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

FDZ·Methodenreport

05/2012
DE

Methodische Aspekte zu Arbeitsmarktdaten

Kombinierte Firmendaten für Deutschland

Möglichkeiten und Konsequenzen der Zusammenführung von Unternehmensdaten unterschiedlicher Datenproduzenten

Elena Biewen,
Anja Gruhl,
Christopher Gürke,
Tanja Hethey-Maier,
Emanuel Weiß

Bundesagentur für Arbeit

Kombinierte Firmendaten für Deutschland

Möglichkeiten und Konsequenzen der Zusammenführung von Unternehmensdaten unterschiedlicher Datenproduzenten

Elena Biewen (Deutsche Bundesbank),

Anja Gruhl (Institut für Arbeitsmarkt- und Berufsforschung),

Christopher Gürke (Forschungsdatenzentrum des Statistischen Bundesamtes),

Tanja Hethey-Maier (Institut für Arbeitsmarkt- und Berufsforschung),

Emanuel Weiß (Forschungsdatenzentrum des Statistischen Bundesamtes)

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

FDZ-Methodenreporte (FDZ method reports) deal with the methodical aspects of FDZ data and thus help users in the analysis of data. In addition, through this series users can publicise their results in a manner which is citable thus presenting them for public discussion.

Inhaltsverzeichnis

Zusammenfassung

Im Rahmen des Projekts „Kombinierte Firmendaten für Deutschland“ (KombiFiD) wurden erstmals Unternehmensdaten verschiedener Institutionen zusammengeführt und der Wissenschaft zur Verfügung gestellt. Sowohl aufgrund der engen rechtlichen Grenzen, denen eine solche Zusammenführung unterliegt, als auch durch einen zum Teil nicht vorhanden eindeutigen Identifikator standen die an dem Projekt beteiligten Institutionen vor großen Herausforderungen. Der folgende Aufsatz gibt einen Überblick über die mit dem Projekt einhergehenden Ziele und den Verlauf des Projekts.

Abstract

In the project “Combined firm data for Germany” (KombiFiD) firm data from different institutions were merged and made available for research for the first time. The institutions involved in the project faced considerable challenges both due to the narrow legal limits underlying such a merging procedure and as a result of the partial lack of a unique identifier. This paper provides an overview of the objectives associated with the project and its progress.

Keywords: KombiFiD, firm level data, Germany

Das Projekt „Kombinierte Firmendaten für Deutschland“ (KombiFiD) wird durch die Förderung des Bundesministeriums für Bildung und Forschung ermöglicht.

1 Einleitung

Die amtliche Statistik konnte in den vergangenen Jahren ihr Mikrodatenangebot im Bereich der Unternehmens- und Betriebsdaten kontinuierlich ausweiten, um der wachsenden Nachfrage seitens der Wissenschaft nach Mikrodaten, die für immer komplexer werdende Analysen geeignet sind, gerecht zu werden.

Das seit Januar 2008 laufende Projekt „Kombinierte Firmendaten für Deutschland“ (KombiFiD) hat die Zusammenführung von Mikrodaten auf Unternehmensebene über die Grenzen von Datenproduzenten zum Gegenstand. An dem Projekt unmittelbar beteiligt sind das Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung, die Leuphana Universität Lüneburg sowie das Forschungsdatenzentrum des Statistischen Bundesamtes. Ferner bringt auch die Deutsche Bundesbank Teile ihres Datenbestandes in das Projekt ein. Im Projektverlauf wurde ein Datensatz erstellt, der von verschiedenen Datenproduzenten bereitgestellte Unternehmensdaten miteinander sowie über die Zeit verknüpft. Als Ergebnis liegt ein Panel über vier Jahre (von 2003 bis 2006) vor. Somit konnte die grundsätzliche juristische und technische Machbarkeit des Vorhabens gezeigt werden, einen entsprechenden institutionenübergreifenden Datensatz zu generieren.

Im Zuge dieses Aufsatzes wird zunächst auf die gegenwärtige Lage bezogen auf das der Wissenschaft zur Verfügung stehende Angebot an Unternehmensdaten und die Zugangswege zu diesen eingegangen und dargelegt, welche Ziele mit der Entwicklung des KombiFiD-Datensatzes verbunden sind. Da die Zusammenführung von Mikrodaten, die von verschiedenen Datenproduzenten erhoben werden, rechtlichen Restriktionen unterliegt, wird im dritten Kapitel die rechtliche Ausgangslage beschrieben sowie auf die sich aus dieser ergebenden Konsequenzen eingegangen. Es folgt die Auseinandersetzung mit den methodischen Aspekten der Zusammenführung der Mikrodaten der verschiedenen Institutionen. Der Aufsatz schließt mit einem Fazit.

2 Datenlage

Die Forschungsdatenzentren stellen für die Wissenschaft Mikrodaten sowohl auf der Betriebs- als auch der Unternehmensebene über verschiedene Zugangswege bereit (vgl. Zühlke et al. (2003)). So gibt es standardisierte Scientific Use Files, deren Mikrodaten faktische Anonymität aufweisen, was bedeutet, dass die Kosten einer potentiellen Aufdeckung einer einzelnen Beobachtungseinheit den mit dieser Deanonymisierung einhergehenden Nutzen überwiegen müssen (§ 16 Abs. 6 BStatG). Ferner haben Wissenschaftler die Möglichkeit, an Gastwissenschaftlerarbeitsplätzen vor Ort in den Forschungsdatenzentren ebenfalls mit Mikrodaten zu arbeiten. Des Weiteren gibt es die kontrollierte Datenfernverarbeitung, in deren Rahmen der Wissenschaftler nicht in direkten Kontakt mit den Daten kommt. Stattdessen schreibt dieser einen Code in einem Statistikprogramm, der anschließend von einem Mitarbeiter eines Forschungsdatenzentrums über nur formal anonymisierte Daten laufen gelassen wird. Bei diesen Daten werden keine Veränderungen vorgenommen, mit Ausnahme der Entfernung direkter Identifikatoren, wie etwa der Adresse des Firmensitzes. Die auf diesem Weg erstellten Ergebnisse werden in den Forschungsdatenzentren so anonymisiert, dass eine Reidentifikation einer einzelnen

Beobachtungseinheit nicht mehr möglich ist. Die Deutsche Bundesbank verfügt ebenfalls über Mikrodaten auf Unternehmensebene. Diese können von Wissenschaftlerinnen und Wissenschaftlern im Forschungszentrum der Bundesbank genutzt werden. Dabei ist die Nutzung nur auf Gastwissenschaftlerplätzen möglich.

Die von der amtlichen Statistik zur Verfügung gestellten Firmendaten weisen u. a. aufgrund der großen Stichproben[1] und der für Unternehmen bestehenden Auskunftspflicht, die sehr wenigen Antwortausfälle zur Folge hat, bereits jetzt ein hohes Informationspotential auf (vgl. Brandt et al. (2008)). Dennoch ist eine ständige Weiterentwicklung der für die Wissenschaft bereitgestellten Mikrodaten auf Betriebs- und Unternehmensebene notwendig, da sich die Analysemöglichkeiten auf Seiten der Wissenschaft stets ausweiten sowie der Bedarf der Politik für auf Mikrodaten basierende Beratung stetig zunimmt.

So wurden im Rahmen des Projektes „Amtliche Firmendaten für Deutschland“ (AFiD) mit Mikrodaten der Statistischen Ämter des Bundes und der Länder gezeigt, beispielsweise bezogen auf Unternehmen des Dienstleistungssektors und Betriebe des Verarbeitenden Gewerbes, dass Unternehmensdaten der amtlichen Statistik erfolgreich als Panel aufbereitet werden können, um entsprechende Längsschnittanalysen zu ermöglichen (vgl. Malchin und Voshage (2009)). Nach der erfolgten Verknüpfung von Betriebs- und Unternehmensdaten über die Zeit, die von einem Datenproduzenten erhoben wurden, stellte die Zusammenführung von Daten verschiedener Datenproduzenten den nächsten logischen Schritt dar.

Im Zuge des Projektes KombiFiD wurden Mikrodaten auf der Unternehmensebene zusammengeführt, die von den Statistischen Ämtern des Bundes und der Länder, dem Institut für Arbeitsmarkt- und Berufsforschung sowie der Deutschen Bundesbank stammen.

Die Statistischen Ämter stellen unterschiedliche Datenbestände aus fünf Bereichen (Verarbeitendes Gewerbe, Bau, Handel, Dienstleistungsbereich, Steuerstatistiken) zur Verfügung. Vom Institut für Arbeitsmarkt- und Berufsforschung geht das Betriebs-Historik-Panel (BHP) in das Projekt ein, das detaillierte Beschäftigteninformationen enthält. Die Daten der Bundesbank schließen Bilanzangaben von Unternehmen sowie Angaben zu Direktinvestitionen ein.

Eingang in den zu generierenden KombiFiD-Datensatz haben die in Tabelle 1 aufgeführten Statistiken der verschiedenen Datenproduzenten gefunden:

[1] Zahlreiche Unternehmensstatistiken stellen sogar Totalerhebungen mit Abschneidegrenzen dar, etwa die Investitionserhebungen bei Unternehmen und Betrieben des Verarbeitenden Gewerbes sowie der Gewinnung von Steinen und Erden.

Tabelle 1: Statistiken, die Eingang in den KombiFiD-Datensatz gefunden haben

Statistiken des FDZ der Statistischen Ämter	Statistiken des IAB und der Deutschen Bundesbank
▪ Unternehmensregister (URS) ▪ Kostenstrukturerhebung im Bereich Verarbeitendes Gewerbe, Bergbau und Gewinnung von Steinen und Erden (KSE-VG) ▪ Kostenstrukturerhebung im Baugewerbe (KSE-Bau) ▪ Strukturerhebung im Dienstleistungsbereich ▪ Jahreserhebung im Handel (sowie in der Instandhaltung und Reparatur von Kfz und Gebrauchsgütern) ▪ Investitionserhebung im Bauhaupt- und Ausbaugewerbe ▪ Investitionserhebung im Bereich Verarbeitendes Gewerbe, Bergbau und Gewinnung von Steinen und Erden ▪ Monatsbericht einschl. Auftragseingangserhebung für Betriebe im Bereich Verarbeitendes Gewerbe, Bergbau und Gewinnung von Steinen und Erden ▪ Umsatzsteuerstatistik ▪ Jahresbericht für Unternehmen im Bereich Verarbeitendes Gewerbe, Bergbau und Gewinnung von Steinen und Erden ▪ Verdienststrukturerhebung im Produzierenden Gewerbe und im Dienstleistungsbereich	**Statistiken des IAB:** ▪ Betriebs-Historik-Panel **Statistiken der Deutschen Bundesbank:** ▪ Bestandserhebung über Direktinvestitionen ▪ Jahresabschlussdaten der Deutschen Bundesbank

Quelle: Eigene Darstellung

Mit der Zusammenführung von Statistiken verschiedener Datenproduzenten verbindet sich wie angesprochen unter anderem das Ziel, neue Analysepotentiale entfalten zu können. So ist es beispielsweise denkbar, mit dem KombiFiD-Datensatz Korrelationen zwischen der Entwicklung von Direktinvestitionen und Beschäftigtenstruktur, oder zwischen der Kostenstruktur eines Unternehmens und der Beschäftigungsentwicklung zu untersuchen (vgl. L'Assainato (2009)). Die gegenwärtige Situation, im Rahmen von zahlreichen Befragungen teilweise gleiche Fragestellungen beantworten zu müssen, stellt eine Belastung sowohl für die befragten Einheiten als auch für die datenerhebenden Institutionen dar, auch wenn die Auskunftspflichten der Wirtschaft gegenüber der amtlichen Statistik nur einen vergleichsweise geringen Teil der gesamten bürokratischen Lasten ausmachen, die auf Unternehmen und Betrieben ruhen (vgl. Vorgrimler et al. (2011)). Ferner kann mit dem im Rahmen des Projektes KombiFiD entstandenen Datensatz evaluiert werden, an welchen Stellen Rationalisierungspotentiale hinsichtlich von Unternehmensbefragungen bestehen.

3 Rechtliche Ausgangssituation und aus dieser resultierende Konsequenzen

Nach der gegenwärtigen Gesetzeslage[2] können ausschließlich Daten, die einer einheitlichen gesetzlichen Grundlage unterliegen, zusammengeführt werden. Da die Daten unterschiedlicher Datenproduzenten auf der Basis von verschiedenen Rechtsgrundlagen erhoben werden, ist eine Zusammenführung von Daten über institutionelle Grenzen ohne vorheriges explizites Einverständnis der betroffenen Erhebungseinheiten nicht möglich. Daher mussten die Unternehmen, deren Daten im Rahmen von KombiFiD zusammengeführt werden sollten, schriftlich um ihr Einverständnis für eben diesen Schritt gebeten werden.

Um ein möglichst genaues Abbild aller Unternehmen zu erhalten, die sich in den für KombiFiD verwendeten Ausgangsstatistiken befinden, musste ein Stichprobenkonzept entworfen werden, welches Unternehmen aus den Wirtschaftsbereichen Produzierendes Gewerbe, Dienstleistung, Handel und Baugewerbe in die Stichprobe mit einbezieht. Aufgrund finanziell begrenzter Mittel konnten nur ca. 2% (ca. 55.000) aller für den Eingang in den zu generierenden KombiFiD-Datensatz potentiell in Frage kommender Unternehmen um Zustimmung gebeten werden. Unternehmen, die auf das Anschreiben nicht reagiert hatten, wurden bis zu dreimal angeschrieben. Aus dieser Vorgehensweise resultierte eine Rücklaufquote von etwa 57%, wobei bezogen auf alle angeschriebenen Unternehmen ca. 30% ihr Einverständnis erteilten. Trotz der deutlich über der ursprünglichen Erwartung liegenden Zustimmungsquote weisen die verknüpften Mikrodaten Verzerrungen auf, die auf die Selektivität zurückzuführen sind. Eine Quelle für Selektivitäten findet sich in einem heterogenen Antwortverhalten, wie es aus den Tabellen 2 und 3 sichtbar wird.

Tabelle 2: Antwortverhalten nach Wirtschaftsbereich und Region in %

Sitz des Unternehmens	Wirtschaftsbereich				Insgesamt
	Verarbeitendes Gewerbe	Bau	Handel	Dienstleistung	
West	39,9	32,1	24,5	28,3	30,6
Ost	39,1	26,4	20,8	27,8	28,2
Insgesamt	39,9	30,7	23,8	28,2	30,7

Quelle: Eigene Berechnung

[2] vgl. §13a Bundesstatistikgesetz (BStatG), Stand: 7. September 2007

Tabelle 3: Antwortverhalten nach Wirtschaftsbereich und Beschäftigtengrößenklasse in %

Beschäftigten-größenklasse	Wirtschaftsbereich				Insgesamt
	Verarbeitendes Gewerbe	Bau	Handel	Dienstleistung	
10/20 – 49	29,2	21,4	19,3	16,7	19,6
50 – 99	33,3	28,9	23,8	21,5	27,0
100 – 249	36,7	35,2	29,1	21,8	30,8
250 – 499	37,6	40,6	32,4	28,0	34,2
500 – 999	42,2	34,0	34,3	33,1	38,4
>= 1000	42,0	19,2	38,9	32,4	38,5
Insgesamt	34,2	25,4	22,1	18,7	30,7

Quelle: Eigene Berechnung

Um das zusätzliche Analysepotential, das ein Mikrodaten verschiedener Datenproduzenten enthaltener Datensatz eröffnen könnte, bedarf es der Möglichkeit, grundsätzlich von verschiedenen Institutionen gehaltene Unternehmensdaten zusammenführen zu können. Dies gilt ebenfalls für die Nutzung von Effizienzpotentialen, die in der Einsparung von den angesprochenen Mehrfachbefragungen bezogen auf die gleichen Merkmale liegen: Auch diese werden wohl nur ausgeschöpft werden können, wenn eine entsprechende Änderung der rechtlichen Rahmenbedingungen erfolgt.

Daher wurde im Zuge des Projekts ein Rechtsgutachten in Auftrag gegeben, mit dem Ziel der Klärung, ob und unter welchen Voraussetzungen die beschriebene Zusammenführung von Unternehmensdaten langfristig ohne gesonderte Zustimmung der betroffenen Unternehmen umgesetzt werden kann. Hervorzuheben ist, dass mit dem §13a des Bundesstatistikgesetzes (BStatG) bereits die Grundlage für eine mögliche Gesetzesänderung vorliegt. Im Sinne der oben genannten Ziele der langfristigen Unternehmensentlastung sowie der Verbesserung der Datennutzung und des Datenzugangs für die empirische Forschung ist eine Ausweitung des Geltungsbereichs des § 13a BStatG auf die entsprechenden Datenbestände der Bundesagentur für Arbeit und der Deutschen Bundesbank sowie die dauerhafte Zulässigkeit der institutionenübergreifenden Zusammenführung von Mikrodaten anzustreben. Gemäß der rechtlichen Begutachtung widerspricht eine Ausweitung des §13a BStatG hin zu einer dauerhaften Etablierung der institutionenübergreifenden Zusammenführung von wirtschaftsstatistischen Mikrodaten nicht grundsätzlich den verfassungsrechtlichen Vorgaben.

4 Verknüpfung der Daten der Statistischen Ämter des Bundes und der Länder, des Institutes für Arbeitsmarkt- und Berufsforschung (IAB) und der Deutschen Bundesbank

Das Projekt KombiFiD wurde in zwei Schritten ausgeführt. Im ersten Schritt wurden die Daten der Statistischen Ämter des Bundes und der Länder und das Betriebs-Historik-Panel des IAB zusammengeführt. Die Verknüpfung fand dabei über die BA-Betriebsnummern statt, die sowohl im Unternehmensregister (URS) des Statistischen Bundesamtes als auch im Betriebs-Historik-Panel des IAB enthalten sind. Im zweiten Schritt wurden die Daten der Bundesbank zugespielt. Da die Datensätze der Bundesbank mit den Daten der anderen Projektpartner keine gemeinsamen numerischen Identifikatoren aufweisen, erfolgte die Zusammenführung in diesem Fall über den Abgleich der Unternehmensnamen und -adressen. In den Abschnitten 4.1 und 4.2 wird das Vorgehen der Verknüpfung der Daten der Statistischen Ämter des Bundes und der Länder und des IAB für die KombiFiD Versionen 1.0 und 2.0 beschrieben[3] Abschnitt 4.3 beschreibt die weitere Verknüpfung mit den Daten der Bundesbank.

4.1 Verknüpfung der Daten für die KombiFiD Version 1.0

Das Ergebnis der nachfolgend dargestellten Arbeitsschritte ist eine Schlüsseldatei, die den Match der Datensätze vereinfacht.

Seitens des Institutes für Arbeitsmarkt- und Berufsforschung gehen die im Betriebs-Historik-Panel (BHP) zu Betriebsangaben zusammengefassten Individualdaten der sozialversicherungspflichtigen Beschäftigten in den KombiFiD-Datensatz ein.[4] Neben grundlegenden Informationen zu Beschäftigten-, Alters- und Lohnstrukturen beinhaltet der Datensatz ebenfalls Merkmale zu Ein- und Austritten von Beschäftigten.

Im Gegensatz zum BHP bildet der KombiFiD-Datensatz die Unternehmensebene ab. Dies stellt zwei wesentliche Anforderungen an den Verknüpfungsschlüssel für das Zusammenspielen der IAB-Daten mit den Datensätzen der Statistischen Ämter des Bundes und der Länder. Zum einen wird ein Identifikator benötigt, der die eindeutige Zuordnung der Einheiten der verschiedenen Datensätze zueinander ermöglicht. Zum anderen ist ein Aggregationsschlüssel erforderlich, der das Zusammenfassen der Betriebsdaten des BHP auf die Unternehmensebene erlaubt. Beide Voraussetzungen erfüllt das Unternehmensregister (URS). Es enthält sowohl die BA-Betriebsnummer, mit deren Hilfe die eindeutige Zuordnung der Einheiten des BHP zu denen der Statistischen Ämter möglich ist als auch eine Unternehmensnummer, die die Aggregation der Betriebseinheiten auf die Unternehmensebene gestattet.[5] Das URS wurde daher als Masterdatei für die Verknüpfung verwendet.

[3] Für die Version 2.0 wurde versucht, die Anzahl der Unternehmen, für die eine Verknüpfung zwischen den Daten der Statistischen Ämter des Bundes und der Länder und dem IAB hergestellt werden kann, zu erhöhen. Aus diesem Grunde wurde das Verfahren der Zusammenführung modifiziert.
[4] Zu näheren Informationen zum BHP vgl. Spengler (2008)
[5] Jeder Betriebsnummer (BNR) ist im Unternehmensregister eine Unternehmensnummer (UNR) zugeordnet. Die Angaben zu allen BNR, die unter einer spezifischen UNR verortet sind, werden

Der für die Verknüpfung verwendete URS-Auszug enthält die Betriebs- und Unternehmensnummern aller Unternehmen, die einer Datenverknüpfung zustimmten. Die Daten wurden in einzelnen Querschnitten für die Jahre 2003 bis 2008 geliefert.[6] Vor dem Match wurden die Unternehmensregisterdaten für die Anforderungen der Verknüpfung aufbereitet. Einheiten ohne im URS verzeichnete BA-Betriebsnummer wurden gelöscht, da in diesen Fällen eine Verknüpfung mit dem BHP nicht möglich war. Unternehmen, die nur unvollständig, d.h. nicht mit allen im URS verzeichneten Betriebsnummern, im BHP identifizierbar waren, wurden ebenfalls vor der Verknüpfung gelöscht. Somit beschränkte sich die Verknüpfung auf Unternehmen, die „vollständig", also mit allen Betriebseinheiten im BHP gefunden wurden. Gleichfalls vor dem Match gelöscht wurden Betriebe, die durch mehrfach auftretende identische BA-Betriebsnummern und Unternehmensnummern gekennzeichnet waren. Diese Betriebe sind im BHP nicht zu identifizieren, da sie unter einer Betriebsnummer subsumiert werden. Im Gegensatz dazu werden die Einheiten solcher Spezialfälle im URS getrennt verzeichnet. Das führt dazu, dass im URS identische Betriebsnummern mehrfach auftreten können.[7]

Somit enthält die aus der Verknüpfung resultierende Schlüsseldatei, die für die Verknüpfungen des BHP mit den Daten der Statistischen Ämter des Bundes und der Länder genutzt wird, nur Identifikatoren zu Unternehmen mit vollständigem Betriebsnummernkanon.

Nach der Aufbereitung der Daten erfolgte der Match von BHP und URS-Auszug über die BA-Betriebsnummer als eindeutigen numerischen Identifikator. Anschließend wurden die Betriebsdaten anhand der Betriebsnummern und der zugehörigen Unternehmensnummern auf Unternehmensebene aggregiert.

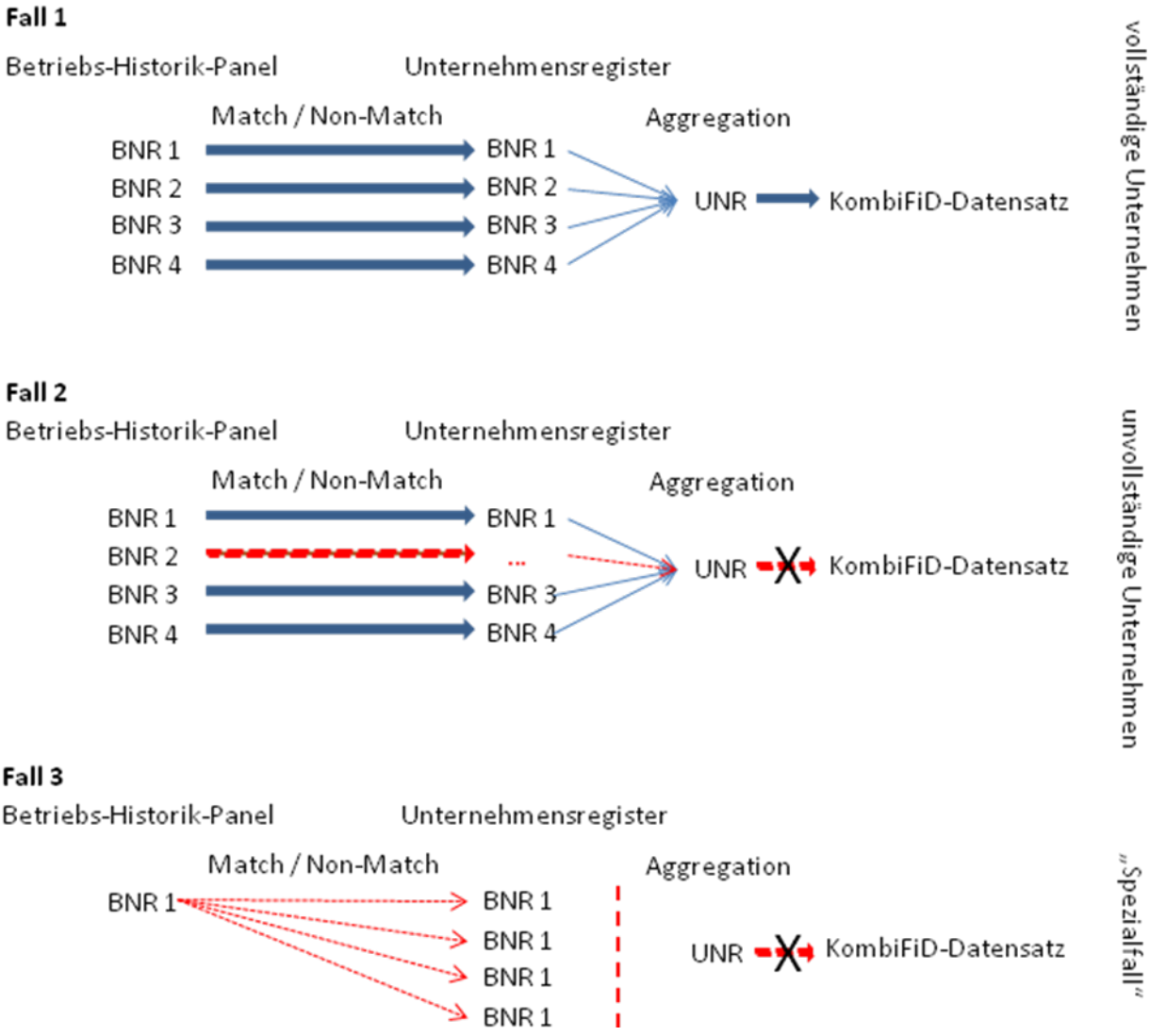
In Abbildung 1 sind diese Bereinigungen, die Verknüpfung und die Aggregation stark vereinfacht dargestellt.

zusammengefasst und in dieser Form in den KombiFiD-Datensatz übernommen (z.B. Summe alle Beschäftigte).

[6] Bei der Verknüpfung mit dem BHP war ein „time lag" von zwei Jahren zu beachten, der sich dadurch begründet, dass qualitativ gesicherte Angaben aus administrativen Quellen zum zwei Jahre zurückliegenden Berichtsjahr vorliegen (Statistisches Bundesamt, 2009).

[7] Für ausführliche Informationen vgl. FDZ-Methodenreport 01/2010, http://fdz.iab.de/187/section.aspx/Publikation/k100311r01.

Abbildung 1: Schematische Darstellung des Verknüpfungsprozesses von URS und BHP, KombiFiD Version 1.0



Quelle: eigene Darstellung

Die Verknüpfung von URS und BHP führte zu sehr guten Ergebnissen. Die Anteile der vollständig im BHP identifizierbaren URS-Unternehmen bewegen sich pro betrachtetem Querschnitt mit deutlich über 90% in einem Bereich, der für die Analysemöglichkeiten des KombiFiD-Datensatzes sehr positiv zu werten ist. Annähernd 80% der Unternehmensnummern ließen sich im gesamten Untersuchungszeitraum beobachten. Lediglich drei Prozent der identifizierten Unternehmensnummern kamen nur in einem Querschnittsjahr vor. Der überwiegende Anteil der Unternehmen des ersten Beobachtungsjahres existiert inklusive aller Betriebseinheiten über den gesamten hier

relevanten Zeitraum. Im letzten Querschnittsjahr waren noch 87,8% der Unternehmen des ersten Untersuchungsjahres vollständig in den Daten abgebildet.

Tabelle 4: Anzahl der im BHP identifizierten vollständigen Unternehmen und Einbetriebsunternehmen

	Im BHP identifizierte *vollständige* Unternehmen		Einbetriebsunternehmen	
Jahr, bezogen auf URS-Auszug	absolut	Anteil an allen *vollständigen* URS-Unternehmen in %	absolut	Anteil an allen URS-Unternehmen in %
2003	13.296	96,2	11.994	73,9
2004	13.600	96,1	12.158	71,4
2005	13.722	95,7	12.256	71,7
2006	13.653	95,0	12.173	71,7

Quelle: KombiFiD-Daten; eigene Berechnungen

4.2 Verknüpfung KombiFiD Version 2.0

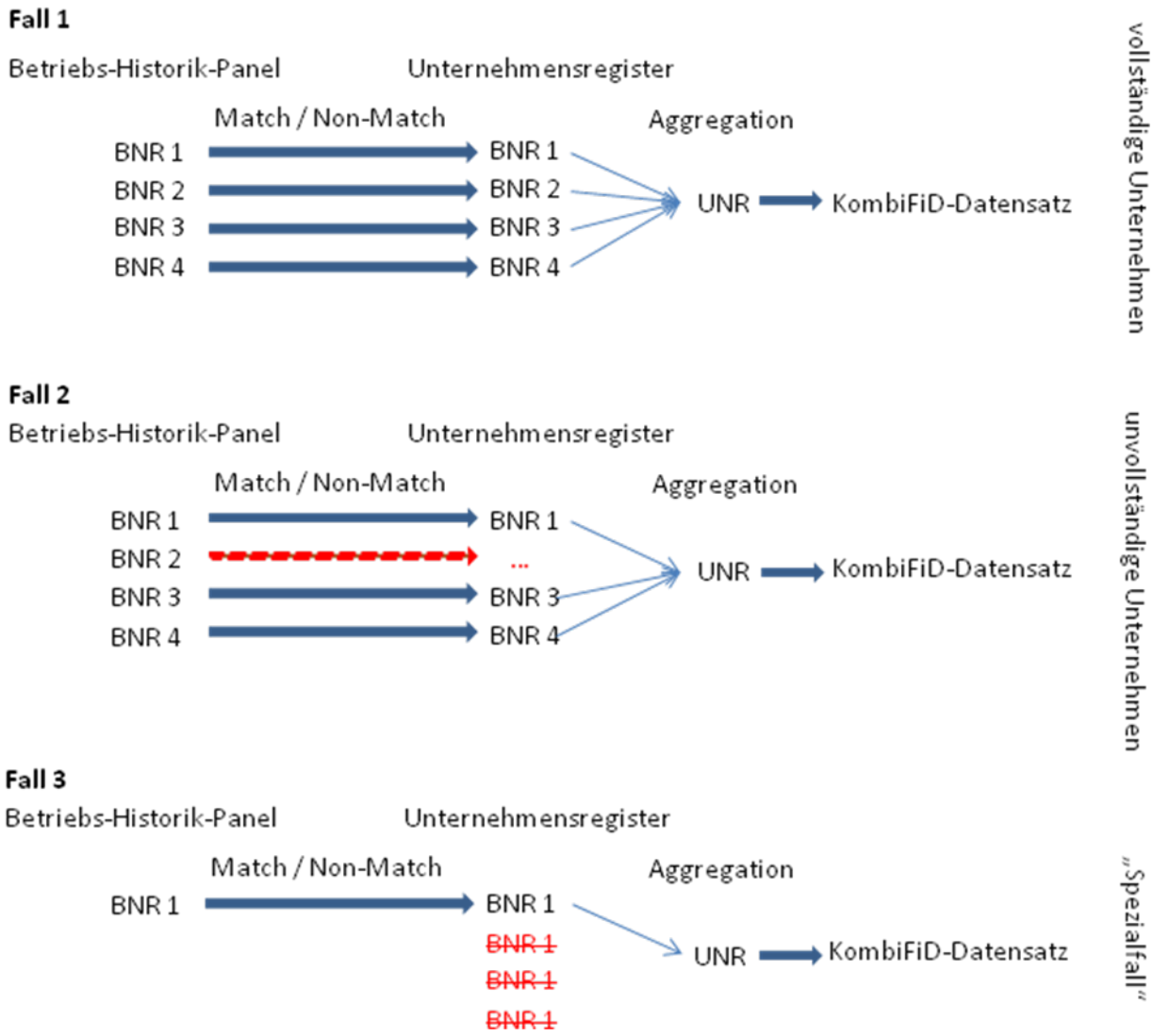
Im Gegensatz zur Verknüpfung bei der KombiFiD Version 1.0 wurden bei der Erstellung der KombiFiD-Version 2.0 auch unvollständige Unternehmen und die genannten Spezialfälle berücksichtigt. Es besteht jedoch weiterhin die Möglichkeit die Gruppe der sogenannten vollständigen Unternehmen zu identifizieren. Dies geschieht über generierte Zusatzvariablen im BHP Variablenspektrum.

Wie schon bei Version 1.0 wurden Einheiten ohne im URS verzeichnete BA-Betriebsnummer (missings) gelöscht, da in diesen Fällen eine Verknüpfung mit dem BHP nicht möglich war. Bei Unternehmen, bei denen nicht alle zugeordneten Betriebsnummern im BHP identifiziert werden konnten, wurden nach der Verknüpfung die nicht zuordenbaren Betriebsnummern gelöscht. Die verknüpften Betriebsnummern dieser Unternehmen wurden im Anschluss aggregiert. Diese Vorgehensweise unterscheidet sich von jener bei der Erstellung der Version 1.0 des KombifiD-Datensatzes, wo solche unvollständigen Unternehmen komplett aus dem Datensatz gelöscht wurden. In Fällen von Mehrfachnennungen einer Betriebsnummer im URS wurden alle Nennungen der Betriebsnummer bis auf eine gelöscht. Diese verbleibende Betriebsnummer wurde mit der entsprechenden Einheit im Betriebs-Historik-Panel verknüpft. Auch hier ergibt sich eine Abweichung zu Version 1.0 in der diese Spezialfälle generell gelöscht und nicht bei der Verknüpfung berücksichtigt wurden.

Nach der Aufbereitung der Daten erfolgte der Match von BHP und URS-Auszug über die BA-Betriebsnummer als eindeutigen numerischen Identifikator. Anschließend wurden die Betriebsdaten anhand der Betriebsnummern und der zugehörigen Unternehmensnummern auf Unternehmensebene aggregiert.

In Abbildung 2 sind diese Bereinigungen, die Verknüpfung und die Aggregation stark vereinfacht dargestellt.

Abbildung 2: Schematische Darstellung des Verknüpfungsprozesses von URS und BHP



Quelle: eigene Darstellung

Pro Querschnittjahr konnten zwischen 85% und 88% der Unternehmen des URS vollständig im BHP identifiziert werden. Zwischen 75% und 79% aller Unternehmen sind sogenannte Einbetriebsunternehmen. Tabelle 5 zeigt die Ergebnisse für die einzelnen Querschnitte der KombiFiD-Daten.

Tabelle 5: Anzahl der im BHP identifizierten vollständigen Unternehmen und Einbetriebsunternehmen

		Im BHP identifizierte *vollständige* Unternehmen		Einbetriebsunternehmen	
Jahr	Anzahl Unternehmen, gesamt	absolut	Anteil an allen URS-Unternehmen in %	absolut	Anteil an allen URS-Unternehmen in %
2003	15.812	13.722	86,8	12.256	77,5
2004	15.888	13.653	85,9	12.173	76,6
2005	15.924	13.580	85,3	12.077	75,8
2006	15.942	13.549	85,0	12.019	75,3

Quelle: KombiFiD-Daten, eigene Berechnungen

4.3 Verknüpfung mit den Daten der Deutschen Bundesbank

Nach der oben beschriebenen Verknüpfung der Daten der Statistischen Ämter mit den Daten des IAB, wurden in einem weiteren Schritt die Datensätze der Deutschen Bundesbank die Mikrodatenbank Direktinvestitionen (MiDi) und die Unternehmensbilanzen (USTAN) hinzugefügt. Die MiDi enthält Informationen über deutsche Direktinvestitionen im Ausland (Outward-FDI) und ausländischen Direktinvestitionen in Deutschland (Inward-FDI), wenn bestimmte Meldegrenzen überschritten werden (vgl. Lipponer 2003, 2009). Die USTAN schließt Jahresabschlüsse nicht-finanzieller Unternehmen ein, die die Bundesbank im Rahmen des Refinanzierungsgeschäfts bekommt (vgl. Stöss 2001). Bei dieser Verknüpfung lag im Gegensatz zur ersten kein gemeinsamer fehlerfreier Schlüssel auf Seiten der Deutschen Bundesbank und der Statistischen Ämter bzw. dem IAB vor. Somit wurde in diesem Falle auf Techniken des Record Linkage zurückgegriffen, die eine Verknüpfung bei fehlerbehafteten Schlüsseln erlauben[8]. Als Schlüssel dienten in diesem Falle Namens- und Adressangaben in den Datenbeständen der Deutschen Bundesbank und des IAB. Technisch umgesetzt wurde der Abgleich mit der an der Universität Duisburg entwickelten RecordLinkage Software MTB (Merge-Toolbox).[9]

Auf Seiten der Deutschen Bundesbank wurde eine Adressdatei mit Namen und Adressen aller Unternehmen aus der MiDi und USTAN zum Stand 2006 bereitgestellt. Aus Datenschutzgründen wurde diese Datei vor der Übermittlung an das IAB mit weiteren Unternehmensadressen aus anderen bei der Deutschen Bundesbank vorliegenden Unternehmensdatenbanken (z.B.: DAFNE, Hoppenstedt) aufgefüllt. Die Datei umfasste damit insgesamt 76.051 Einträge. Auf Seiten des IAB wurde die sogenannte IAB Betriebedatei zum Stand 2006 aufbereitet. Hierbei handelte es sich um eine Adressdatei aller zum Zeitpunkt 2006 aktiven Betriebe in Deutschland mit mindestens einem sozialversicherungspflichtig bzw. geringfügig Beschäftigten. Die Datei beinhaltete 2.734.332 Einträge. Ähnlich wie bei der weiter oben beschriebenen Verknüpfung zwischen den Daten der Statistischen Ämter und des IAB, bestand auch hier das Problem, dass sich der IAB

[8] Ein Einblick in das Themengebiet Record Linkage gibt Winkler 1995.
[9] Eine ausführliche Beschreibung der Software findet sich in Schnell et al. 2005.

Datensatz nicht auf die Unternehmensebene sondern auf die feingliedrigere Betriebsebene bezog. Aus Vorgängeruntersuchungen war jedoch bekannt, dass bei Unternehmen mit sozialversicherungspflichtig Beschäftigten die Unternehmensadresse in der Regel auch eine Betriebsadresse darstellt. Somit wurde bei der Verknüpfung versucht, den Unternehmen aus dem Datenbestand der Deutschen Bundesbank jeweils eine Betriebsadresse aus der IAB Betriebedatei und damit eine Betriebsnummer zuzuordnen. Über den bereits oben beschriebenen URS-Auszug, welcher für jedes zustimmende Unternehmen alle Betriebsnummern beinhaltet, wurde anschließend ermittelt, ob der Betrieb und damit das dahinterstehende Unternehmen Teil der KombiFiD Stichprobe ist.

Vor der eigentlichen Verknüpfung wurden die Datenbestände im Preprocessing weitgehend standardisiert, um unterschiedliche Schreibweisen derselben Namens- und Adressinformation zu minimieren. Bestandteil dieser Aufbereitung waren unter anderem eine Umwandlung aller Buchstaben in Großbuchstaben, die Umkodierung von Umlauten (Ä = AE, …), die Löschung von Sonderzeichen (z.B.: >!/;), die Standardisierung von Adressbestandteilen (z.B.: Straße, Weg,), die Löschung von Leerzeichen sowie die Bereinigung von fehlerhaften Postleitzahlen. Damit lagen am Ende auf beiden Seiten die Namens- und Adressinformationen in folgenden Variablen vor: Unternehmensname, Rechtsform, Straße, Hausnummer, Ortsname, Postleitzahl.

Vor der Anwendung von fehlertoleranten Verknüpfungsverfahren wurden die beiden Datenbestände einem exakten Abgleich unterzogen. Hierbei musste eine Beobachtung aus dem Datenbestand der Deutschen Bundesbank auf ausgewählten Variablenkombinationen zu 100 Prozent mit einer Beobachtung aus dem IAB Datenbestand übereinstimmen, um als gültige Verknüpfung eingestuft zu werden. Um die Rechenzeit dieser Abgleiche zu reduzieren, wurde eine sogenanntes „Blocking“ angewendet. Hierbei werden nicht alle Beobachtungen der beiden Datensätze miteinander verglichen, sondern nur jene, die identische Einträge auf der Blockingvariable besitzen. Als Blockingvariable wurde die zwei- bzw. dreistellige Postleitzahl verwendet. Tabelle 6 listet die drei verschiedenen Variablenkombinationen, die beim exakten Abgleich in gegebener Reihenfolge verwendet wurden, sowie die Anzahl gültiger Verknüpfung die daraus resultierten.

Tabelle 6: Ergebnisse des exakten Namens- und Adressabgleichs

Modell	Blocking	Variablen	Anzahl gültige Verknüpfungen
1	3-Steller PLZ	Unternehmensname, Rechtsform, Ort, Straße, Hausnummer	22.080
2	3-Steller PLZ	Unternehmensname, Rechtsform, 5-Steller PLZ, Straße, Hausnummer	631
3	2-Steller PLZ	Unternehmensname, Rechtsform, Ort, Straße, Hausnummer	20

Nach Abschluss der ersten drei exakten Abgleiche konnten somit 22.731 Unternehmen (ca. 30 Prozent) im Bundesbank Datensatz einer Betriebsnummer zugeordnet werden.

Anschließend wurde der exakte Abgleich durch den Einsatz einer fehlertoleranten Ähnlichkeitsfunktion ersetzt. In diesem Falle wurde auf die sogenannten Bi-gramme[10] beim Abgleich der Unternehmensnamen zurückgegriffen. Auf den Einsatz von weiteren Ähnlichkeitsfunktionen bei anderen Verknüpfungsvariablen wurde verzichtet und restriktiv auf der Hausnummernebene bzw. Straßenebene geblockt. Damit war es möglich Unternehmen auch in Industrieparks und großen Bürokomplexen, in denen alle Unternehmen dieselbe Straßenanschrift haben und sich lediglich in ihrem Unternehmensnamen voneinander unterscheiden, richtig zu verknüpfen. Tabelle 7 zeigt die Ergebnisse der in gegebener Reihenfolge durchgeführten fehlertoleranten Abgleiche.

Tabelle 7: Ergebnisse des fehlertoleranten Namensabgleichs

Modell	Blocking	Variablen	Ähnlichkeitsfunktion	Anzahl gültige Verknüpfungen
4	5-Steller PLZ, Straße, Hnr.	Unternehmensname	Bi-gramme	15.435
5	2-Steller PLZ, Ort, Straße, Hnr.	Unternehmensname	Bi-gramme	114
6	5-Steller PLZ, Straße	Unternehmensname	Bi-gramme	2.976

Nach Abschluss des fehlertoleranten Abgleichs konnten somit insgesamt 41.256 Unternehmen (ca. 55 Prozent) im Bundesbank Datensatz einer Betriebsnummer zugeordnet werden. Die Suche dieser Betriebsnummern im URS-Schlüssel ergab, dass von diesen, 4.170 Unternehmen Teil der KombiFiD Stichprobe sind.

Weiter wurden die nach dem Adressabgleich identifizierten KombiFiD-Unternehmen mit den Datensätzen der Deutschen Bundesbank Mikrodatenbank Direktinvestitionen (MiDi) und Unternehmensbilanzen (USTAN) verknüpft. Vor der Verknüpfung wurden zuerst Dubletten bereinigt. Unter Dubletten sind Fälle zu verstehen, wenn eine URS-Unternehmensnummer mehreren Unternehmen auf der Seite der Bundesbank zugeordnet wurde. Dies trat besonders häufig dann auf, wenn neben einer Firma eine gleichnamige Holding-Gesellschaft existiert. Da Holdings in den Daten der Bundesbank eine eigene Unternehmensnummer haben, oft aber keine eigenen sozialversicherungspflichtige Beschäftigte aufweisen, wurden sie somit in der Adressdatei des IAB mit der gleichnamigen Firma verknüpft, unter der die sozialversicherungspflichtig Beschäftigten registriert waren. Nach manueller Überprüfung solcher Fälle reduzierte sich die Anzahl der Unternehmen der KombiFiD-Stichprobe von 4.170 auf 3.788.

Tabelle 8 zeigt die Anzahl von Unternehmen der KombiFiD-Stichprobe, über die es Informationen in der Mikrodatenbank Direktinvestitionen (MiDi) gibt. Die Firmenzahlen variieren je nach Jahr zwischen 702 (in 2003) und 780 (in 2005), wobei 681 Unternehmen in allen vier Jahren 2003-2006 vorkommen (Zeile Panel in der Tabelle). Folglich zeigt Tabelle 8 Anzahlen von MiDi-Firmen, die mit dem Betriebs-Historik-Panel des IAB und ausgewählten

[10] Hierbei wird der Unternehmensname in Zeichenketten der Länge 2 (Bi-gramme) aufgeteilt. Die Anzahl der übereinstimmenden Bi-gramme auf beiden Seiten entscheidet über die Ähnlichkeit der Beobachtungen.

Statistiken der Statistischen Ämter des Bundes und der Länder verknüpft werden konnten. Ca. 97% der Firmen sind im BHP, ca. 76% in der Umsatzsteuerstatistik aufzufinden. Des Weiteren wird am besten der Bereich des Verarbeitenden Gewerbes abgedeckt. Ca. 63% von Unternehmen in der MiDi findet sich im Monatsbericht, ca. 60% in der Kostenstrukturerhebung, im Jahresbericht für Mehrbetriebsunternehmen und in der Investitionserhebung. Die Verknüpfung mit den Statistiken aus anderen Wirtschaftsbereichen (Handel, Dienstleistungen, Bau) fiel allerdings viel geringer aus.

Tabelle 8: Verknüpfte MiDi-Unternehmen

Jahr	MiDi	Verknüpfung MiDi mit								
		BHP	MB VG	KSE VG	JB VG	IE VG	JE H	USS	SE D	VSE
2003	702	675	416	410	402	404	115	546	90	
2004	732	716	447	430	437	437	117	565	96	
2005	780	772	475	461	463	466	129	595	106	
2006	771	769	472	459	455	461	130	578	100	268
Panel	681	651	388	377	373	493	110	496	84	

BHP: Betriebs-Historik-Panel; MB VG: Monatsbericht im Verarbeitenden Gewerbe; KSE VG: Kostenstrukturerhebung im Verarbeitenden Gewerbe; JB VG: Jahresbericht für Mehrbetriebsunternehmen im Verarbeitenden Gewerbe; IE VG: Investitionserhebung im Verarbeitenden Gewerbe; JE H: Jahreserhebung im Handel; USS: Umsatzsteuerstatistikpanel; SE D: Strukturerhebung im Dienstleistungsbereich; VSE: Verdienststrukturerhebung im Produzierenden Gewerbe und im Dienstleistungsbereich (Querschnitt 2006). Fallzahlen für weitere Statistiken fallen geringer aus und werden aus Platzgründen hier nicht berichtet.

Quelle: Eigene Berechnungen

Anhand von verknüpften Unternehmenszahlen kann noch keine Aussage über die Qualität des Datensatzes gemacht werden. Aus diesem Grund werden im nächsten Arbeitsschritt Replikationsstudien und weitere Qualitätsanalysen durchgeführt und somit überprüft, ob Analysen auf Basis von Originaldaten denjenigen auf Basis von KombiFiD-Stichprobe gleichen.

5 Fazit der Machbarkeitsstudie „KombiFiD“

Die Machbarkeitsstudie KombiFiD hat den Grundstein für neue Überlegungen hinsichtlich von Entlastungspotentialen für Auskunftspflichtige und neue Analysemöglichkeiten für die empirisch arbeitende Wissenschaft gelegt. Insbesondere die Forderung nach einer Novellierung des §13a BStatG zur Schaffung einer dauerhaften Gesetzesgrundlage zur Verknüpfung von wirtschaftsstatistischen Daten über die Grenzen von Datenproduzenten hinweg, sollte künftig im Fokus aller Beteiligten, den Datenproduzenten, Verbänden und Interessengruppen, Politik und Wissenschaft, stehen. Eine entsprechende Novellierung würde der Erfüllung zweier Ziele entsprechen: So würde das Wegfallen der Notwendigkeit das Einverständnis bei den Erhebungseinheiten für die Datenzusammenführung einholen zu müssen mit der Beseitigung derjenigen potentiellen Selektionseffekte einhergehen, die ihren Ursprung in dem zuvor beschriebenen heterogenen Antwortverhalten haben. Des Weiteren ist das Anschreiben aller Unternehmen, für eine Zusammenführung ihrer bei verschiedenen Datenproduzenten liegenden Mikrodaten in Frage kommen, mit einem sehr hohen Ressourcenaufwand verbunden. Somit würde die angesprochenen Gesetzesänderungen mit großen Effizienzgewinnen bei möglichen künftigen Projekten zur Erstellung von Datensätzen einhergehen, die sich Datenbeständen mehrere Datenproduzenten bedienen.

Unabhängig von der Schaffung einer Gesetzesgrundlage ist ebenfalls die Harmonisierung von Begriffsdefinitionen und Methodiken zwischen den nationalen Datenproduzenten ein wichtiger Meilenstein für die Schaffung einer qualitativ hochwertigen und Datenproduzenten übergreifenden vergleichbaren Mikrodatengrundlage für die empirisch arbeitende Wissenschaft.

Es folgen weitere Qualitätsanalysen bezogen auf die den KombiFiD-Datensatz betreffenden Selektionseffekte, was Lerneffekte für künftige Erhebung ohne Auskunftspflicht mit sich bringen wird.

Der bereits vorliegende KombiFiD-Datensatz steht der Wissenschaft in der Version 2.0 zur Nutzung bis Ende 2021 zur Verfügung, da aufgrund der rechtlichen Restriktionen eine längere Nutzungsdauer nicht ermöglicht werden konnte. Die entsprechenden Nutzungsanträge und die dem Zugang zu den Daten zugrundeliegenden Regeln sowie die Metadaten sind unter www.kombifid.de zu finden.

Literatur

Brandt, Maurice; Oberschachtsiek, Dirk; Pohl, Ramona (2008): Neue Datenangebote in den Forschungsdatenzentren – Betriebs- und Unternehmensdaten im Längsschnitt, in: AStA Wirtschafts- und Sozialstaatliches Archiv, 2008/2, S. 193-207.

L'Assainato, Sandro (2009): KombiFiD – Kombinierte Firmanetdaten für Deutschland: Institutionen-übergreifende Zusammenführung von Unternehmensdaten, in: DRV-Schriften, Band 55, 2009, S. 39-54.

Lipponer, Alexander (2003): Deutsche Bundesbank's FDI Micro Database. European Data Watch-Articles Schmollers Jahrbuch / Journal of Applied Social Science Studies 123, 593-600.

Lipponer, Alexander (2009): Microdatabase Direct Investment – MiDi a brief guide. Deutsche Bundesbank, http://www.bundesbank.de/vfz/vfz_forschungsdaten_einzeldaten.php

Malchin, Anja und Ramona, Voshage (2009): Official Firm Data for Germany, Schmollers Jahrbuch / Journal of Applied Social Science Studies 129, S. 501-513, Berlin.

Schnell, Rainer; Bachteler, Tobias; Reiher, Jörg (2005): MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung. in: ZA-Information, 56 S. 93-103.

Spengler, Anja (2008): The Establishment History Panel. European Data Watch Articles Schmollers Jahrbuch / Journal of Applied Social Science Studies 128, 501-509.

Spengler, Anja (2010): Verknüpfung und Abgleiche von Unternehmensregisterdaten des Statistischen Bundesamtes mit Betriebsdaten des Instituts für Arbeitsmarkt- und Berufsforschung. FDZ-Methodenreport 1/2010, Nürnberg.

Stöss, Elmar (2001): Deutsche Bundesbank's Corporate Balance Sheet Statistics and Areas of Application. European Data Watch-Articles Schmollers Jahrbuch / Journal of Applied Social Science Studies 121, 131-137.

Vorgrimler, Daniel; Spengler, Florian; Schüßler, Simone (2011): Konzeption und erste Ergebnisse des Belastungsbarometers für Wirtschaftsstatistiken, in: Wirtschaft und Statistik 06/2011, S. 528-535.

Winkler, William E. (1995): Matching and Record Linkage, in: Cox, Brenda G. et al. (Hrsg.) Business Survey Methods. New York, S. 355-384.

Zühlke, Sylvia; Zwick, Markus; Scharnhorst, Sebastian; Wende, Thomas (2003): Die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, in: Wirtschaft und Statistik 10/2003, S. 906-911.

Impressum

FDZ-Methodenreport 05/2012

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

Redaktion
Stefan Bender, Iris Dieterich

Technische Herstellung
Iris Dieterich

Bezugsmöglichkeit
http://doku.iab.de/fdz/reporte/2012/MR_05-12.pdf

Internet
http://fdz.iab.de/

Rückfragen zum Inhalt an:

Anja Gruhl
Forschungsdatenzentrum (FDZ)
Regensburger Str. 104
90478 Nürnberg
Tel.: 0911 / 179-5669
E-Mail: anja.gruhl@iab.de