daher kann die Grundlogik der Empfehlungen der Mikrozensusanonymisierung auf die Querschnittsanonymisierung der IAB-Beschäftigtenstichprobe angewendet werden (vgl. Kap. 4.2.1). Bei der Querschnittsanonymisierung der Betriebe wird das gewählte Anonymisierungsverfahren von den Maßnahmen der vorgeschlagenen Personenanonymisierung abweichen (vgl. Kap. 4.3.4). Die Längsschnittanonymisierung geht von der Grundidee aus, den gesamten Erwerbsverlauf jeder Person um eine Konstante zu verschieben. Eine exakte zeitliche Verortung von Ereignissen würde somit erschwert sein, gleichzeitig bleiben aber alle exakten Dauern erhalten (vgl. Kap. 5.1).

Bei der Entwicklung der Anonymisierungsmaßnahmen werden die drei Anonymisierungsebenen (Personen-, Betriebs- und Längsschnittanonymisierung) aufeinander bezogen. So wird bei der Anonymisierung besonders darauf geachtet, daß keine der drei Anonymisierungsebenen eine Schwachstelle besitzt, da davon ausgegangen werden muß, daß die Deanonymisierung einer der drei Ebenen das Reidentifikationsrisiko der anderen beiden Ebenen erheblich erhöhen würde.

4.2 Die Querschnittsanonymisierung der Personenangaben

4.2.1 *Anonymisierungspraxis beim Mikrozensus*

Die Anonymisierungsmaßnahmen für den Mikrozensus basieren im wesentlichen auf Merkmalsvergröberungen und der eingeschränkten Weitergabe von Regionalinformationen:

- Regionalangaben werden nur vergröbert weitergegeben. Dies schließt Angaben über das Bundesland und eine Klassifikation des siedlungstrukturellen Typs ein.
- Angaben über Staatsangehörigkeit werden nur so aggregiert weitergegeben, daß eine Nationalität oder identifizierbare Gruppe von Nationalitäten in der Bundesrepublik Deutschland mindestens 50.000 Einwohner umfaßt.
- Alle übrigen („sichtbaren") Variablen sollen so aggregiert werden, daß in den univariaten Randverteilungen jede ausgewiesene Merkmalsausprägung für die Grundgesamtheit der Bundesrepublik Deutschland mindestens 5.000 Fälle umfaßt.

  Bei öffentlich wenig bekannten oder über die Zeit wenig stabilen, jedoch differenziert erfaßten Merkmalen, sollen die höchsten und niedrigsten Ausprägungen nur als Mittelwert dieser Ausprägungen ausgewiesen werden. Diese Anonymisierungsmaßnahme wird insbesondere für die Einkommens- und Verbrauchsstichprobe empfohlen, da in diesem Datensatz die Vermögenswerte und die Ausgabenbeträge von Haushalten sehr detailliert erfaßt werden (vgl. Müller et al. 1991:443f.).

422 Die Umsetzung der Anonymisierungsmaßnahmen bei der IAB-Beschäftigtenstichprobe

Als Grundgesamtheit werden alle zum 30.6. des jeweiligen Jahres Sozialversicherungspflichtig Beschäftigten[37] (ca. 40% der Einwohner der Bundesrepublik Deutschland) definiert; d.h. es werden bei allen „sichtbaren" Variablen (z.B. Berufsordnung) nur Merkmalsausprägungen ausgewiesen, deren univariate Randverteilung zum 30.6. des jeweiligen Jahres mindestens 5.000 Sozialversicherungspflichtig Beschäftigte umfaßt. Für die Querschnittsanonymisierung werden demnach für 16 Stichtage (30.6.1975 bis 30.6.1990) die univariaten Randverteilungen betrachtet.

Erfüllt eine Merkmalsausprägung zum 30.6. des jeweiligen Jahres das Anonymisierungskriterium nicht, so wird die Merkmalsvergröberung für alle Versicherungsmeldungen vorgenommen, unabhängig vom Zeitpunkt der Meldung. Das hier auferlegte Kriterium ist härter als die Mikrozensus-Regelung, da die Basis Wohnbevölkerung durch die Basis 'sozialversichert beschäftigt' ersetzt wird. Dieses strenge Anonymisierungskriterium wird deshalb gewählt, da man davon ausgeht, daß die Sozialversicherungspflicht ein leicht zugängliches bzw. gut sichtbares Merkmal ist.

Eine Ausnahme bildet die Anonymisierung der Staatsangehörigkeit, da hier nicht die Wohnbevölkerung durch die sozialversicherungspflichtig Beschäftigten ersetzt wird. Vielmehr wird für das Merkmal 'Staatsangehörigkeit' das Anonymisierungskriterium für die Ausweisung von Nationalitäten bzw. identifizierbare Gruppen von Nationalitäten auf 30.000 Personen festgesetzt. Dieses Anonymisierungskriterium ist allerdings immer noch härter als bei der Mikrozensusanonymisierung.

Betrachtet man die einzelnen Merkmale der IAB-Beschäftigtenstichprobe, so können die meisten Merkmale[38] *(Stellung im Beruf, Schul-/ Berufsausbildung, Versicherungsträger*[39] und *Geschlecht)* ohne Merkmalsvergröberungen weitergegeben werden, da sie die aufgestellten Anonymisierungskriterien erfüllen. Stärkere Merkmalsvergröberungen werden nur bei den Variablen Berufsordnung und Nationalität vorgenommen.

Berufskennziffer (Berufsordnung)

Von den 334 Berufen besitzen nach der Anonymisierung noch 234 ihre ursprüngliche Klassifikation. Die verbleibenden 100 Berufe werden zu 41 Berufsgruppen zusammengefaßt.

Um möglichst viele Informationen zu erhalten und die Vergleichbarkeit mit anderen Statistiken zu gewährleisten, erfolgt die Zusammenfassung von Berufen nicht nach einem festen Schema. Vielmehr wird in jedem einzelnen Fall geprüft, welche Zusammenfassung von Merkmalsausprägun-

[37] Die einzige Ausnahme bildet hierbei das Merkmal Einkommen, da es nicht als sinnvolle Querschnittsinformation zur Verfügung steht. Bei diesem Merkmal wird unmittelbar auf die IAB-Stichprobe zurückgegriffen.

[38] Da die Wirtschaftsklassenzuordnungen der Sozialversicherungspflichtig Beschäftigten direkt über die Betriebsnummer erfolgt, ist der Wirtschaftszweig für eine Betriebsnummer und für die Beschäftigten, die unter einer Betriebsnummer gemeldet sind, gleich. Bei den Wirtschaftsklassen hat sich die zulässige Datenweitergabe an den Kriterien für die Anonymisierung der Betriebsangaben zu orientieren. Es wäre sonst ein leichtes, die Anonymisierung der Wirtschaftsklassen hinsichtlich der Betriebe aufzuheben, indem diese einfach durch die anonymisierten Wirtschaftsklassen bzgl. der Personen ersetzt würden.

[39] Beim Merkmal Versicherungsträger sind im Laufe der Zeit Änderungen der zulässigen Merkmalsausprägungen vorgenommen worden. Seit 1981 werden nur noch zwei Versicherungsträger (vorher sieben) erfaßt. Die vor 1981 zusätzlich erfaßten Merkmalsausprägungen werden entsprechend zugeordnet und nicht extra ausgewiesen.

gen inhaltlich sinnvoll ist. Grundsätzlich wird versucht, die Aggregation innerhalb der Berufe (Dreisteller) vorzunehmen, d.h. es sollen nur Berufe zusammengefaßt werden, die zur gleichen Berufsgruppe (Zweisteller) gehören[40].

Staatsangehörigkeit (Nationalität)

Von den 188 ausgewiesenen Nationalitäten (Stand 1990) bleiben neun Nationalitäten (deutsch, griechisch, italienisch, jugoslawisch, portugiesisch, spanisch, türkisch, französisch und österreichisch) vollständig erhalten. Alle übrigen Nationalitäten werden in sieben Nationalitätengruppen (Benelux-Staaten, sonstige EG, sonstige Industriestaaten, osteuropäische Staaten, Afrika, Asien und übrige Welt) zusammengefaßt.

Bei drei Nationalitätengruppen ist das vorgeschlagene Anonymisierungskriterium verletzt. Da diese Nationalitätengruppen sehr unspezifisch sind ('Süd-/Mittelamerika und Ozeanien' , sowie 'staatenlos und Staatsangehörigkeit ungeklärt') und eine Reidentifikation dieser Merkmalsausprägungen nicht möglich ist, ist dies vertretbar. Die Nationalitätengruppe 'sonstige Industriestaaten' liegt in den Jahren 1975 bis 1978 nur knapp unterhalb des Anonymisierungskriteriums, so daß auch diese nicht noch weiter zusammengefaßt werden. Somit ist ein Vergleich der zusammengefaßten Nationalitätengruppen mit anderen Stichproben (z.B. Mikrozensus) möglich.

Geburtsjahr

Das Merkmal Alter bzw. Geburtsjahr weist unterhalb von 16 Jahren und oberhalb von 66 Jahren zu geringe Fallzahlen auf. Daher werden diese Altersangaben jeweils in einer Kategorie zusammengefaßt. Eine weitere Merkmalsvergröberung ist nicht notwendig, so daß das Alter in Jahresabständen vorliegt.

Grund der Meldung

Die Merkmalsausprägung 'Tod', die oft nicht korrekt gemeldet wird, weist sehr geringe Fallzahlen auf, so daß sie mit der Ausprägung 'Abmeldung' zusammengefaßt wird.

Durchschnittliches Tagesentgelt im Meldezeitraum

In der Stichprobe ist das Sozialversicherungspflichtige Bruttoentgelt bis zur Beitragsbemessungsgrenze der Rentenversicherung – für den jeweiligen Meldezeitraum – enthalten (Lohnsumme). Dieses wird auf ein durchschnittliches Bruttoentgelt pro Kalendertag umgerechnet (Bruttoentgelt im Meldezeitraum durch die Anzahl der Tage innerhalb des Meldezeitraums). Hierbei werden die Nachkommastellen abgeschnitten, so daß das Tagesentgelt mit DM-Genauigkeit vorliegt. Das anonymisierte Einkommen weicht vom Einkommen in der Historikdatei geringfügig ab (höchstens um 50 Pfennig pro Tag bzw. um 180 DM pro Jahr). Für Entgeltbeträge, die über bzw. unter der Beitragsbemessungsgrenze liegen, werden eindeutige Kennziffern ausgewiesen.

[40] Wir danken dem Forschungsbereich VII/4 des IAB für seine Unterstützung – insbesondere Herrn Troll – bei der Erstellung eines Vorschlages für die Merkmalsvergröberung Berufsordnung.

Versicherungs- bzw. Betriebsnummer

Die ursprünglichen Versicherungs- bzw. Betriebsnummern werden durch eine systemfreie Personen- bzw. Betriebsnummer ersetzt. Durch eine Umsortierung der gesamten IAB-Beschäftigtenstichprobe mit Hilfe einer Zufallsvariable wird gewährleistet, daß auch die interne Sortierung keine Anhaltspunkte für eine mögliche Reidentifikation bietet.

4.3 Die Anonymisierung der Betriebsangaben

4.3.1 Bestimmung des response knowledge für einzelne Betriebsgrößenklassen

Bei einer gezielten Suche ist es für einen potentiellen Datenangreifer von großer Bedeutung, ob er Kenntnisse darüber besitzt, ob der gesuchte Betrieb im Datensatz enthalten ist oder nicht. Ist dies der Fall, so erhöht sich die Reidentifikationwahrscheinlichkeit (vgl. Kap. 4.1) erheblich.

Um eine Abschätzung des response knowledge geben zu können, soll die Anzahl der Betriebe in der IAB-Beschäftigtenstichprobe für einzelne Betriebsgrößenklassen mit der Anzahl der in der aggregierten Beschäftigtenstatistik vorhandenen Betriebe verglichen werden. Es ist generell anzunehmen, daß mit steigender Betriebsgröße auch die Wahrscheinlichkeit wächst, in der IAB-Beschäftigtenstichprobe enthalten zu sein. Die „kritische" Betriebsgröße liegt bei etwa 300 sozialversichungspflichtig Beschäftigten, da ab diesem Wert response knowledge angenommen werden muß (vgl. Tab. 4). Dieser Sachverhalt muß bei der Beurteilung der späteren Anonymisierungsmaßnahmen berücksichtigt werden.

Tabelle 4: Wahrscheinlichkeit eines Betriebes für einzelne Betriebsgrößenklassen, in der IAB-Beschäftigtenstichprobe enthalten zu sein (in %)

Betriebsgröße	1980	1985	1989
5000-u.m.	100,0	100,0	100,0
2000-4999	100,0	99,8*	99,8
1000-1999	100,0	100,0	99,9
500-999	99,6	99,7	99,8
300-499	97,4	96,5	97,0
200-299	89,5	90,3	89,5
150-199	81,1	82,0	81,1
100-149	68,8	68,9	68,9
75-99	57,5	56,5	56,6
50-74	44,9	45,4	44,7
40-49	35,3	35,9	35,3
30-39	28,7	28,4	28,5
20-29	20,8	20,8	20,6
15-19	15,2	15,4	15,1
10-14	10,8	10,9	10,9
5-9	6,2	6,3	6,2
2-4	2,8	2,8	2,8
1	1,0	1,4	1,0

* Abweichungen sind auf nicht zulässige Betriebsnummern in der Quartalsdatei zurückzuführen.

4.3.2 Dateninkompatibilitäten bei Betriebsnummern

Wie schon in Kap. 1.4 erwähnt, müssen die Betriebsnummern, wie sie in der Beschäftigtenstatistik vorliegen, nicht unbedingt den tatsächlichen Betrieben entsprechen. Erkenntnisse über die Dateninkompatibilität zwischen Betriebsnummern in der Beschäftigtenstatistik und einer Betriebsnummernerhebung liegen durch das IAB-Betriebspanel vor (vgl. Projektgruppe Betriebspanel 1994). Allerdings muß einschränkend gesagt werden, daß diesem Projekt die Betriebsnummernproblematik bekannt war, und daher schon in der Kontaktaufnahme mit den Befragungseinheiten darauf geachtet wurde, die Betriebsnummer der Beschäftigtenstatistik in den Merkmalen Branchen- und Betriebsgrößenklasse möglichst exakt zu reproduzieren. Daher werden mögliche Dateninkompatibilitäten zwischen der Beschäftigtenstatistik und dem IAB-Betriebspanel – im Vergleich zu anderen Betriebsbefragungen – unterschätzt. Ein Vergleich der Ergebnisse des IAB-Betriebspanels mit den amtlichen Daten läßt aber trotzdem einige grobe Abschätzungen über mögliche Dateninkompatibilitäten zu.

Betrachtet man den Rücklauf des IAB-Betriebspanels (vgl. Infratest 1994), so waren von den 6.237 kontaktierten Betriebsnummern 2,5% (n=159) nicht mehr existent[41]. Von den 6.132 existierenden, und somit befragungsfähigen Betrieben waren 1,3% (n=79) nicht auffindbar, bei 0,7% (n=40) die Betriebsnummer nicht identifizierbar, bei 1,6% (n=96) die Zentrale des Betriebs für die Betriebsnummer zuständig[42].

Eine Übereinstimmung zwischen den Ergebnissen der Befragung und der Betriebsdatei der Beschäftigtenstatistik wurde hinsichtlich der Merkmale „Betriebsgrößenklasse der sozialversicherungspflichtig Beschäftigten“[43] und „Branchenklasse“[44] ausgewiesen. Eine Übereinstimmung zwischen Branchenklasse und Betriebsgrößenklasse ergab sich in 83% aller Fälle (n=3.599). Eine falsche Betriebsgrößenklasse war in 5% (n=206), eine falsche Branche in 3% (n=139) und eine Abweichung beider Dimensionen war in 7% (n=321) aller Fälle feststellbar. In 2% aller Fälle (n=91) war die Betriebseinheit bzw. die Betriebsgröße nicht feststellbar (vgl. Bellmann et al. 1994:10).

Dateninkompatibilitäten der Wirtschaftsklassenbezeichnungen lassen sich auch mit Hilfe der Beschäftigtenstatistik feststellen. Bei ca. 0,2% aller Betriebe in 1990 ist eine Änderung der Wirtschaftsbezeichnung zum letzten Jahr (2.959 von 1.540.126 Betriebsnummern), bei ca. 0,2% aller Betriebe in 1990 ist eine Änderung der Wirtschaftsbezeichnung zum folgenden Jahr (3.485 von 1.540.126) zu verzeichnen[45].

Somit ist mit einem hohen Grad an Dateninkompatibilitäten zwischen der IAB-Beschäftigtenstichprobe und anderen Betriebsdatensätzen zu rechnen. Diese Dateninkompatibilitäten würden die Deanonymisierungsversuche eines möglichen Datenangreifer erheblich unsicher machen.

4.3.3 Bestimmung des Deanonymisierungspotentials der Betriebsangaben

Informationen über Betriebe liegen in der IAB-Beschäftigtenstichprobe aus drei unterschiedlichen Datenquellen vor:

1. Die Merkmale Region und Wirtschaftsklasse werden über die Betriebsnummer aus der Betriebsdatei der Bundesanstalt für Arbeit zu den einzelnen Sozialversicherungspflichtig Beschäftigten zugespielt (vgl. Kap. 1.4).

2. Die IAB-Beschäftigtenstichprobe verfügt über Merkmale, die den Quartalsdateien der Beschäftigtenstatistik entstammen (Grundgesamtheitsaggregation).

3. Theoretisch sind auf der Basis der IAB-Beschäftigtenstichprobe Aggregationen durchführbar, da die Betriebsnummer als Merkmal in der IAB-Beschäftigtenstichprobe zur Verfügung steht (Stichprobenaggregation).

Eine Stichprobenaggregation bietet nur dann ein Potential zu einer möglichen Reidentifikation, wenn die Abweichungen einzelner Merkmale zwischen einer Stichproben- und einer Grundgesamtheitsaggregation gering sind. Diese Form der Betriebsinformationsgewinnung würde dann ein hohes Potential für mögliche Deanonymisierungen darstellen, da dann jede beliebige Merkmalskombination aus der IAB-Beschäftigtenstichprobe zur Deanonymisierung herangezogen werden könnte. Im Rahmen des Projekts wurde dieser Sachverhalt überprüft, sowohl durch direkte Vergleiche als auch über Rangvergleiche[46] von Merkmalen.

Für eine Betriebsgröße ab 500 Sozialversicherungspflichtig Beschäftigten sind Merkmale, die durch eine Stichprobenaggregation gewonnen werden, von großem Nutzen für einen potentiellen Datenangreifer. Da gleichzeitig response knowledge für diese Betriebsgrößen vorliegt, wird auf eine Ausweisung von Regionalinformationen ab einer Betriebsgröße von 500 Sozialversicherungspflichtig Beschäftigten verzichtet.

Durch die Rangplatzuntersuchungen wird klar, daß in einigen Wirtschaftszweigen sehr leicht exponierte Großbetriebe zu identifizieren sind. Diese Betriebe werden unkenntlich gemacht, indem sie mit anderen Betriebsnummern künstlich zusammengelegt bzw. zerlegt werden (vgl. Kap. 4.3.8).

4.3.4 Darstellung der Anonymisierungsregeln

Grundlage der Anonymisierungsmaßnahmen von Betrieben ist eine Kreuztabelle, die die Merkmale Wirtschaftszweig und Regionalmerkmal in Abhängigkeit von der Betriebsgröße beinhaltet.

Für Betriebe mit weniger als 500 sozialversicherungspflichtig Beschäftigten wird dabei eine dreidimensionale Tabelle mit der Merkmalskombination Wirtschaftszweig * Betriebsgröße * Regionalinformation erstellt (vgl. Tab. 5). Für Betriebe ab 500 Sozialversicherungspflichtig Beschäftigte wird auf eine zweidimensionale Kreuztabelle zurückgegriffen.

---

[41] Zwischen der Ziehung der Betriebsnummer aus der Beschäftigtenstatistik und der eigentlichen Befragung liegen fast drei Monate.

[42] Weitere 1,1% (n=68) fielen aus nicht näher genannten Gründen aus.

[43] Die Größenklassen umfassen folgende Sozialversicherungspflichtige Beschäftigungszahlen: 1-9, 10-19, 20-49, 50-99, 100-499, 500-999, über 1000.

[44] Die Branchenklasse gliedert sich wie folgt: Land- und Forstwirtschaft (00-03); Energie- und Wasserversorgung (04-08); Grundstoffindustrie (09-22); Investitionsgüterindustrie (23-39); Verbrauchsgüterindustrie (40-58); Bauhaupt- und Ausbaugewerbe (59-61); Groß-, Einzel- und Versandhandel (620-625); Verkehr (63-68); Kredit- und Finanzwesen (690); Versicherung (691); Gaststätten, Heime und Reinigung (70-73); Hochschulen, Schulen und Kultur (74-77); Gesundheitswesen (780-785); Freiberufl. Dienstleistungen (79-85); sonst. Dienstleistungen (86); Organisationen ohne Gebietskörperschaft, Sozialversicherungen (87-94).

[45] Genaugenommen sind auch tatsächliche Wirtschaftszweigwechsel vorstellbar, so daß sich die Angaben von Wechseln in tatsächliche Wechsel und Dateninkompatibilitäten aufgliedern. Das genaue Verhältnis ist allerdings nicht bekannt.

[46] Ränge stellen leicht bildbare Zusatzinformationen dar, die anonymisierte Merkmale aufheben können. Ein Beispiel wäre das Merkmal „Größe des Betriebes“, da sich durch die einfache Rangbildung (z.B. „größter Betrieb“) das anonymisierte Merkmal „Betriebsgröße“ teilweise wieder aufheben läßt. Für diesen Hinweis danken wir Bernhard Schimpl-Neimanns (ZUMA).

Tabelle 5: Anonymisierungsgrundlage für Betriebe mit weniger als 500 Beschäftigten

Wirtschaftsbereich	Betriebsgrößenklasse (Anzahl der Beschäftigten am 30.06.)																	
	1			2-9			10-19			20-49			50-99			100-499		
	Regionaltyp			Regionaltyp			Regionaltyp			Regionaltyp			Regionaltyp			Regionaltyp		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
01																		
02																		
03																		
...																		

Für den Aufbau beider Tabellen ist es nicht unerheblich, in welcher Kategorisierung die einzelnen Merkmale vorliegen. Für den Wirtschaftszweig wird von den Wirtschaftsbereichen (94 Zweisteller) ausgegangen. Die Betriebsgrößenklassen liegen in acht Kategorien vor (1, 2-9, 10-19, 20-49, 50-99, 100-499, 500-999, ab 1000 Sozialversicherungspflichtig Beschäftigte). Als Regionalangabe ist die Gebietstypisierung der Bundesforschungsanstalt für Landeskunde und Raumordnung (BfLR) enthalten. Es handelt sich hierbei um eine flächendeckende Raumtypisierung, die nicht nur nach Verdichtungsansätzen vorgenommen wird, sondern auch nach dem Kriterium der „Zentralität". Die Typisierung der BfLR dient der vergleichenden Beobachtung der regionalen Situation und Entwicklung derjenigen Lebens- und Arbeitsbereiche, die eine Abhängigkeit von der Siedlungsstruktur vermuten lassen (Görmar/Irmen 1991:387). Allerdings muß gesagt werden, daß die verwendeten Merkmalsausprägungen sich auf drei Kategorien, nämlich Regionen mit großen Verdichtungsansätzen, Regionen mit Verdichtungsansätzen und ländlich geprägte Regionen, beschränken.

In diesen beiden Kreuztabellen werden nun die einzelnen Zellbesetzungen betrachtet. Findet sich eine Zellbesetzung mit weniger als drei Betrieben, so werden die Merkmalskombinationen geeignet vergröbert. Die Betriebe können dann als geschützt gelten, wenn die Zellenbesetzung der dreidimensionalen Tabelle in jedem Jahr mindestens drei beträgt. Dies gilt auch für die Randverteilung (entspricht der zweidimensionalen Tabelle Wirtschaftszweig * Betriebsgrößenklasse). Die Vergröberung der Merkmalskategorien wird demnach solange fortgeführt, bis die Anzahl der Betriebe in jeder Zelle mindestens den Wert drei erreicht.

Für die zwei bzw. dreidimensionale Kreuztabelle wird eine unterschiedliche Heuristik für die Vergröberung der Merkmale angewendet.

Bei den Betriebsgrößenklassen unter 500 Beschäftigten werden in einem ersten Schritt benachbarte Betriebsgrößenklassen zusammengefaßt, da durch eine Stichprobenaggregation die tatsächliche Betriebsgrößenklasse nicht hinreichend genau zu schätzen ist. Gleichzeitig bleiben die beiden anderen Merkmale Wirtschaftszweig und Regionalangabe über die einzelnen Betriebsgrößenklassen konstant erhalten, wodurch das Problem von Betrieben, die im Zeitverlauf die Betriebsgrößenklasse wechseln, (vgl. Kap. 4.3.7) vermindert wird.

Ab einer Betriebsgröße über 500 sozialversicherungspflichtig Beschäftigter bietet die Zusammenfassung von benachbarten Betriebsgrößenklassen keinen ausreichenden Deanonymisierungsschutz, da sich die zugehörige Betriebsgrößenklasse zu einer Betriebsnummer durch eine Stichprobenaggregation hinreichend genau schätzen läßt. Es werden daher überwiegend Wirtschaftsbereiche zusammengefaßt.

Für die Zusammenfassung von Wirtschaftsbereichen (Zweisteller) sind eine Vielzahl von Möglichkeiten denkbar. Neben der Zusammenfassung innerhalb von Wirtschaftsabteilungen (Einsteller) wird im IAB noch auf eine spezielle Zusammenfassung der Wirtschaftszweige zurückgegriffen, die zur Beschreibung des sektoralen Strukturwandels entwickelt wurde (vgl. Dietz 1988). Diese erlaubt Zusammenfassungen innerhalb von Wirtschaftsabteilungen (l -Steller) nach Art der erzeugten Güter bzw. der erbrachten Dienstleistungen (vgl. Dietz 1988:151)[47]. Diese Systematik ermöglicht weiterhin eine sinnvolle Zusammenfassung innerhalb der Wirtschaftsabteilung „Verarbeitendes Gewerbe".

Daneben wurde die Zusammenfassung von Wirtschaftsbereichen so gewählt, daß sich die Anzahl der Betriebe, die zerlegt werden müssen (vgl. Kapitel 4.3.8), in vertretbarem Rahmen hält.

4.3.5 Anonymisierungsmaßnahmen bei Betrieben mit weniger als 500 Sozialversicherungspflichtig Beschäftigten

Ausgangspunkt für die Anonymisierung von Betrieben unter 500 Sozialversicherungspflichtig Beschäftigten ist die folgende dreidimensionale Tabelle[48]:

*Wirtschaftszweig (2-Steller) * siedlungsstruktureller Typ der BFLR (3 Ausprägungen) * Betriebsgrößenklasse (6 Ausprägungen)*

Zur Anonymisierung von Betrieben wird folgende Heuristik angewendet:

1. Ist die einzelne Zellenbesetzung der dreidimensionalen Tabelle kleiner als drei und die Randverteilung größer als zwei, so wird auf die Ausweisung des siedlungsstrukturellen Typs der BFLR ganz oder teilweise verzichtet.

2. Ist die Randverteilung kleiner als drei, so werden zunächst benachbarte Betriebsgrößenklassen zusammengefaßt, danach wird nochmals geprüft, ob die Zellenbesetzung das oben genannte Kriterium erfüllt (ggf. weiter mit L).

Auf die Ausweisung des siedlungsstrukturellen Typs der BFLR wird ganz verzichtet, wenn bei allen drei Ausprägungen des siedlungsstrukturellen Typs die Zellenbesetzung der dreidimensionalen Tabelle zu gering ist oder wenn die Zusammenfassung von zwei Ausprägungen des siedlungsstrukturellen Typs nicht ausreicht, um eine Zellenbesetzung von drei zu erhalten. Ausnahmen werden hier zugelassen, wenn die notwendige Zellenbesetzung für weit zurückliegende Jahre (vor 1984) nicht erfüllt ist.

[47] Wir danken Herrn Walter Müller für wertvolle Hinweise bei der Kategorisierung der Wirtschaftszweige.

[48] Die Festlegung auf diese Tabelle als Ausgangspunkt für die Anonymisierung von Betrieben erfolgte auf einer Arbeitsgruppensitzung von IAB (Stefan Bender, Jürgen Hilzendegen), WZB (Roland Habich), ZUMA (Bernhard Schimpl-Neimanns) am 15.10.1994 in Nürnberg

Insgesamt werden bei fünf von 94 Wirtschaftsklassen[49] benachbarte Betriebsgrößenklassen zusammengefaßt und bei ca. 10% der Zellen der dreidimensionalen Tabelle wird auf die Ausweisung des siedlungsstrukturellen Typs verzichtet.

4.3.6 Anonymisierungsmaßnahmen bei Betrieben mit mehr als 500 Sozialversicherungspflichtig Beschäftigten

Als Grundlage für die Anonymisierung von Betrieben ab 500 Sozialversicherungspflichtig Beschäftigten dient die zweidimensionale Tabelle:

*Wirtschaftszweig (2-Steiler) * Betriebsgrößenklasse (2 Ausprägungen)*

Ausgehend von dieser Tabelle können Betriebe (Betriebsnummern), dann als geschützt gelten, wenn die Zellenbesetzung der zweidimensionalen Tabelle in jedem Jahr mindestens drei beträgt. Ist dies nicht der Fall, werden Wirtschaftsbereiche zusammengefaßt.

Bei der Betriebsgrößenklasse 500-999 Beschäftigte werden insgesamt 21 von 94 Wirtschaftsbereichen zusammengefaßt. Bei der Betriebsgrößenklasse ab 1000 Beschäftigte erhöht sich die Zahl der zusammengefaßten Wirtschaftsbereiche auf 41.

4.3.7 Wechsel der Betriebsgrößenklassen

Einen Sonderfall stellen Betriebe dar, die im Laufe der Zeit die Betriebsgrößenklasse wechseln. Da für verschiedene Betriebsgrößenklassen unterschiedliche Anonymisierungsregeln bzw. Zusammenfassungen von Wirtschaftsbereichen gelten, sind diese Betriebsgrößenwechsel zu untersuchen. So ergibt z.B. ein Wechsel von der Größenklasse unter 500 sozialversicherungspflichtig Beschäftigte in die Klasse über 500 Sozialversicherungspflichtig Beschäftigte eine nicht zulässige Regionalinformation, die unmittelbar das Reidentifikationsrisiko steigen läßt.

Bei Betrieben, die die Größenklasse wechseln, wird deshalb die Anonymität wie folgt gesichert:

1. Wird ein Betrieb einmal in einer Zelle der dreidimensionalen Tabelle ausgewiesen, in der kein siedlungsstruktureller Typ angegeben wird, so wird bei diesem Betrieb kein siedlungsstruktureller Typ ausgewiesen (betrifft nur Betriebe unter 500 Sozialversicherungspflichtig Beschäftigte).

2. Die Ausweisung von Wirtschaftsbereichen erfolgt für jeden Betrieb über alle Jahre nach der Wirtschaftsklassifikation der größten Betriebsgrößenklasse. Dies bedeutet, daß z.B. ein Betrieb des Wirtschaftsbereiches 67, der einmal in der Betriebsgrößenklasse 1000 u.m. und sonst in der Betriebsgrößenklasse 500-1000 enthalten ist, in jedem Jahr in den Wirtschaftsbereich 65 klassifiziert wird, da ab einer Betriebsgröße von 1000 die zwei Wirtschaftsbereiche zusammengefaßt werden.

43.8 Eindeutig identifizierbare Betriebe

In der IAB-Beschäftigtenstichprobe sind Betriebe ein Problem, deren Reidentifikation aufgrund ganz bestimmter mittelbarer Identifikatoren möglich ist. Durch eine Stichprobenaggregation (vgl. Kap. 4.3.3) ist es möglich, den größten Betrieb in einem Wirtschaftszweig zu identifizieren, wenn dieser einen ausreichenden Abstand bzgl. der Betriebsgröße zu den übrigen Betrieben besitzt. Eine Identifikation kann sehr leicht über eine Rangplatzanalyse oder über eine graphische Darstellung von Größenverläufen von Betrieben erfolgen (zu Einschränkungen vgl. Kap. 3.2.2).

Solche eindeutig identifizierbaren Betriebe werden entweder künstlich zerlegt oder es werden in diesem Wirtschaftsbereich Betriebe zusammengelegt. Es muß darauf hingewiesen werden, daß bei der Zerlegung bzw. dem Zusammenlegen von Betriebsnummern nicht mehr die originäre Einheit „Betriebsnummer“ erhalten ist. Da aber nach Möglichkeit die originäre Stichprobeneinheit erhalten werden sollte (vgl. Laaksonen 1995:7), wird für einige Betriebe – anstelle einer Zerlegung bzw. Zusammenlegung von Betriebsnummern - eine niedrigere Betriebsgrößenklasse ausgewiesen. Es wird hierbei aber sichergestellt, daß diese Anonymisierung durch eine Stichprobenaggregation nicht wieder rückgängig gemacht werden kann. Durch dieses Vorgehen sind bei über 99% aller Betriebsnummern die anonymisierte und die tatsächliche Betriebsnummer identisch.

Einen zweiten Problemkreis stellt das Streik- bzw. Aussperrungsjahr 1984 dar. Betrachtet man die Größenentwicklung einzelner Betriebe über die Jahre, so ist ein Abfallen der Betriebsgröße von bestimmten Betrieben in bestimmten Wirtschaftszweigen und Regionen feststellbar. Da diese „Beschäftigungseinbrüche“ auch durch eine Stichprobenaggregation zu erkennen sind, werden auf der Personenebene Streikende bzw. Ausgesperrte im Jahre 1984 auf Beschäftigung gesetzt. Diese Datenbereinigung stellt keinen Fehler dar, da sonst Streikende bzw. Ausgesperrte für diesen Zeitraum als nicht sozialversicherungspflichtig Beschäftigte gezählt werden würden.

5 Längsschnittanonymisierung

5.1 Wahl des Verfahrens

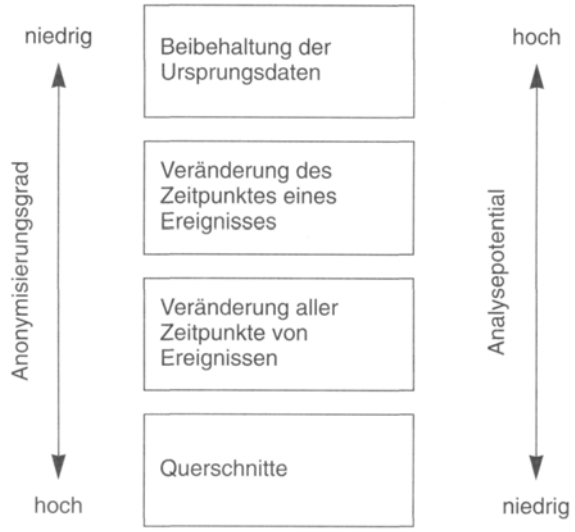
Die IAB-Beschäftigtenstichprobe beinhaltet von 426.363 Versicherungsnummern Informationen, die den Zeitraum vom l. l. 1975 bis 31.12.1990 abdecken. Die tagesgenaue Datierung von meldepflichtigen Ereignissen wird unmittelbar aus dem Meldeverfahren gewonnen und unterscheidet sich damit wesentlich von allen anderen sozialwissenschaftlichen Datensätzen, die zum großen Teil auf den Erinnerungsleistungen der Befragten basieren (z.B. Retrospektivbefragung). Die IAB-Beschäftigtenstichprobe ist somit einer der wenigen Datensätze, in denen datumsgenaue Dauern von bestimmten Zuständen (z.B. Betriebszugehörigkeit) über einen längeren Beobachtungzeitraum verfügbar sind. Die Längsschnittanonymisierung sollte nach Möglichkeit diese Besonderheiten erhalten.

Es sind eine Reihe von Anonymisierungsmaßnahmen (verschiedenste Anonymisierungstechniken sind in Kelly-McNulty/Unger 1993:487ff. bzw. Little 1993 zu finden) bzgl. des Längsschnitts vorstellbar, so daß kurz auf einige Alternativen eingegangen werden soll, die im Rahmen dieses Anonymisierungsprojektes realisiert werden könnten. Diese Möglichkeiten spannen ein Spektrum auf, das durch ein abnehmendes Analysepotential bei gleichzeitigem Anstieg des Anonymisierungsgrades gekennzeichnet ist (vgl. Abb. 2).

Das geringste Analysepotential hat die Zerlegung des Datensatzes in einzelne Querschnitte (z.B. Monatskalendarien). Bei dieser Art der Anonymisierung würde man allerdings alle An-

[49] So wird z.B. bei dem Wirtschaftsbereich 06 (Erzbergbau) nur noch eine Betriebsgrößenklasse ausgewiesen (0-499 Beschäftigte)

Abbildung 2: Der Zusammenhang von Anonymisierungsgrad und Analysepotential bei verschiedenen Längsschnittanonymisierungen



gaben über Dauern zerstören, so daß die IAB-Beschäftigtenstichprobe nur sehr bedingt für Längsschnittanalysen verwendet werden könnte.

Bei der Veränderung aller Beginn- und Endzeitpunkte jedes einzelnen Ereignisses ist die Gefahr sehr groß, daß die bestehende Datenstruktur (speziell die Korrelationsstruktur) zerstört wird. Analysen, die diesen Sachverhalt berücksichtigen, wären mit einem erheblichen statistischen Aufwand verbunden, so daß ein „Normalnutzer" den so anonymisierten Datensatz nicht handhaben könnte.

Eine mögliche Längsschnittanonymisierung, die sowohl eine ausreichende Anonymisierung als auch ein ausreichendes Analysepotential gewährleistet, ist die einmalige zeitliche Verschiebung des gesamten Erwerbsverlaufs jeder einzelnen Person. Hierbei würden alle Angaben für jede einzelne Person auf der Zeitachse um einen konstanten Betrag nach vorne oder hinten verschoben.

Eine exakte zeitliche Verortung aller Ereignisse einer Person würde dadurch erschwert werden, gleichzeitig blieben aber alle weiteren Dauern tagesgenau[50] erhalten. Somit würde die anonymisierte IAB-Beschäftigtenstichprobe bzgl. Dauerbetrachtungen wie die Originalangaben auswertbar sein, allerdings wäre der tagesgenaue Bezug auf einen Querschnitt nur unsicher.

Eine Deanonymisierung des Längsschnitts wird zusätzlich dadurch erschwert, daß Ereignisse, die aufgrund des Meldeverfahrens in der IAB-Beschäftigtenstichprobe vorliegen, nicht unbedingt mit tatsächlichen bzw. wahrgenommenen Ereignissen übereinstimmen. Auch hängt die Deanonymisierung von Personen über die Dauer dieser Ereignisse u.a. von der Erinnerungsleistung des Datenangreifers oder seiner Quellen ab. Aus diesen Gründen erscheinen die zeitlichen Verschiebungen des Erwerbsverlaufs um eine Konstante ausreichend für die Längsschnittanonymisierung.

[50] Mit Ausnahme der ersten oder letzten Dauer, von denen ein Teil bereits aufgrund der Konstruktion der Datensätze links- oder rechtszensiert ist.

[51] Eine zweite Längsschnittanonymisierung stellt die Bereinigung von Streiks/Aussperrungen dar (vgl. Kap. 4.3.8).

[52] Aus Gründen des Datenschutzes wird der Wert der Verschiebekonstante nicht veröffentlicht.

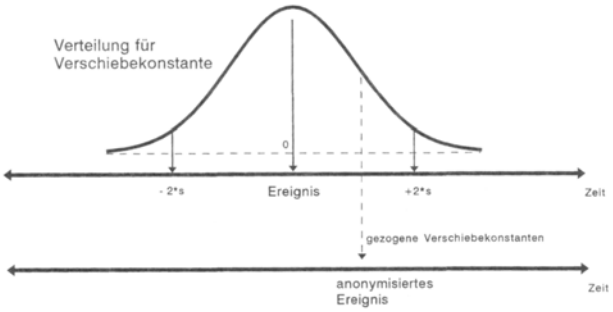
5.2 Konkretisierung der Längsschnittanonymisierung

Im Meldeverfahren sind meldepflichtige von nicht-meldepflichtigen Ereignissen zu unterscheiden. Nicht-meldepflichtige Ereignisse müssen ex definitione nicht unmittelbar angezeigt werden und werden zum Jahresende mit der Jahresmeldung oder im Zusammenhang mit meldepflichtigen Vorgängen gemeldet (vgl. Bender/Dietrich 1994, Veiling/Bender 1994). Dies bedeutet, daß nur für meldepflichtige Ereignisse in der IAB-Beschäftigtenstichprobe Tagesgenauigkeit vorliegt.

Als meldepflichtige Ereignisse gelten der Betriebswechsel und der Beginn bzw. das Ende von Sozialversicherungspflichtiger Beschäftigung. Alle weiteren möglichen Änderungen (z.B. Berufswechsel, Wechsel der beruflichen Stellung, Lohnerhöhung) sind keine meldepflichtigen Ereignisse. Die zugespielten Informationen aus der Leistungsempfängerdatei, insbesondere Beginn und Ende von Leistung, besitzen ebenfalls Tagesgenauigkeit und werden deshalb wie meldepflichtige Ereignisse behandelt.

Für die Längsschnittanonymisierung[51] ist es sinnvoll, die vorgegebene Unterscheidung in meldepflichtige bzw. nicht-meldepflichtige Ereignisse zu übernehmen. Eine Längsschnittanonymisierung sollte also die exakte zeitliche Verortung von tagesgenauen (meldepflichtigen) Ereignissen verhindern. Meldepflichtige Ereignisse liegen bei fast allen Personen vor. Jeder Erwerbsverlauf eines Sozialversicherungspflichtig Beschäftigten wird einmalig um einen konstanten Betrag verschoben. Als Verteilung für die Verschiebekonstante wird die Normalverteilung gewählt. Der Wertebereich der Verteilung reicht von minus unendlich bis plus unendlich, allerdings liegen 99,0% der Verteilung in 2,57-facher Standardabweichung vom Erwartungswert. Die individuelle Verschiebungskonstante einer jeden Person wird durch das Ziehen einer normalverteilten Zufallsgröße festgelegt. Der Erwartungswert der Normalverteilung ist Null, die Standardabweichung[52] legt den Verschiebebereich der Konstanten fest (vgl. Abbildung 3).

Abbildung 3: Darstellung der Längsschnittanonymisierung



Durch die Längsschnittanonymisierung werden Querschnittsanalysen, die an bestimmte historische Zeitpunkte gebunden sind, problematisch. „Sind jedoch mit dem Stichtag massive Effekte verbunden – etwa saisonale Effekte auf die Arbeitslosigkeit in bestimmten Wirtschaftszweigen – dann wird dieser Effekt durch die Anonymisierung im Zeitverlauf 'verschmiert' und erscheint um so weniger prägnant, je größer die Varianz der Zufallsverschiebung ist." (Wiedenbeck/ Schimpl-Neimans 1994: 1f)

Eine Überprüfung der Robustheit von Ereignisanalysen mit obigen Verteilungsparametern wurde von ZUMA durchge-

führt (vgl. Wiedenbeck/Schimpl-Neimanns 1994 und Schimpl-Neimanns 1995). Hierzu wurden exponentiell verteilte Wartezeiten in verschiedenen homogenen Klassen angenommen und zwei Modelle geschätzt:

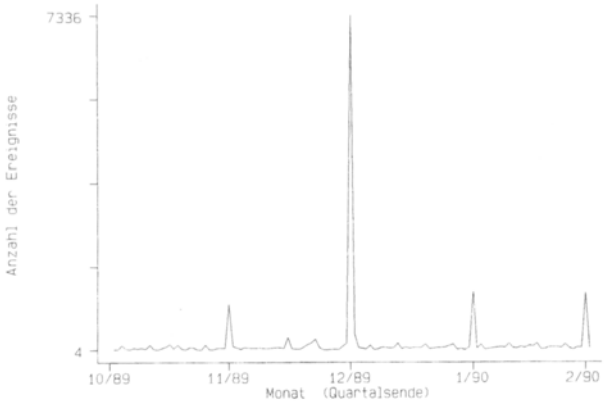
1. Eine Schätzung für jede Klasse.
2. Schätzung in einem gemeinsamen Regressionsansatz, wobei für die klassenweisen Raten ein log-lineares Modell angenommen wurde.

Die Autoren kommen bzgl. der Robustheit von Ereignisanalysen zu dem folgenden Ergebnis: „Die Simulationsergebnisse deuten darauf hin, daß sich die Anonymisierung nicht gravierend auf die Schätzung der Parameter auswirkt. Die Befunde beziehen sich jedoch nur auf die gewählte Modellklasse und die vorgegebene Verteilung künstlicher Daten."

5.3 Empirisches Beispiel

Im folgenden soll die Längsschnittanonymisierung praktisch demonstriert werden. Dazu werden alle meldepflichtigen Ereignisse (Betriebswechsel und Beginn bzw. Ende von Sozialversicherungspflichtiger Beschäftigung) für einen viermonatigen Zeitraum (11/89 bis 2/90) aus der IAB-Beschäftigtenstichprobe verwendet. In diesen Zeitraum fallen 17.755 meldepflichtige Ereignisse. Die meisten Ereignisse finden sich am Jahresende (12/89) mit 7.736 (vgl. Abbildung 4). Drei kleinere Peaks sind zu den jeweiligen Monatsenden zu erkennen.

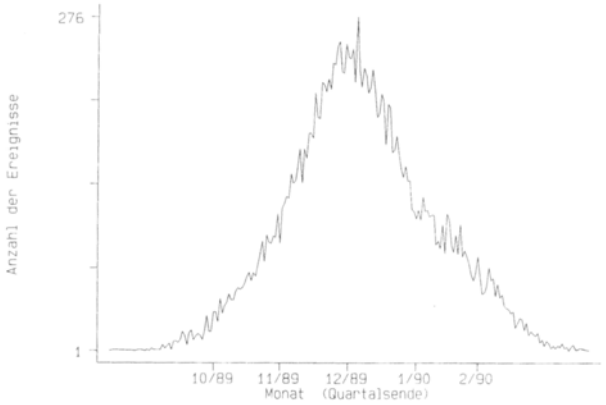
Abbildung 4: Ursprüngliche Verteilung von Betriebswechseln und Erwerbsunterbrechungen – Nov. 89 bis Feb. 90 (absolut)



Nun wird zu jedem meldepflichtigen Ereignis eine gerundete Konstante aus der gewählten Normalverteilung addiert. Aus der tatsächlichen Verteilung der meldepflichtigen Ereignisse (Abbildung 4) wird so eine anonymisierte Verteilung der meldepflichtigen Ereignisse (vgl. Abbildung 5).

Die Gesamtverteilung der anonymisierten meldepflichtigen Ereignisse wird von dem Stauchen des Jahrespeaks dominiert[53], d.h. auf der einen Seite ist der ehemalige Jahrespeak erkennbar, auf der anderen Seite erscheint aber eine exakte Zuordnung aller Personen unmöglich, die vor der Anonymisierung auf dem Jahrespeak lagen. Bezogen auf Querschnittsauswertungen wird durch dieses Beispiel noch einmal deutlich, daß historische Zeitpunkte, wie die Häufung von meldepflichtigen Ereignisse am Jahresende, durch die Längsschnittanonymisierung „verschmiert" werden.

Abbildung 5: Anonymisierte Verteilung von Betriebswechseln und Erwerbsunterbrechungen – Nov. 89 bis Feb. 90 (absolut)



6 Schluß

Die anonymisierte IAB-Beschäftigtenstichprobe wird dem Zentralarchiv in Köln als Rohdatensatz übergeben. Es ist geplant[54], daß der interessierte Forscher einen Antrag mit einer Beschreibung des Forschungsvorhabens an das Zentralarchiv sendet, in dem die Verwendung der IAB-Beschäftigtenstichprobe beschrieben sein muß. Das IAB wird dann über den Antrag entscheiden. Der Datenempfänger muß weiterhin vertraglich zusichern, jeden Reidentifikationsversuch zu unterlassen und durch geeignete technische und organisatorische Maßnahmen sicherzustellen, daß externe Zugriffe verhindert und interne Deanonymisierungsversuche rechtzeitig erkannt werden.

Die anonymisierte Beschäftigtenstichprobe wird von Götz Rohwer (Universität Bremen) in das Programmpaket TDA implementiert (vgl. Rohwer 1994). Dies hat für den Nutzer mehrere Vorteile:

- Die IAB-Beschäftigtenstichprobe liegt in gepackter Form vor.
- Der Datensatz liegt für Analysen aufbereitet vor.
- Der Nutzer bekommt anhand von Beispielprogrammen einen leichten Zugang zum Datenmaterial.
- In TDA sind eine Reihe von analytischen Hauptanwendungen (z.B. Ereignisanalyse) für die IAB-Beschäftigtenstichprobe vorhanden.
- Da das Sozioökonomische Panel in RZOO – das Retrievalprogramm von TDA – implementiert ist (vgl. Pischner/Rohwer 1994), wird für einen Teil der Nutzer die „Zugangsbarierre" gering sein.

Eine zentrale Anonymisierungsmaßnahme der jetzigen IAB-Beschäftigtenstichprobe ist die eingeschränkte Weitergabe von Regionalinformationen. Dies bedingt natürlich, daß der weitergegebene Datensatz für wissenschaftliche Fragestellungen, die differenzierte Regionalinformationen benötigen, wenig geeignet ist. Daher wird in Kooperation und mit Unterstützung durch Mittel der GESIS und dem WZB im IAB ein anonymisiertes Regionalfile aus der IAB-Beschäftigtenstichprobe erstellt. Dieses Projekt wird 1996 abgeschlossen sein.

[53] Das Ausklingen der Verteilung an den Enden (10/89 und 3/90) hängt von der Begrenztheit des ausgewählten Zeitraums ab, da keine Ereignisse außerhalb des viermonatigen Zeitraums in den betrachteten Zeitraum geschoben werden können.

[54] Das genaue Verfahren wird z.Z. noch erarbeitet.

Es ist weiterhin geplant, künftig in Abständen von mehreren Jahren die IAB-Beschäftigtenstichprobe zu aktualisieren bzw. eine neue Stichprobe zu ziehen und zu anonymisieren. Diese kann dann auch Daten aus den neuen Bundesländern enthalten. Bis zu einer Aktualisierung werden sicher eine Reihe von Erfahrungen mit der jetzigen Datei vorliegen, die dann berücksichtigt werden können. Mit Abschluß dieses Projekts und der Bereitstellung der IAB-Beschäftigtenstichprobe besteht nun für die empirische Sozialforschung die Möglichkeit, Daten der Beschäftigtenstatistik zu verwenden.

Literatur

Alba, R./Müller, W./Schimpl-Neimanns, B. (1994): Secondary Analysis of Official Microdata. In: Borg, I./ Mohler, P. Ph. (Hrsg.): Trends and Perspectives in Empirical Social Research, Berlin: 57-78.

Bender, S./ Dietrich, H. (1995): Berufliche Mobilität von männlichen Absolventen aus dem Dualen System. Manuskript, Nürnberg.

Bundesanstalt für Arbeit (1973): Verzeichnis der Wirtschaftszweige für die Statistik der Bundesanstalt für Arbeit, Nürnberg.

Bundesanstalt für Arbeit (1981): Schlüsselverzeichnis für die Angaben zur Tätigkeit in den Versicherungsnachweisen, Nürnberg.

Bellmann, L./ Kohaut, S./ Kühl, J. (1994): The Establishment Panel of the German Institute for Empoyment Research. Manuskript, Nürnberg.

Bethlehem, J.G./Keller, W.J./Pannekoek, J. (1990): Disclosure Control of Microdata. In: Journal of the American Statistical Association, 85, 38-45.

Boeri, T./ Cramer, U. (1991): Betriebliche Wachstumsprozesse: Eine statistische Analyse mit der Beschäftigtenstatistik 1977-1987. In MittAB 1/91, 70-80.

Cochran, W. G. (1972): Stichprobenverfahren. Berlin: De Gruyter.

Cramer, U. (1985): Probleme der Genauigkeit der Beschäftigtenstatistik. In: Allg. Statist. Archiv, 69, 56-68.

Cramer, U. (1987): Klein- und Mittelbetriebe: Hoffnungstäger der Beschäftigungspolitik?. In MittAB 1/87, 15-29.

Cramer, U. (1992): Regionale Mobilität im Beschäftigungsverlauf, in: Akademie für Raumforschung und Landesplanung (Hrsg.): Regionale und biographische Mobilität im Lebensverlauf, Hannover, 69-89.

Cramer, U./ Majer, W./ John, K. (1989): Zwischenbericht zum IAB-Projekt 6-370 V „Bestandsaufnahme der Beschäftigtenstatistik. Internes Manuskript, Nürnberg.

Cramer, U./ Majer, W/John, K. (1990): Endbericht zum IAB-Projekt 6-370 V „Bestandsaufnahmen der Beschäftigtenstatistik, Internes Manuskript, Nürnberg.

Cramer, U./ Majer, W. (1991): Ist die Beschäftigtenstatistik revisionsbedürftig?. In: MittAB 1/91, 81-90.

Dietz, F. (1988): Strukturwandel auf dem Arbeitsmarkt. In: MittAB 1/88,115-152.

Duncan, G.T./ Lambert, D. (1986): Disclosure-Limited Data Dissemination (mit Diskussion). In: Journal of the American Statistical Association, Vol. 81, Nr. 393, 10-28.

Duncan, G.T./ Lambert, D. (1989): The Risk of Disclosure for Mikrodata. In: Journal of Business and Economic Statistic, Vol. 7, 207-217.

Fritsch, M./ König, A./ Weißhuhn, G. (1992): Probleme und Vorgehensweise bei der Bereinigung der in eine Betriebsdatei transformierten Beschäftigtenstatistik. Manuskript, Berlin.

Fritsch, M./ König, A./ Weißhuhn, G. (1994): Die Beschäftigtenstatistik als Betriebspanel – Ansatz, Probleme und Analysepotential. Beitrag für die Tagung „Firmenpanelstudien in Deutschland". Manuskript, Tübingen.

Görmar, W./ Irmen E. (1991): Nichtadministrative Gebietsgliederungen und -kategorien für die Regionalstatistik. Die siedlungsstrukturelle Gebietstypisierung der BfLR. In: BfLR-Mitteilungen, Nr. 6.

Herberger, L./ Becker, B. (1983): Sozialversicherungspflichtig Beschäftigte in der Beschäftigtenstatistik und im Mikrozensus. In: Wirtschaft und Statistik, Nr. 4, 290-304.

Hoffmann, H.P./Wermter, W. (1976): Die Beschäftigtenstatistik der Bundesanstalt für Arbeit – ihr Informationsgehalt, das Auswertungsprogramm und seine Realisierung. In: Arbeit und Beruf, 2, 33-36.

Infratest Sozialforschung (1994): Beschäftigungstrends Arbeitgeberbefragung 1993. Methodenband, München.

Karr, W./Rudolph, H./Cramer, U. (1986): Analytische Möglichkeiten der Beschäftigtenstatistik unter besonderer Berücksichtigung von Verlaufsuntersuchungen: Einführung und erste Ergebnisse zur Fluktuation. Internes Manuskript, Nürnberg.

Keller-McNulty, S./Unger E. (1993): Database Systems: Inferential Security. In: Journal of Official Statistics, Vol. 9, Nr.2, 475-517.

Laaksonen, S. (1995): Handling longitudinal micro data files of enterprise surveys. 2nd International Seminar on Statistical Confidentiality, Nov. 28-30, 1995, Luxembourg, Manuskript.

Lambert, D. (1993): Measures of Disclosure Risk and Harm. In: Journal of Official Statistics, Vol. 9, Nr.2, 313-331.

Little, R.J.A. (1993): Statistical Analysis of Masked Data. In: Journal of Official Statistics, Vol. 9, Nr.2, 407-426.

Marsh, C./ Skinner, C./ Arber, S./ Penhale, B./ Openshaw, S./ Hobcroft, J./ Lievesley, D./ Walford, N. (1991): The Case for Samples of Anonymized Records from the 1991 Census. In: Journal of the Royal Statistical Society, 154, 305-340.

Mayer, H.-L./ Becker, B. (1984): Sozialversicherungspflichtig Beschäftigte nach Beschäftigungsdauer, Bruttoarbeitsentgelt und Art der Beschäftigung. In: Wirtschaft und Statistik, Nr. 12, 994-1009.

Müller, W/ Blien, U./ Knoche, P./ Wirth, H. unter der Mitarbeit von Beckmann, P./ Bender, S./ Helmcke, T./ Müller, M. (1991): Die faktische Anonymität von Mikrodaten, Stuttgart: Metzler-Poeschel.

Müller, W/ Blien, U./ Wirth, H. (1995): Identification Risks of Microdata: Evidence from Experimental Studies. In: Sociological Methods & Research, im Erscheinen.

Pischner, R./ Rohwer, G. (1994): RZOO – Ein Retrievalprogramm für das Sozio-ökonomische Panel. Berlin, DIW.

Projektgruppe Betriebspanel (1994): Das IAB-Betriebspanel – Ergebnisse der ersten Welle 1993. In: MittAB 1/94, 20-32.

Rohwer, G. (1994): TDA Working Papers, Manuskripte, Bremen.

Rudolph, H. (1993): Darstellung der Bereinigungsläufe zur Historikdatei. Manuskript, Nürnberg.

Schimpl-Neimanns, B. (1995): Ergänzung zur Stellungnahme „Robustheit von Ereignisanalysen bei längsschnittanonymisierten Daten der Beschäftigtenstatistik des IAB" vom 17.10.94. Internes Manuskript, Mannheim.

Schmähl, W. (1985): Prozeßproduzierte Längsschnittinformationen zur Einkommensanalyse – Anmerkungen zu den Datenquellen. In: Allg. Statist. Archiv, 69, 275-285.

Schmähl, W./ Fachinger, U. (1994): Prozeßproduzierte Daten als Grundlage für sozial- und verteilungspolitische Analysen – Erfahrungen mit Daten der Rentenversicherungsträger für Längsschnittanalysen. In: Hauser, R./ Ott, N./ Wagner, G. (Hrsg.): Mikroanalytische Grundlagen der Gesellschaftspolitik, Band 2: Erhebungsverfahren, Analysemethoden und Mikrosimulation. Berlin: Akademie Verlag, 179-200.

Skinner, CJ./ Marsh, C./ Openshaw, S./ Wymer, C. (1994): Disclosure Control for Census Microdata. In: Journal of Official Statistics, Vol. 10, Nr. 1, 31-51.

Statistisches Bundesamt (1993): Bevölkerung und Erwerbstätigkeit, Fachserie l, Reihe 4.2. l, Struktur der Arbeitnehmer 30. Juni 1992. Wiesbaden: Metzler und Poeschel.

Statistisches Bundesamt (1960): Stichproben der amtlichen Statistik. Wiesbaden.

Veiling, J./ Bender, S. (1994): Berufliche Mobilität zur Anpassung struktureller Diskrepanzen am Arbeitsmarkt. In: MittAB 3/94, 212-231.

Wermter, W. (1981): Die Beschäftigtenstatistik der Bundesanstalt für Arbeit. In: MittAB 4/81, 428-435.

Wermter, W./ Cramer, U. (1988): Wie hoch war der Beschäftigtenanstieg seit 1983? – Ein Diskussionsbeitrag aus der Sicht der Beschäftigtenstatistik der Bundesanstalt für Arbeit. In: MittAB 4/88, 468-482.

Wiedenbeck, M./ Schimpl-Neimanns, B. (1994): Robustheit von Ereignisanalysen bei längsschnittanonymisierten Daten der Beschäftigtenstatistik des IAB. Internes Manuskript, Mannheim.