

Forschungsdatenzentrum

der Bundesagentur für Arbeit
im Institut für Arbeitsmarkt-
und Berufsforschung

FDZ

FDZ-Methodenreport

05/2014

DE

Methodische Aspekte zu Arbeitsmarktdaten

ReLOC-Linkage: Eine neue Methode zur Verknüpfung von Unternehmensdaten mit den Betriebsdaten des IAB

Johannes Schäffler



Bundesagentur für Arbeit

ReLOC-Linkage: Eine neue Methode zur Verknüpfung von Unternehmensdaten mit den Betriebsdaten des IAB

Johannes Schäffler (Institut für Arbeitsmarkt- und Berufsforschung)

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

Tabellenverzeichnis	4
Zusammenfassung	5
Abstract	5
1 Einleitung	6
2 Ausgangslage	6
3 Datenbanken	8
3.1 ReLOC-Datenbank	8
3.2 Betriebsdaten der BA	9
4 Verknüpfungsmöglichkeiten	11
4.1 Einzigartigkeit von Unternehmensnamen	11
4.1.1 Unternehmen ohne Kaufmannseigenschaft	11
4.1.2 Unternehmen mit Kaufmannseigenschaft	12
4.2 Überprüfung der zu verknüpfenden Datenbanken	14
4.2.1 Einzigartigkeit der Unternehmensnamen	14
4.2.2 Fehler in den Betriebs-/Unternehmensnamen	14
5 Datenverknüpfung	16
5.1 Verwendung der Namen und Adressen	19
5.2 Nur Verwendung der Namen	21
5.3 Ergebnis aus beiden Verknüpfungsschritten	22
6 Zusammenfassung und Ausblick	25
Literatur	26

Tabellenverzeichnis

Tabelle 1: Rechtsformen der ReLOC-Unternehmen	9
Tabelle 2: Fiktive Beispiele für die bei der BA erfassten Betriebsnamen	10
Tabelle 3: Fehler in den Betriebs-/Unternehmensnamen der BA	15
Tabelle 4: Fiktive Preprocessing-Beispiele	18
Tabelle 5: Ergebnis des exakten Namens- und Adressabgleichs	20
Tabelle 6: Ergebnis des fehlertoleranten Namens- und Adressabgleichs	21
Tabelle 7: Namenseinzigartigkeit potentieller Verknüpfungen	22
Tabelle 8: Quelle der gültigen Verknüpfungen	23
Tabelle 9: Anzahl der Betriebsnummern pro Unternehmen	24

Zusammenfassung

Dieser Artikel beschreibt eine neue Methode zur Verknüpfung von Unternehmensdaten mit den Betriebsdaten des Instituts für Arbeitsmarkt- und Berufsforschung (IAB). Diese wurde im Rahmen des Projekts „Research on Locational and Organisational Change“ (ReLOC) entwickelt und erstmalig angewendet. Hierbei wird der Umstand genutzt, dass bei der Vergabe der Betriebsnummern durch die Bundesagentur für Arbeit (BA) für einen Betrieb der zugehörige Unternehmensname erfasst wird. Dies ermöglicht die Zusammenführung von Unternehmens- mit Betriebsdaten allein auf Basis des Unternehmensnamens. Der Erfolg dieses Vorgehens hängt aber von der Korrektheit und Einzigartigkeit der Unternehmensnamen ab. Deshalb wird vor der Datenverknüpfung untersucht, inwieweit diese Voraussetzungen erfüllt sind und worauf dabei zu achten ist. Da diese Methode für viele Projekte eine innovative Erweiterung bei der Bearbeitung von Forschungsfragen darstellen kann und mittlerweile auch vom Forschungsdatenzentrum (FDZ) der BA und dem German Record Linkage Center (GRLC) für erste Projekte übernommen wurde, erläutert dieser Artikel die Verknüpfung der ReLOC-Datenbank und die dafür zugrundeliegenden Informationen im Detail.

Abstract

This article describes a new method for the linkage of firm-level data with establishment-level data of the Institute for Employment Research (IAB). It has been developed and for the first time applied in the project „Research on Locational and Organisational Change“ (ReLOC). The method makes use of the fact that in the course of assigning IDs to establishments the Federal Employment Agency (BA) records the associated firm name. This enables the linkage of firm-level data with information available at the establishment level solely on the basis of the firm name. However, the success of this approach depends on the correctness and uniqueness of firm names. Before conducting the record linkage, it was investigated whether and to what extent these requirements are fulfilled and what aspects have to be taken into account. As this method provides an innovative extension for tackling research topics in many projects and has already been adapted by the Research Data Center (FDZ) of the BA and the German Record Linkage Center (GRLC), this report explains the linkage of the ReLOC database and the underlying information in detail.

Keywords: Record linkage, firm-level data, establishment-level data, ReLOC

Ich möchte mich ganz besonders bei Tanja Hethy-Maier und Anja Gruhl bedanken, die mich als Mitarbeiterinnen des German Record Linkage Center (GRLC) zu Beginn meiner Arbeiten beraten und mir ihr Stata Do-File für das Preprocessing als Vorlage zur Verfügung gestellt haben. Herzlichen Dank auch an Manfred Antoni für die Mitlese und die sehr hilfreichen Anmerkungen zu diesem Methodenbericht. Das GRLC wird durch die Deutsche Forschungsgemeinschaft (DFG) gefördert.

Ich bedanke mich auch sehr herzlich bei Matthias Dörner vom Forschungsdatenzentrum (FDZ), der mir seine Modifizierungen des Stata Do-Files von Tanja Hethy-Maier und Anja Gruhl zur Verfügung gestellt hat. Vielen Dank auch an Cerstin Rauscher, Ali Athmani und Alaa Jasim vom Geschäftsbereich IT- und Informationsmanagement (ITM) für die Bereitstellung der benötigten Betriebsdaten und ihre Beratung hierzu.

1 Einleitung

Die Verknüpfung von Unternehmensdaten mit den Betriebsdaten des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) kann eine wichtige Grundlage für die Validität von Analysen sein. So stellt sich im Rahmen des Projekts „Research on Locational and Organisational Change“ (ReLOC) das Problem, dass für die adäquate Bearbeitung von Forschungsfragen eine Unternehmensdatenbank mit den Betriebsdaten des IAB verknüpft werden muss. Es handelt sich hierbei um eine Datenbank von deutschen Unternehmen, die zwischen 1990 und 2009 tschechische Unternehmen gegründet oder Anteile an deren Eigenkapital übernommen haben. Für eine möglichst vollständige Messung der Auswirkungen der Investitionen und den Verbindungen zwischen deutschen und tschechischen Unternehmen war es notwendig, möglichst alle Betriebe der investierenden deutschen Unternehmen zu identifizieren. Da die Daten des IAB keinen eindeutigen Unternehmens-Identifikator enthalten, stand bisher ohne eine Kooperation mit dem Statistischen Bundesamt kein Verfahren zur Verfügung, mit dem diese Zuordnung in systematischer Weise geleistet werden konnte. Aus diesem Grund wurde im Zuge des ReLOC-Projekts die Zusammenführung von Unternehmensdaten mit den Betriebsdaten des IAB um eine neue Methode ergänzt, welche in diesem Aufsatz beschrieben wird. Diese nutzt den Umstand, dass bei der Vergabe der Betriebsnummern durch den Betriebsnummern-Service der Bundesagentur für Arbeit (BA) für einen Betrieb der zugehörige Unternehmensname erfasst wird. Dies ermöglicht grundsätzlich die Verknüpfung von Betrieben und Unternehmen allein über den Unternehmensnamen und ohne Verwendung einer Adresse. Diese Methode findet mittlerweile auch in anderen Projekten am IAB Anwendung und stellt unter der gegenwärtigen Gesetzes- und Datenlage eine sinnvolle Erweiterung dar. Der Erfolg dieses Vorgehens hängt aber zum einen von der korrekten Erfassung der Unternehmensnamen und zum anderen von ihrer bundesweiten Einzigartigkeit ab. Deshalb wurden vor der Datenverknüpfung auf Basis der gesetzlichen Regelungen zur Namensgebung von Unternehmen die Korrektheit und die Struktur der Betriebs- und Unternehmensnamen näher untersucht. Da sich auch für andere Projekte Anknüpfungspunkte ergeben können, werden im Folgenden die dabei erzielten Erkenntnisse und das für diese Datenverknüpfung gewählte Vorgehen im Detail erläutert.

2 Ausgangslage

Das IAB ist die Forschungseinrichtung der BA. Zur Erfüllung seiner Forschungsaufgaben nutzt es unter anderem die im Rahmen der Datenerfassungs- und -übermittlungsverordnung (DEÜV) von den Arbeitgebern bzw. Unternehmen an die Sozialversicherungsträger und damit auch an die BA übermittelten Beschäftigtenmeldungen. Bei diesen Meldungen wird immer auch eine eindeutige Identifikationsnummer, die sogenannte Betriebsnummer, für den Betrieb, bei dem der jeweilige Arbeitnehmer beschäftigt ist, übermittelt. Dabei ist ein Betrieb im Sinne des Meldeverfahrens eine regional und wirtschaftsfachlich abgegrenzte Einheit mit mindestens einem sozialversicherungspflichtig oder geringfügig¹ Beschäftigten. Für die regionale Begrenzung ist die Gemeinde maßgebend. Daher können auch räumlich getrennte Niederlassungen² desselben Unternehmens, wenn sie innerhalb derselben Gemeinde liegen und denselben wirtschaftlichen Schwerpunkt aufweisen, unter einer Betriebsnummer zusammengefasst werden (Betriebsnummern-Service 2013). Ein Unternehmen kann wiederum mehr als eine Betriebsnummer umfassen, wenn es aus mindestens zwei Niederlassungen besteht und diese sich entweder auf unterschiedliche Gemeinden verteilen oder innerhalb

¹ Die geringfügige Beschäftigung unterliegt erst seit 01.04.1999 der Meldepflicht. Davor benötigten Betriebe mit lediglich geringfügig Beschäftigten keine Betriebsnummer.

² Grundsätzlich kann ein Unternehmen aus einer oder auch aus mehreren örtlichen Einheiten (Niederlassungen), die rechtlich nicht eigenständig, sondern Teil des Unternehmens sind, bestehen.

einer Gemeinde schwerpunktmäßig in unterschiedlichen Wirtschaftszweigen tätig sind.³ Die Vergabe der Betriebsnummer erfolgt über den Betriebsnummern-Service der BA.⁴ Dabei werden u.a. die Betriebsadresse und eine Bezeichnung für den Betrieb erfasst. Ein Unternehmens-Identifikator wird aber nicht vergeben. Das Unternehmensregister (URS) des Statistischen Bundesamtes enthält hingegen neben der BA-Betriebsnummer auch eine eindeutige Unternehmensnummer. Die im URS enthaltene Zuordnung von Betrieben zu Unternehmen kann aber aufgrund gesetzlicher Regelungen anderen Institutionen nicht ohne Weiteres zur Verfügung gestellt werden. So ist eine Verknüpfung der Datenbestände anderer Institutionen mit denen des Statistischen Bundesamtes ohne einer ausdrücklichen Zustimmung der betroffenen Erhebungseinheiten nicht möglich (Bender et al. 2007). Ein Projekt, in dem erstmals eine solche Datenzusammenführung erfolgte, ist das Projekt „Kombinierte Firmendaten für Deutschland“ (KombiFiD) (siehe u.a.: Spengler und Lorek 2010; Biewen et al. 2012). Dabei wurden Daten des IAB, des Statistischen Bundesamtes und der Deutschen Bundesbank miteinander verknüpft. Das URS fungiert dabei als Masterdatei, durch die in den Betriebsdaten des IAB alle Betriebsnummern desselben Unternehmens identifiziert werden können. Hierfür wurden rund 55.000 Unternehmen schriftlich um ihr Einverständnis gebeten, die Zustimmungquote beträgt rund 30%. Ziel des Projekts ist einerseits die Bereitstellung eines neuen und umfangreichen Datensatzes, der durch die Kombination der unterschiedlichen Datenbestände neue Analysemöglichkeiten bietet. Andererseits leistet es im Sinne einer Machbarkeitsstudie für zukünftige institutionsübergreifende Verknüpfungsprojekte Pionierarbeit. So wurde auch mithilfe eines Rechtsgutachtens geklärt, ob und unter welchen Voraussetzungen eine solche Datenverknüpfung ohne vorherige Zustimmung der betroffenen Erhebungseinheiten möglich ist (Biewen et al. 2012). Der daraus resultierende Vorschlag, den Geltungsbereich des §13a des Bundesstatistikgesetzes (BstatG) auf Datenbestände anderer Institutionen wie der BA und der Deutschen Bundesbank auszuweiten, hat bisher leider noch nicht zu einer Änderung der Gesetzeslage geführt.

Für kleinere Datenbestände und Projekte kann ein Vorgehen wie im KombiFiD-Projekt unter Umständen nicht geeignet sein. Denn zum einen ist der Ressourcenaufwand für die Kooperation und das Einholen der Zustimmung sehr hoch, zum anderen kann für statistische Analysen mancher Fragestellungen die Anzahl der Zustimmungen zu gering sein. Zudem besteht das Risiko durch ein heterogenes Antwort- oder Zustimmungsverhalten eine selektive Auswahl von Unternehmen zu erhalten (Biewen et al. 2012). Wenn die Selektivität von nicht beobachtbaren Variablen abhängt, kann sie die Ergebnisse in den darauf basierenden Forschungsanalysen verzerren. Das ReLOC-Projekt des IAB ist vor diesem Hintergrund auf einen alternativen Lösungsansatz angewiesen. Im Rahmen des Projekts wurde als Basis einer Unternehmensbefragung auf beiden Seiten der Grenze (Hecht et al. 2013a) eine Datenbank, bestehend aus deutschen Unternehmen mit tschechischem Tochter- oder Schwesterunternehmen, erstellt. Aus dem tschechischen Handelsregister, das öffentlich zugänglich ist, liegen u.a. Informationen über das Datum, zu dem die Investitionen der deutschen Unternehmen getätigt wurden, vor. Durch eine Identifizierung der deutschen Betriebe, deren Unternehmen in Tschechien investiert haben, kann ihre Beschäftigungsentwicklung vor und nach der Investition verfolgt werden. Hierbei ist es wichtig, möglichst alle betroffenen Betriebe zu identifizieren. Denn Investitionsentscheidungen werden in der Regel auf der Unternehmensebene getroffen und können alle Betriebe eines Unternehmens beeinflussen (Hecht et al. 2013b). Zudem können innerhalb eines Unternehmens die Betriebe je nach ihrer Größe, Tätigkeit und regionaler Lage unterschiedlichen Effekten ausgesetzt sein. So ist beispiels-

³ Dies ist abzugrenzen vom Begriff des Konzerns. Ein Konzern ist ein Zusammenschluss von Unternehmen unter einheitlicher Leitung, bei dem diese aber ihre rechtliche Selbstständigkeit behalten (§ 18 AktG).

⁴ Seit 01.01.2008 ist der Betriebsnummern-Service mit Sitz in Saarbrücken bundesweit für die Vergabe und Aktualisierung der Betriebsnummern zuständig. Zuvor erfolgte die Vergabe noch dezentral über die Arbeitsagenturen, in denen die jeweiligen Betriebe ihren Sitz hatten.

weise die Nutzung von Produktivitäts- und Faktorpreisunterschieden weniger attraktiv, wenn die Transportkosten hoch sind (Helpman 1984; Markusen 2002). Zugleich kann der von Feenstra und Hanson (1996) und Grossman und Rossi-Hansberg (2008) identifizierte Produktivitätseffekt bei geringerer Distanz zum tschechischen Betrieb stärker zum Tragen kommen. Betriebe an der deutsch-tschechischen Grenze unterliegen damit anderen Auswirkungen und Anreizen als Betriebe, die von ihren tschechischen Standorten weiter entfernt sind. Auch der Zweck und die Tätigkeit eines deutschen Betriebs beeinflussen, ob dieser mit dem tschechischen Standort in der Ausführung der Aufgaben konkurriert. Wenn in Tschechien bestimmte Aufgaben kostengünstiger erledigt werden können, erhöht dies die Wahrscheinlichkeit einer Reduzierung der Beschäftigten im deutschen Betrieb. Aufgrund komparativer Vorteile wäre zudem zu erwarten, dass deutsche Unternehmenszentralen, die zentrale Steuerungs-, Marketing- und F&E-Aufgaben erfüllen, anderen Effekten ausgesetzt sind als Produktionsstätten mit einem hohen Anteil von Routinetätigkeiten. Daher kann die alleinige Identifikation des Betriebs am offiziellen Sitz des Unternehmens zu Verzerrungen in der Messung der Auswirkung der Auslandsinvestition führen.

3 Datenbanken

3.1 ReLOC-Datenbank

Die zu verknüpfende ReLOC-Datenbank umfasst 3406 deutsche Unternehmen mit einem tschechischen Tochter- oder Schwesterunternehmen.⁵ Ausgangsbasis für die Identifikation der deutschen Unternehmen ist das tschechische Handelsregister. Dieses weist für alle tschechischen Unternehmen die Namen und Adressen inklusive des Landes der Eigentümer auf. Nach der Identifikation aller deutschen Eigentümer im Zeitraum von Januar bis August 2010 wurden die aktuellen Namen und Adressen der entsprechenden deutschen Unternehmen insbesondere aus kommerziellen Unternehmensdatenbanken wie GENIOS und FirmenWissen entnommen. Diese Datenbanken geben u.a. den im deutschen Handels- oder Genossenschaftsregister eingetragenen Unternehmenssitz kostenfrei an und greifen auch auf Daten von Creditreform zurück (Hecht et al. 2013b).

Die multinationalen Unternehmen bilden auch die Basis für eine in Deutschland und Tschechien durchgeführte Unternehmensbefragung. Bei dieser Befragung wurden in Deutschland die in Tschechien investierenden Unternehmen sowie auch als Referenzgruppe deutsche Unternehmen ohne Auslandsinvestitionen befragt. In Tschechien richtete sich die Befragung an die tschechischen Tochter-/Schwesterunternehmen sowie auch hier als Referenzgruppe an Unternehmen ohne ausländischen Eigentümer (Hecht et al. 2013b). Nach der erfolgreichen Verknüpfung der deutschen Unternehmen mit tschechischem Tochter- oder Schwesterunternehmen, die in diesem Artikel exemplarisch erläutert wird, wurden auch die deutschen Unternehmen ohne Auslandsinvestitionen nach demselben Vorgehen verknüpft.

Bei den 3406 deutschen Unternehmen mit tschechischer Tochter- oder Schwestergesellschaft handelt es sich um privatrechtliche Unternehmen mit folgenden Rechtsformen (siehe Tabelle 1):⁶ Gesellschaft mit beschränkter Haftung (GmbH), Gesellschaft mit beschränkter Haftung & Compagnie Kommanditgesellschaft (GmbH & Co. KG), Aktiengesellschaft (AG), Kommanditgesellschaft (KG), eingetragener Kaufmann bzw. eingetragene Kauffrau (e.K.), offene Handelsgesellschaft (OHG), Aktiengesellschaft & Compagnie Kommanditgesellschaft (AG & Co. KG), eingetragene Genossenschaft (eG), Unternehmergesellschaft (UG) (haftungsbeschränkt), Aktiengesellschaft & Compagnie Kommanditgesellschaft auf Aktien (AG &

⁵ In Hecht et al. (2013a) und Hecht et al. (2013b) werden aufgrund von zwei nachträglich identifizierten Doppelungen 3408 Unternehmen gezählt.

⁶ Siehe Kapitel 4.1 für nähere Erläuterungen zur Bedeutung der Kaufmannseigenschaft und den Rechtsformen aus Tabelle 1.

Co. KGaA), Limited Company (Ltd.), Limited Company & Compagnie Kommanditgesellschaft (Ltd. & Co. KG), Kommanditgesellschaft auf Aktien (KGaA), Kleingewerbetreibende ohne Handelsregister-Eintragung und freiberuflich tätige Unternehmer (inklusive Partnerschaftsgesellschaft).

Tabelle 1: Rechtsformen der ReLOC-Unternehmen

Rechtsform	Anzahl	Anteil
Unternehmen mit Kaufmannseigenschaft		
GmbH	2164	63,68%
GmbH & Co. KG	584	17,15%
AG	259	7,60%
KG	48	1,41%
e.K.	25	0,73%
OHG	15	0,44%
AG & Co. KG	13	0,38%
eG	6	0,18%
UG (haftungsbeschränkt)	5	0,15%
AG & Co. KGaA	4	0,12%
Ltd.	3	0,09%
Ltd. & Co. KG	1	0,03%
KGaA	1	0,03%
Unternehmen ohne Kaufmannseigenschaft		
Kleingewerbetreibende ohne Handelsregister-Eintragung und Freiberufler	278	8,16%
Summe	3406	100,00%

Quelle: ReLOC-Datenbank.

Damit ist die deutliche Mehrheit der Unternehmen (91,84%) im Handels- oder Genossenschaftsregister eingetragen bzw. besitzt die Kaufmannseigenschaft. Dies wird, wie sich in den späteren Ausführungen zeigen wird, für das Vorgehen und die Umsetzbarkeit der Datenverknüpfung eine wichtige Rolle spielen.

3.2 Betriebsdaten der BA

Alle Betriebe mit einem sozialversicherungspflichtig oder geringfügig Beschäftigten sind in einer zentralen historisierten Betriebsdatei der BA u.a. mit ihrer Betriebsnummer, ihrer Adresse und einer Betriebsbezeichnung, im Folgenden Betriebsname genannt, gespeichert. Nach der erstmaligen Erfassung durch den Betriebsnummern-Service werden auch Änderungen bestimmter Betriebsmerkmale wie dem Namen oder der Adresse, die grundsätzlich von den Arbeitgebern mitzuteilen sind, als neuer Eintrag gespeichert. So lassen sich auch alte Stände der Betriebsdatei wiederherstellen. In den zu Forschungszwecken genutzten Daten des IAB sind solche sensiblen Informationen aus Datenschutzgründen nicht enthalten. Stattdessen wird zum Zwecke der Anonymisierung die Betriebsnummer durch eine systemfreie und eindeutige Betriebsnummer ersetzt. Über eine Zuordnungstabelle lässt sich aber die systemfreie Betriebsnummer der echten zuordnen. Damit werden durch die Verknüpfung einer Unternehmensdatenbank mit den Betriebsdaten der BA auch die entsprechenden systemfreien Betriebsnummern identifiziert. Diese können dann in einem weiteren Schritt ohne Verwendung sensibler Merkmale den Betriebsdaten des IAB zugeordnet werden.

Für die Erfassung der Betriebsnamen stehen 3 Felder mit je 30 Stellen zur Verfügung (siehe Tabelle 2). Dabei kann das nächste Feld beginnen, ohne dass das vorhergehende vollständig gefüllt ist. Für Betriebe eines Unternehmens mit Kaufmannseigenschaft ist als erstes die

Firma, also der in den entsprechenden Registern (Handels-, Genossenschafts- oder Vereinsregister) verzeichnete und somit offizielle Unternehmensname (§ 17 HGB), erfasst. Dahinter folgen teilweise noch zusätzliche Angaben, die den Betrieb näher beschreiben. Diese geben oft den Niederlassungsort oder die Funktion des Betriebs und speziell bei e.K.s den Vor- und Nachname des Inhabers, insofern dieser nicht Bestandteil der Firma ist, wieder. Die Unternehmensnamen und die zusätzlichen Angaben verteilen sich nicht systematisch auf unterschiedliche Felder. Da die Unternehmensnamen aber an erster Stelle stehen, können sie durch die Identifizierung ihres Endes von den Zusätzen getrennt werden. Da eine Firma immer einen Rechtsform-Zusatz enthalten muss, der sich in der Regel an ihrem Ende befindet, kann der Name nach der Rechtsform abgeschnitten und damit die Firma vollständig identifiziert werden. Dies wird auch durch Gesetze begünstigt, die sicherstellen, dass der Rechtsform-Zusatz in der Firma entweder vollständig ausgeschrieben oder allgemein verständlich abgekürzt ist (siehe z.B. §19 HGB, §4 GmbHG, §4 AktG oder §3 GenG).

Bei Partnerschaftsgesellschaften und Gesellschaften bürgerlichen Rechts (GbR), die grundsätzlich nicht die Kaufmannseigenschaft und damit auch keine Firma besitzen, ist ebenfalls eine Bezeichnung für das Unternehmen angegeben.

Bei Einzelunternehmen⁷ ohne Kaufmannseigenschaft bzw. ohne Eintragung im Handelsregister sind der Vor- und Nachname des Inhabers und oftmals auch Branchen-, Tätigkeits-, Etablissement- oder andere Werbebezeichnungen erfasst. Dabei ist der Name⁸ des Inhabers nicht zwingend als erstes eingetragen, er befindet sich aber in einem separaten Feld.

Tabelle 2: Fiktive Beispiele für die bei der BA erfassten Betriebsnamen

Feld 1	Feld 2	Feld 3
Bauunternehmen Meier GmbH Malerwerkstatt Huber Gebäudereinigung Max Meier Maschinenbau GmbH TKF Bayerische MMAX Max Mustermann KG Werkzeug- und Maschinenbau Mustermann Service Deutschland Oliver Meier e.K. KTV Maschinentechnik e.K. Michael Schmidt Maxl's Grillstube ANL-Anlagentechnik Thomas Hoffmann	GmbH Hoffmann GmbH & Co. KG Elektroinstallations- und Technischer Großhandel Mustermann GmbH GmbH Niederlassung Martin Muster Garten- und Landschaftsbau Maximilian Meier Andreas Lieberknecht	-handels GmbH Werk Düsseldorf Berlin

Quelle: Ausgewählte Beispiele aus den Betriebsdaten der Statistik der BA zur Illustration der Verteilung und Struktur der Namensbestandteile von Betrieben privatrechtlicher Unternehmensformen.

⁷ Einzelunternehmen werden von einer einzelnen natürlichen Person (Gewerbetreibender, Freiberufler oder Landwirt) betrieben, die unbeschränkt mit ihrem Geschäfts- und Privatvermögen haftet.

⁸ Der Vorname steht jeweils vor dem Nachnamen.

4 Verknüpfungsmöglichkeiten

In den Betriebsdaten der BA ist der Unternehmensname der einzig mögliche Identifikator für die Unternehmenszugehörigkeit von Betrieben. Ob dies eine Zusammenführung allein über die Unternehmensnamen ermöglicht, hängt zum einen von ihrer bundesweiten Einzigartigkeit und zum anderen von ihrer korrekten Erfassung in den zu verknüpfenden Datenbanken ab. Die Grundlagen und die Erfüllung dieser Voraussetzungen werden in diesem Kapitel untersucht.

4.1 Einzigartigkeit von Unternehmensnamen

4.1.1 Unternehmen ohne Kaufmannseigenschaft

Die Regelungen zur Namensgebung und deren Einfluss auf die Einzigartigkeit von Unternehmensnamen unterscheiden sich bei privatrechtlichen Unternehmensformen grundlegend zwischen zwei Gruppen: Unternehmen ohne und Unternehmen mit Kaufmannseigenschaft. Der Name eines Unternehmens mit Kaufmannseigenschaft ist die sogenannte Firma (§ 17 HGB). Unternehmen, die nicht im Handelsregister eingetragen sind und damit nicht als Kaufmann gelten, haben hingegen keine Firma, sie tragen aber eine Unternehmensbezeichnung. Kleingewerbetreibende ohne Handelsregister-Eintragung und freiberuflich tätige Unternehmer sind dieser zweiten Kategorie zuzuordnen (IFB Nürnberg 2007).

Ein Kleingewerbe ist ein Unternehmen, „das nach Art oder Umfang einen in kaufmännischer Weise eingerichteten Geschäftsbetrieb nicht erfordert“ (§ 1 (2) HGB). Erfolgt keine freiwillige Eintragung in das Handelsregister, durch den auch ein Kleingewerbe die Kaufmannseigenschaft gemäß § 2 HGB erhält, gelten bei der Verwendung einer Unternehmensbezeichnung relativ strenge Regeln.⁹ So sollte ein Kleingewerbetreibender im Geschäftsverkehr grundsätzlich mit seinem Vor- und Nachnamen auftreten (IHK Köln 2011). Dies leitet sich aus verschiedenen Spezialvorschriften ab. Bis zum 24.03.2009 war noch in § 15b der Gewerbeordnung (GewO) übergreifend geregelt, in welcher Form Kleingewerbetreibende ohne Handelsregister-Eintragung im allgemeinen Geschäftsverkehr aufzutreten haben. So mussten sie auf allen Geschäftsbriefen, die an einen bestimmten Empfänger gerichtet waren, ihren Familiennamen mit mindestens einem ausgeschriebenen Vornamen angeben. Danach trat das Dritte Mittelstandsentlastungsgesetz (MEG III) zum Abbau bürokratischer Hemmnisse insbesondere in der mittelständischen Wirtschaft in Kraft, was den Wegfall dieses Paragraphen zur Folge hatte. Aber auch nach dessen Abschaffung wird u.a. von den Industrie- und Handelskammern empfohlen, mit dem Vor- und Nachnamen im Geschäftsverkehr aufzutreten, insofern keine Eintragung des Unternehmens im Handelsregister erfolgt. Dies liegt daran, dass sich weiterhin in verschiedenen Spezialvorschriften Rechtsgrundlagen für Pflichtangaben im Geschäftsverkehr finden. Meist handelt es sich dabei um vorvertragliche Informationspflichten. Beispiele finden sich unter anderem im Umsatzsteuergesetz (UStG), der Dienstleistungs-Informationspflichten-Verordnung (DL-InfoV) oder dem Telemediengesetz (TMG). Dabei setzt die Erfüllung der gesetzlichen Informationspflichten die vollständige Angabe von Vor- und Nachnamen voraus. Die Nennung nur eines Namensbestandteils reicht hier nicht aus (IHK Dresden 2013). Daneben ist es aber erlaubt zusätzlich zum Namen Branchen-, Tätigkeits-, Etablissement- oder sonstige Werbebezeichnungen (Buchstabenkombinationen, Fantasiebegriffe, etc.) zu verwenden (IFB Nürnberg 2007; IHK Dresden 2013).

Beispiele: „Max Mustermann Elektroanlagen“, „Max Mustermann Apotheke Sonnenstern“, „Max Mustermann IT-MMax“.

Eine andere Unternehmensform ist die freiberufliche Tätigkeit. Unter diese fällt nach § 18 des Einkommensteuergesetzes (EStG) insbesondere „die selbstständig ausgeübte wissen-

⁹ Durch eine Eintragung in das Handelsregister wird eine natürliche Person zu einem e.K..

schaftliche, künstlerische, schriftstellerische, unterrichtende oder erzieherische Tätigkeit“. Typische Freie Berufe sind unter anderem Arzt, Architekt, Anwalt, Steuer- und Unternehmensberater, Wissenschaftler und Künstler.¹⁰ Freiberufler müssen in ihrer Unternehmensbezeichnung nur ihren Nachnamen enthalten. Auch sie können Zusätze in den Namen einbauen (IFB Nürnberg 2007). Schließen sich zwei oder mehrere Freiberufler in einer sogenannten Partnerschaftsgesellschaft zusammen, so muss gemäß § 2 PartnGG der Nachname von mindestens einer Person, die in der Partnerschaft vertretenen Berufe sowie der Zusatz "und Partner" oder „Partnerschaft" in der Unternehmensbezeichnung enthalten sein.

Beispiele: „Dr. Martin Mustermann Steuerberater“, „Mustermann Freie Kulturwissenschaftlerin“, „Mustermann & Partner Rechtsanwälte“, „Dr. Huber & Meier Rechtsanwälte in Partnerschaft“.

Bei der Wahl der Unternehmensbezeichnung kann der Markenschutz eine wichtige Rolle spielen. So ist eine Marke durch einen Eintrag im Markenregister geschützt (MarkenG § 4 Nr. 1). Der Schutz einer eingetragenen Marke gilt innerhalb der entsprechenden Waren- und Dienstleistungsklassen bundesweit. Wenn sich die Unternehmensbezeichnung mit der Marke teilweise überschneidet, erhöht sich damit die Wahrscheinlichkeit, dass die Unternehmensbezeichnung bundesweit einzigartig ist. Das Markengesetz schützt unterscheidungskräftige Unternehmenskennzeichen (§5 (1) und (2) MarkenG) wie den Namen oder auch Teile davon auch ohne Eintragung als Marke. Dieser Schutz ist aber regional begrenzt, wenn das Unternehmen nach Zweck und Zuschnitt nur regional tätig ist und nicht auf eine darüber hinausgehende Expansion ausgelegt ist. Typische Beispiele sind Gaststätten, Hotels, Apotheken, Fahrschulen oder Friseursalons. Zudem erstreckt sich der Schutz nur über ähnliche Branchen (Ströbele und Paul 2012: S. 163-165, 1004 und 1011-1024). Wird nur der Personennamenname als Unternehmensbezeichnung verwendet, besteht kein Schutz, da ohne einer Eintragung im Handelsregister der Unternehmer sein Unternehmen selbst darstellt und die Verwendung des eigenen Namens anderen nicht untersagt werden darf (§12 BGB).

Damit hängt die bundesweite Einzigartigkeit der Unternehmensbezeichnung von Kleingewerbetreibenden ohne Handelsregister-Eintragung und Freiberuflern von der Häufigkeit der Inhabernamen und der verwendeten Zusätze ab. Der Schutz der Unternehmensbezeichnung durch das Markengesetz ohne Eintragung im Markenregister ist bei diesen Unternehmensformen mit ihrer oft sehr begrenzten regionalen Ausrichtung und Produktvielfalt eher als gering einzuschätzen. Angesichts der Vielzahl dieser Unternehmen ist daher häufig von der Nicht-Eindeutigkeit der Unternehmensbezeichnung auszugehen. Da diese Arten von Unternehmen aber per Definition sehr klein sind, ist es meist unwahrscheinlich, dass sie über mehrere Betriebe verfügen. Daher kann hier zur Identifikation des Unternehmens in der Betriebsdatei der BA die Adresse herangezogen werden.

4.1.2 Unternehmen mit Kaufmannseigenschaft

Für Unternehmen mit Kaufmannseigenschaft gelten hinsichtlich der Namensgebung weniger strenge Regeln. Hierzu gehört jedes Gewerbe, das „einen in kaufmännischer Art und Weise eingerichteten Geschäftsbetrieb“ erfordert (§1 HGB). GmbHs (§13 (3) GmbHG), AGs (§3 (1) AktG), KGs (§ 161 (1) HGB) und OHGs (§ 105 (1) HGB) gelten als Handelsgesellschaften und damit gemäß § 6 (1) HGB immer als Kaufmann. Dies gilt auch für GmbH & Co. KGs, AG & Co. KGs und Ltd. & Co. KGs, die Sonderformen der KG darstellen, wie auch für KGaAs und AG & Co. KGaAs¹¹, für die die Vorschriften der AG gelten (§ 278(3) AktG), und für UGs (haftungsbeschränkt), eine Sonderform der GmbH (§5 GmbHG). Auch eGs besitzen immer die Kaufmannseigenschaft (§ 17 (2) GenG). Natürliche Personen erhalten diese hingegen

¹⁰ 2011 gab es rund 1,15 Millionen Freiberufler in Deutschland (Brehm et al. 2012).

¹¹ Die AG & Co. KGaA ist eine Sonderform der KGaA.

nur, wenn das Erfordernis des kaufmännischen Geschäftsbetriebs vorliegt oder wenn sie von der Möglichkeit der freiwilligen Eintragung in das Handelsregister gemäß § 2 HGB Gebrauch machen. In diesem Fall müssen sie in ihrer Firma den Zusatz „eingetragener Kaufmann“ bzw. „eingetragene Kauffrau“ oder eine Abkürzung davon tragen (§ 19 (1) HGB).

Da diese Unternehmen im Handels- oder Genossenschaftsregister eingetragen sind, aus denen wiederum die rechtlichen Verhältnisse wie z.B. Inhaber oder Gesellschafter eines Unternehmens ersichtlich sind, besitzen sie eine wesentlich größere Wahlfreiheit hinsichtlich ihres Unternehmensnamens. Dieser setzt sich aus mehreren Teilen zusammen (IHK Köln 2011):

- Einer Bezeichnung, die einen Namen, einen Sach- oder einen Phantasiebegriff enthalten kann und aus einem oder mehreren Wörtern bestehen darf,
- dem Rechtsform-Zusatz sowie
- gegebenenfalls noch weiteren Zusätzen.

Beispiele: „Max Mustermann GmbH“, „Max Mustermann EDV-Service GmbH“, „Max Mustermann GmbH EDV-Service“, „MMax Mustermann GmbH“, „KTV EDV-Service GmbH“, „Mustermann EDV-Service Berlin GmbH“.

Damit aber die Identifizierbarkeit des Unternehmens und die eindeutige Zuordnung des Vertragspartners gewährleistet sind, gibt es auch hier bei der Namenswahl Einschränkungen. So muss gemäß § 18 (1) HGB die Firma zur „Kennzeichnung des Kaufmanns geeignet sein und Unterscheidungskraft besitzen“. Für den Unternehmensnamen wäre damit alles zulässig, was unterscheidungskräftig ist. Er muss keinen Sinn ergeben (eine „YXCVCBN GmbH“ wäre zulässig) und auch nicht zwingend den Namen des Inhabers, den Ort oder die Tätigkeit enthalten. Allgemeine Sachbezeichnungen oder Tätigkeitsbeschreibungen reichen zur Unterscheidbarkeit nicht aus. Eine „Handwerker GmbH“ würde beispielsweise kein Merkmal aufweisen, das es von anderen Handwerksbetrieben abgrenzt. Ein Unternehmensname, der nur aus Tätigkeits- und Regionalbezeichnungen besteht (*Beispiel: „Immobilienvermittlung Berlin GmbH“*), besitzt ebenfalls keine hinreichende Unterscheidungskraft. In solchen Fällen besteht die Möglichkeit über ein weiteres Wort (*Beispiele: „Mustermann Immobilienvermittlung Berlin GmbH“, „KTV Handwerker GmbH“*) die Unterscheidungskraft herzustellen (IHK Köln 2011).

Aber auch ein Unternehmensname, der grundsätzlich unterscheidungskräftig ist, kann unzulässig sein. Denn eine neue Firma muss sich „von allen an demselben Ort oder in derselben Gemeinde bereits bestehenden und in das Handelsregister oder in das Genossenschaftsregister eingetragenen“ deutlich unterscheiden (§ 30 (1) HGB). Im Gegensatz zum Kleingewerbe ohne Handelsregister-Eintragung und Freien Beruf gilt dies auch dann, wenn der Unternehmensname neben dem Rechtsform-Zusatz nur aus dem Vor- und Nachnamen des Inhabers besteht (HGB § 30 (2)). In einem solchen Fall muss dem Personennamen ein zur Unterscheidung fähiger Zusatz hinzugefügt werden. Allein ein anderer Rechtsform-Zusatz reicht hingegen nicht aus (IHK Köln 2011). Durch die Beschränkung auf die Gemeinde ist aber der Einfluss des § 30 (1) HGB auf die bundesweite Einzigartigkeit einer Firma wohl eher gering.

Auch für diese Unternehmensformen greift natürlich das in Kapitel 4.1.1 erläuterte Marken-gesetz. Damit existieren für Unternehmen mit Kaufmannseigenschaft einige Regelungen und Anreize, die auf einen bundesweit einzigartigen Namen hinwirken. Dazu zählt die Unterscheidungskraft eines Namens, die Einzigartigkeit innerhalb der politischen Gemeinde, der Schutz des Namens durch das Markengesetz wie auch ein grundsätzliches Interesse an einem Namen mit Wiedererkennungswert. Dennoch können Unternehmensnamen mehrfach vergeben sein. So ist die Eindeutigkeit einer Firma umso wahrscheinlicher, je größer der

Wirkungskreis eines Unternehmens und je größer das Unternehmen selbst ist. Denn mit der Größe des Unternehmens erhöhen sich auch tendenziell der Bekanntheitsgrad und die Produktvielfalt und dadurch wiederum der Namensschutz wie auch die Wahrscheinlichkeit der Eintragung einer Marke im Markenregister. Da die ReLOC-Datenbank multinationale und tendenziell größere Unternehmen umfasst, ist hier mit einem verhältnismäßig großen Schutz der Unternehmensnamen zu rechnen.

4.2 Überprüfung der zu verknüpfenden Datenbanken

4.2.1 Einzigartigkeit der Unternehmensnamen

Im vorhergehenden Abschnitt wurden die rechtlichen Grundlagen der Namensgebung von Unternehmen erläutert. Um einen Einblick in die tatsächliche bundesweite Einzigartigkeit der Unternehmensnamen in der zu verlinkenden Datenbank zu erhalten, wurde eine Zufallsstichprobe von 250 Unternehmen aus der ReLOC-Datenbank gezogen und die Häufigkeit dieser Namen über das deutsche Handels- und Genossenschaftsregister überprüft. Da die Verknüpfung der Datenbank über alle seit Bestehen des Meldeverfahrens (01.01.1973) verwendeten Betriebsnummern erfolgen soll, wurden alle jemals im Handels- oder Genossenschaftsregister eingetragenen Namen mitgezählt. Für Kleingewerbe ohne Handelsregister-Eintragung und Freiberufler existiert kein zentrales Register, deshalb wurden sie von vornherein aus diesem Test ausgeschlossen.¹²

Insgesamt wurden 18 Unternehmensnamen identifiziert, die mindestens zweimal im Handels- oder Genossenschaftsregister auftreten. 7,2% der Unternehmen haben damit keinen einzigartigen Namen. In 11 Fällen besteht der Name neben der Rechtsform aus nur einem Wort, wobei sich dieses wiederum in 10 Fällen aus maximal 6 Buchstaben zusammensetzt. In den verbleibenden 7 Fällen beinhaltet der Unternehmensname immer den Nachnamen einer Person, dabei in 3 Fällen auch den vollständigen Vornamen.

Mit der Länge und Komplexität der Unternehmensnamen nimmt erwartungsgemäß ihre Einzigartigkeit zu. Bestehen sie neben dem Rechtsform-Zusatz aus mehr als einem Wort und enthalten keinen Personennamen, sind sie bundesweit einzigartig. Das Risiko ist am größten, wenn sie neben dem Rechtsform-Zusatz nur ein kurzes Wort enthalten. Wird ein Teil des Personennamens verwendet, erhöht sich ebenfalls die Gefahr, dass der Unternehmensname mehrfach vergeben ist. Dies wird natürlich auch von der Verbreitung des Personennamens und möglichen weiteren Zusätzen beeinflusst.

Ist man nicht an der Unternehmensebene, sondern an der Konzernebene interessiert, dann reduziert sich das Risiko des mehrfach auftretenden Unternehmensnamens, da für 7 der 18 Fälle Internet- und Unternehmensregisterrecherchen ergeben haben, dass die Unternehmen mit dem identischen Namen zum selben Konzern gehören.

4.2.2 Fehler in den Betriebs-/Unternehmensnamen

Da die ReLOC-Datenbank bereits vor der Befragung aufbereitet wurde, steht die Korrektheit der Betriebsnamen im Vordergrund. Entsprechend ergibt sich auch bei einer Zufallsstichprobe aus der ReLOC-Datenbank, bestehend aus 100 Unternehmen mit Kaufmannseigenschaft, in 97% eine komplette Übereinstimmung mit dem jeweiligen Namen im deutschen Handels- oder Genossenschaftsregister.

¹² Es gibt lediglich für Partnerschaftsgesellschaften ein Register, das Partnerschaftsregister (PR), das wie auch das Handels- und Genossenschaftsregister über das Gemeinsame Registerportal der Länder (Zugang über: <https://www.handelsregister.de>) einsehbar ist.

Aus den Betriebsdaten der BA wurde ebenfalls eine Stichprobe von 100 Betrieben gezogen. Die Grundgesamtheit wurde auf Betriebe beschränkt, die zum 30.06.2009 mindestens einen sozialversicherungspflichtig oder geringfügig Beschäftigten hatten und die gemäß ihrem Rechtsform-Zusatz einem Unternehmen mit Kaufmannseigenschaft zuzuordnen sind. Die Betriebsnamen wurden im Handels- und Genossenschaftsregister recherchiert und dabei in 85 Fällen jeweils ein vollständig identisch lautender Unternehmensname gefunden. Hierbei wurde eine Standardisierung und Aufbereitung, das sogenannte Preprocessing, der Betriebsnamen, wie es auch im späteren Verlauf der Datenverknüpfung angewendet wird (siehe Kapitel 5), bereits berücksichtigt. Es zeigt sich auch, dass das Preprocessing, durch das der Name nach dem Rechtsform-Zusatz abgeschnitten wird, für die Datenverknüpfung sehr wichtig ist. Denn von den 85 Betriebsnamen weisen 22 (25,9%) nach dem Rechtsform-Zusatz weitere Wörter auf, die wiederum in 19 Fällen (22,4%) nicht Bestandteil der Firma sind.

Aus einer Analyse der Fehler bei der Namenserfassung lassen sich auch Strategien für die Datenverknüpfung ableiten. Daher wurde im Handels- und Genossenschaftsregister für die 15 fehlerhaften Betriebs-/Unternehmensnamen jeweils das Unternehmen gesucht, dem diese aller Voraussicht nach zuzuordnen sind. Demnach wurden am häufigsten ein oder mehrere Wörter weggelassen, hinzugefügt oder an eine falsche Position gestellt (siehe Tabelle 3). Eine Veränderung von Worten, entweder durch Rechtschreibfehler, durch eine absichtliche Abkürzung eines Wortes oder durch einen falschen Rechtsform-Zusatz, treten wesentlich seltener auf.

Tabelle 3: Fehler in den Betriebs-/Unternehmensnamen der BA

Art des Fehlers	Häufigkeit
Wort/Wörter weggelassen	7
Verdrehung der Reihenfolge der Wörter	4
Rechtschreibfehler	3
Wort/Wörter hinzugefügt	2
Wort/Wörter abgekürzt	2
Rechtsform falsch	1

Quelle: Betriebsdaten der Statistik der BA (Nürnberg 05/2012): Zufallsstichprobe von 100 Betriebsnummern mit mindestens einem sozialversicherungspflichtig oder geringfügig Beschäftigten zum 30.06.2009; Mehrfach-Nennungen sind möglich.

Das Hauptproblem besteht also darin, dass in fehlerhaften Betriebs-/Unternehmensnamen entweder zu viele oder zu wenige Wörter enthalten sind. Daher ist es sinnvoll, für die Berechnung der Ähnlichkeit der Namen ein Distanzmaß zu wählen, das fehlende oder überflüssige Namensbestandteile nicht zu stark bestraft und bei dem ihre genaue Position nur eine geringe Rolle spielt. N-Gramme stellen hier eine gute Alternative dar. Bei diesen wird der Name in Zeichenketten der Länge N aufgeteilt. Die Anzahl der übereinstimmenden Zeichenketten der verglichenen Namen entscheidet über deren Ähnlichkeit.

Beispiel mit Bi-Grammen (N=2) unter Anwendung von Padding¹³: Ähnlichkeit von „AHAB Hausbau“ und „AHAB Bau“ bei Entfernung von Leerstellen und Großschreibung aller Buchstaben.

Aus „AHABHAUSBAU“ werden die 12 Zeichenketten „_A“, „AH“, „HA“, „AB“, „BH“, „HA“, „AU“, „US“, „SB“, „BA“, „AU“ und „U_“. Aus „AHABBAU“ werden die 8 Zeichenketten „_A“, „AH“, „HA“, „AB“, „BB“, „BA“, „AU“ und „U_“. Damit ergeben sich 7 übereinstimmende („_A“,

¹³ Beim Padding werden der Anfang und das Ende des Ausdrucks mitberücksichtigt, indem an das erste und das letzte Zeichen eine Leerstelle angehängt wird.

„AH“, „HA“, „AB“, „BA“, „AU“ und „U_“) Zeichenketten und ein Dice-Koeffizient, als Maß für die Ähnlichkeit, von 0,7.¹⁴

N-Gramme haben den Vorteil, dass sie beim Vergleich zweier Namen die Namensbestandteile nicht nur im Umkreis einer bestimmten Stelle suchen. Im obigen Beispiel verdeutlicht sich das an den Zeichenketten „BA“, „AU“ und „U_“. Diese sind in beiden Namen enthalten, durch das Fehlen von „HAUS“ in „AHABBAU“ befinden sich diese Zeichenketten jeweils an unterschiedlichen Positionen. Dennoch erhöht sich durch sie der Dice-Koeffizient. Die Ähnlichkeit wird zwar auch dadurch verringert, dass das fehlende Wort in angrenzende Zeichenketten hineinreicht (hier: „BB“ anstatt „BH“ bzw. „SB“), die genaue Position der N-Gramme erfährt aber darüber hinaus keine zusätzliche Gewichtung in der Zählung der übereinstimmenden Zeichenketten. Das Fehlen eines Wortes verringert den Dice-Koeffizienten umso stärker, umso größer N gewählt wird. Denn mit größerem N tauchen die fehlenden Zeichen häufiger in den gebildeten Zeichenketten auf, während zugleich die Anzahl der gebildeten Zeichenketten sinkt. Unigramme (N=1) eignen sich hier nicht, da bei ihnen die Abgrenzung zwischen falschen und richtigen Paaren relativ gering ist. Man würde mit ihnen sogar bei zwei Namen, die aus denselben Zeichen bestehen, die sich aber alle an unterschiedlichen Positionen befinden, einen Dice-Koeffizienten von 1 erreichen. Daher werden später für die Berechnung der Ähnlichkeit von Betriebs-/Unternehmensnamen Bi-Gramme verwendet.

5 Datenverknüpfung

In den vorhergehenden Abschnitten hat sich gezeigt, dass die Verknüpfung einer Unternehmensdatenbank mit den Betriebsdaten der BA allein auf Basis des Unternehmensnamens möglich ist. Da für manche Betriebe die Unternehmensnamen nicht vollständig korrekt erfasst sind und zudem das Problem der bundesweit nicht einzigartigen Unternehmensnamen besteht, wird die Verlinkung in zwei Schritten erfolgen:

Im ersten Schritt werden sowohl die Namen als auch die Adressen der Unternehmen mit denen aus den Betriebsdaten abgeglichen. Folglich wird im ersten Schritt am Ort des Unternehmenssitzes nach Betrieben, die zu diesem Unternehmen gehören, gesucht.

Im zweiten Schritt werden alle Betriebe identifiziert, deren Namen mit dem eines Unternehmens aus der ReLOC-Datenbank übereinstimmen. Die Adressen der Unternehmen und Betriebe werden hierbei nicht verwendet. Zugleich werden die Unternehmensnamen, für die ein Betrieb gefunden wurde, durch Abfrage des Handels- und Genossenschaftsregisters auf ihre bundesweite Einzigartigkeit hin überprüft. Ist diese nicht erfüllt, wird der Betrieb nachträglich ausgeschlossen.

Das Record Linkage wurde von August bis Dezember 2012 durchgeführt. Die Betriebsdaten der BA umfassen alle bis zum 31.05.2012 jemals für das Meldeverfahren verwendeten Betriebsnummern, dies sind ca. 11,8 Millionen. Die Namen und Adressen aus der Betriebsdatei werden auf ihren Stand zum 31.12.2009 gebracht. Bei Betriebsnummern, die erst nach dem 31.12.2009 vergeben wurden, wird der erste Stand der dabei erfassten Namen und Adressen verwendet.

Die ReLOC-Datenbank setzt sich aus 3128 Unternehmen mit Kaufmannseigenschaft und 278 Unternehmen ohne Kaufmannseigenschaft zusammen. Diese Unterscheidung ist für die folgenden Verlinkungsschritte wichtig, da beide Gruppen aufgrund der unterschiedlichen Namensgebungsregeln (siehe Kapitel 4.1) teilweise unterschiedlich behandelt werden. So wird der zweite Verknüpfungsschritt, der allein auf Basis der Betriebs- und Unternehmens-

¹⁴ Dice-Koeffizient: $\text{Anzahl der Übereinstimmungen} * 2 / (\text{Anzahl der Zeichenketten im ersten String} + \text{Anzahl der Zeichenketten im zweiten String}) = 7 * 2 / (12 + 8) = 0,7$.

namen erfolgt, nur für Unternehmen mit Kaufmannseigenschaft durchgeführt. Die Identifikation der Betriebe von Unternehmen ohne Kaufmannseigenschaft erfolgt hingegen immer unter Einbezug ihres Betriebs-/Unternehmensnamens und ihrer Adresse.

Eine Verknüpfung der Betriebs- und Unternehmensdatensätze ist nur unter Anwendung einer Standardisierung und Aufbereitung, dem sogenannten Preprocessing, sinnvoll. Das Preprocessing erfolgt mittels eines Stata Do-Files. Hierfür konnte auf ein Stata Do-File von Tanja Hethy-Maier und Anja Gruhl (Biewen et al. 2012), mit zusätzlichen Modifizierungen von Matthias Dorner, zurückgegriffen und eigene Änderungen darin eingearbeitet werden.

Das Preprocessing der Betriebs-¹⁵ und Unternehmensnamen beinhaltet folgende aufeinander aufbauende Schritte:¹⁶

1. Anführungszeichen, Klammern und Bindestriche werden durch Leerstellen ersetzt.
2. Es werden allgemein bekannte Abkürzungen von oder innerhalb von Rechtsform-Zusätzen identifiziert und in ihre korrekt ausgeschriebene Form gebracht:

Beispiele: „GmbH“ wird zu „Gesellschaft mit beschränkter Haftung“, „KG“ zu „Kommanditgesellschaft“, „AG“ zu „Aktiengesellschaft“, „OHG“ und „oHG“ zu „offene Handelsgesellschaft“, „KGaA“ zu „Kommanditgesellschaft auf Aktien“, „StG“ zu „stille Gesellschaft“, „eG“, „e.G.“ und „e. G.“ zu „eingetragene Genossenschaft“ etc.

3. Standardisierung von „und“:

„u.“, „und“, „u“, „and“ und „+“ werden durch „&“ ersetzt.

4. Folgende Sonderzeichen werden durch Leerstellen ersetzt:

Punkte, Ausrufezeichen, Kommas, Doppelpunkte, Schrägstriche, umgekehrte Schrägstriche und Vergleichszeichen („<“ und „>“).

5. Großschreiben aller Buchstaben
6. Standardisierung von Umlauten:

Beispiele: „Ä“ zu „AE“, „Ü“ zu „UE“, „ß“ zu „SS“ etc.

7. Entfernung von Akzenten:

Beispiele: „Á“ zu „A“, „À“ zu „A“, „Â“ zu „A“ etc.

¹⁵ Der Betriebsname setzt sich aus den drei aufeinanderfolgenden Feldern, die für die Namenserfassung zur Verfügung stehen (siehe Kapitel 3.2), zusammen: Feld1 + Leerstelle + Feld2 + Leerstelle + Feld3.

¹⁶ Zwischen den Schritten werden wenn nötig Leerstellen am Anfang oder Ende des Betriebs-/Unternehmensnamens entfernt und mehrfach aufeinanderfolgende Leerstellen durch eine einzelne ersetzt.

8. Teilweise ausgeschriebene Rechtsform-Zusätze werden in ihre korrekt ausgeschriebene Form gebracht:

Beispiele: „GESELLSCHAFT MIT BESCHR HAFTUNG“ wird zu „GESELLSCHAFT MIT BESCHRAENKTER HAFTUNG“, „KOMMANDITGESELL“ zu „KOMMANDITGESELLSCHAFT“, „AKTIENGES“ zu „AKTIENGESELLSCHAFT“, „EINGETR KAUFM“ zu „EINGETRAGENER KAUFMANN“, „OFFENE HANDELSG“ zu „OFFENE HANDELSGESELLSCHAFT“ etc.

9. Die ausgeschriebenen Rechtsform-Zusätze werden identifiziert und in abgekürzter Schreibweise in einer neuen Variable gespeichert.
10. Alle Zeichen, die im Betriebs-/Unternehmensnamen nach dem (identifizierten) Rechtsform-Zusatz auftreten, werden entfernt. Der resultierende Name wird in einer neuen Variable gespeichert.
11. Entfernung von allen Leerstellen

Bei diesen Schritten werden für die Unternehmensnamen aus der ReLOC-Datenbank und für die Betriebsnamen aus den Betriebsdaten der BA u.a. folgende Variablen generiert (siehe Tabelle 4):

- *Name_ohne_Rechtsform*: aufbereiteter Betriebs-/Unternehmensname bis zum Rechtsform-Zusatz, der aber aus dieser Variable ausgeschlossen wird.
- *Rechtsform*: standardisierter Rechtsform-Zusatz in abgekürzter Schreibweise.
- *Name_mit_Rechtsform*: aufbereiteter Betriebs-/Unternehmensname bis zum Rechtsform-Zusatz, der in dieser Variable ebenfalls enthalten ist.

Tabelle 4: Fiktive Preprocessing-Beispiele

Betriebs-/Unternehmensname vor dem Preprocessing	Variablen nach dem Preprocessing		
	<i>Name_ohne_Rechtsform</i>	<i>Rechtsform</i>	<i>Name_mit_Rechtsform</i>
AKB Fräs-Technik GmbH Niederlassung Berlin	AKBFRAESTECHNIK	GmbH	AKBFRAESTECHNIKGmbH
AKB Frästechnik Gesell. mbH	AKBFRAESTECHNIK	GmbH	AKBFRAESTECHNIKGmbH
René Muster EDV Service	RENEMUSTEREDVSERVICE		RENEMUSTEREDVSERVICE
Rene Muster EDV-Service eingetr. Kaufmann	RENEMUSTEREDVSERVICE	e.K.	RENEMUSTEREDVSERVICEe.K.
Rene Muster EDV-Service e.K. Berlin	RENEMUSTEREDVSERVICE	e.K.	RENEMUSTEREDVSERVICEe.K.
"Insigno" Immobilienges. mbh	INSIGNOIMMOBILIEN	GmbH	INSIGNOIMMOBILIENGmbH
Klaus+Peter Huber KG Technischer Großhandel	KLAUS&PETERHUBER	KG	KLAUS&PETERHUBERKG
Klaus & Peter Huber Kommanditges. Zentrale Berlin	KLAUS&PETERHUBER	KG	KLAUS&PETERHUBERKG
Klaus u. Peter Huber KG	KLAUS&PETERHUBER	KG	KLAUS&PETERHUBERKG

Es werden unterschiedliche Schreibweisen folgender Rechtsformen berücksichtigt: GmbH, AG, KG, GmbH & Co. KG, OHG, eG, e.K., AG & Co. KG, AG & Co. KGaA, AG & Co. OHG, GmbH & Co. OHG, KGaA, GbR, e.V., StG, UG (haftungsbeschränkt), UG (haftungsbeschränkt) & Co. KG, Ltd., Ltd. & Co. KG, SE, SE & Co. KG.

Basierend auf den erläuterten Hintergründen zu den Betriebs- und Unternehmensnamen werden für diese Variablen gebildet, in denen alle Zeichen, die nach dem Rechtsform-Zusatz auftreten, nicht mehr enthalten sind (Schritt 10). Denn die zusätzlichen Angaben nach der Firma, wie sie in der Betriebsdatei teilweise zu finden sind (siehe Kapitel 3.2), würden ansonsten die errechneten Ähnlichkeiten verzerren und die Anzahl richtiger Verknüpfungen reduzieren. Wenn die Firma eines Unternehmens mit dem Rechtsform-Zusatz endet, in der ReLOC-Datenbank ist dies für 96% der Unternehmen mit Kaufmannseigenschaft erfüllt, dann wird in den Betriebsnamen der BA der Unternehmensname vollständig identifiziert, insofern er korrekt erfasst wurde.

Die Identifizierung von vollständig abgekürzten (Schritt 2) und teilweise ausgeschriebenen (Schritt 8) Rechtsform-Zusätzen erfolgt getrennt. So kann für die Abkürzungen der Informationsgehalt von Punkten und der Groß- und Kleinschreibung genutzt werden, während sich bei den teilweise ausgeschriebenen Rechtsformen durch die vorher durchgeführten Standardisierungen die Gefahr verringert, dass ein Rechtsform-Zusatz aufgrund unerwarteter Sonderzeichen oder Groß- und Kleinschreibung nicht erkannt wird.

Die Straßen- und Ortsnamen werden ebenfalls standardisiert. Auch hier werden alle Sonderzeichen, Akzente und Leerstellen entfernt, Buchstaben großgeschrieben und Umlaute standardisiert. Bei den Straßennamen wird zudem die Hausnummer von der Straße getrennt. Hieraus resultiert die Variable *Strasse_ohne_Hnr*, die den aufbereiteten Straßennamen ohne der Hausnummer enthält, und die Variable *Ort* mit der aufbereiteten Ortsangabe. Hierzu wurde ein Stata Do-File von Tanja Hethy-Maier ohne eigene Änderungen verwendet (Biewen et al. 2012).

5.1 Verwendung der Namen und Adressen

Zuerst erfolgt die Datenverknüpfung auf Basis der Namen und Adressen. Dabei wird schrittweise eine Lockerung der Namens- und Adresskriterien zugelassen. Eine Lockerung der Adresskriterien ist zum einen wegen möglicher Fehler in den Adressangaben sinnvoll. Zum anderen kann es auch sein, dass im Umkreis eines Unternehmenssitzes noch weitere Betriebe des Unternehmens angesiedelt sind.

Ohne dem Preprocessing gibt es nur sehr wenige Unternehmen in der ReLOC-Datenbank mit demselben Namen, demselben Straßennamen und demselben 3-stelligen Postleitzahlbezirk wie ein Betrieb aus der Betriebsdatei (siehe Tabelle 5). Es wird nur für 460 (14,7%) der Unternehmen mit und 8 (2,9%) der Unternehmen ohne Kaufmannseigenschaft ein Betrieb mit identischen Merkmalen gefunden. Das Preprocessing der Betriebs- und Unternehmensnamen wie auch der Adressangaben ist demnach zwingend notwendig. Dadurch erhöht sich die Anzahl der Unternehmen mit Kaufmannseigenschaft, für die mindestens eine Betriebsnummer (BNR) identifiziert wurde, auf 1904, was einem Anteil von 60,9% entspricht. Für Unternehmen ohne Kaufmannseigenschaft ist die Anzahl der Unternehmen mit mindestens einer Verknüpfung weiterhin sehr gering. Dies erklärt sich auch dadurch, dass es sich hierbei per Definition um sehr kleine Unternehmen handelt. Diese haben seltener einen sozialversicherungspflichtig oder geringfügig Beschäftigten und damit auch seltener eine Betriebsnummer.

Tabelle 5: Ergebnis des exakten Namens- und Adressabgleichs

Modell	Matching-Kriterien	Anzahl gültiger Verknüpfungen			
		mit Kaufmannseigenschaft		ohne Kaufmannseigenschaft	
		BNR	Unternehmen	BNR	Unternehmen
1	Ohne Preprocessing Name, Straße und 3-stellige PLZ identisch	489	460	8	8
2	Mit Preprocessing <i>Name_mit_Rechtsform,</i> <i>Strasse_ohne_Hnr</i> und 3-stellige PLZ identisch	2305	1904	17	16

Quelle: ReLOC-Datenbank und Betriebsdaten der Statistik der BA (Nürnberg 05/2012).

Da weiterhin ein großer Anteil von Unternehmen, für die in Modell 1 und 2 keine gültige Verknüpfung mit den Betriebsdaten gefunden wurde, verbleibt, wird in den nächsten Modellen der exakte Abgleich durch den Einsatz einer fehlertoleranten Ähnlichkeitsfunktion ersetzt (siehe Tabelle 6). Für die Betriebs-/Unternehmensnamen und Straßennamen werden die Ähnlichkeiten mittels Bi-Grammen berechnet. Dabei geht die Rechtsform nicht in die Berechnung mit ein. Der zur Bewertung der Ähnlichkeit herangezogene Index ergibt sich aus den aggregierten Dice-Koeffizienten der Betriebs-/Unternehmensnamen und Straßennamen. Da der Dice-Koeffizient der zu vergleichenden Variablen jeweils maximal 1 ergibt, wird bei vollständiger Übereinstimmung zweier Datensätze hinsichtlich ihres Namens und Straßennamens ein Wert von 2 erreicht. Die Berechnung erfolgt mit der Record Linkage Software Merge Toolbox (MTB) V0.73, die an der Universität Duisburg entwickelt wurde (Schnell et al. 2005).¹⁷ Hierbei wird auch das sogenannte „Blocking“ angewendet. Beim Blocking werden nur Beobachtungen aus den beiden Datenbanken miteinander verglichen, die in der entsprechend festgelegten Blocking-Variable auch denselben Wert aufweisen.

In Modell 3 werden die Unternehmen nur mit Betrieben, die entweder dieselbe 3-stellige Postleitzahl oder denselben Ortsnamen aufweisen, verglichen. Es werden alle Paare mit einem Ähnlichkeitsindex von mindestens 1,4 manuell überprüft und als „Match“ oder „Nicht-Match“ klassifiziert. Die Rechtsform wird zwar von der Berechnung des Ähnlichkeitsindex ausgeschlossen, in die Bewertung, ob es sich um einen richtigen „Match“ handelt, wird sie aber miteinbezogen. So werden 2908 Betriebsnummern 2422 Unternehmen mit Kaufmannseigenschaft und 174 Betriebsnummern 161 Unternehmen ohne Kaufmannseigenschaft zugeordnet. Damit erhöht sich die Verknüpfungsquote der Unternehmen auf 77,4% und 57,9%. Für Unternehmen ohne Kaufmannseigenschaft erhöht sich die Quote besonders stark. Denn bei diesen Unternehmensformen weichen die Namen in den beiden Datenbanken stärker voneinander ab, vor allem was die Erfassung von Branchen-, Tätigkeits-, Etablissement- oder sonstigen Werbebezeichnungen betrifft. Diese sind nicht verpflichtend im Namen zu führen und werden deshalb nicht immer konsistent angegeben bzw. verwendet. Dementsprechend beträgt in Modell 3 der durchschnittliche Ähnlichkeitsindex für Unternehmen ohne Kaufmannseigenschaft 1,71, für Unternehmen mit Kaufmannseigenschaft hingegen 1,95.¹⁸

In den Modellen 4 und 5 werden die Straßennamen nicht berücksichtigt. Der Ähnlichkeitsindex entspricht damit dem Dice-Koeffizienten für die Betriebs-/Unternehmensnamen und kann folglich einen Wert von maximal 1 erreichen. Da dadurch die Anzahl der möglichen Verknüpfungen steigt, werden in Modell 4 nur Verknüpfungen mit einem Ähnlichkeitswert von min-

¹⁷ Die Software ist kostenlos downloadbar unter: <http://record-linkage.de>.

¹⁸ Zwar gehen in den Ähnlichkeitsindex auch die Dice-Koeffizienten für die Straßennamen mit ein, diese erklären aber nicht diesen großen Unterschied.

destens 0,7 in die Klassifizierung als „Match“ oder „Nicht-Match“ miteinbezogen. In Modell 5 erweitern sich die Postleitzahlgebiete und damit auch die Anzahl der zu überprüfenden Verknüpfungen. Deshalb wird der Schwellenwert hier auf 0,8 gesetzt. Ein weiterer Unterschied zu den vorhergehenden Modellen ist, dass die Modelle 4 und 5 nur für Unternehmen mit Kaufmannseigenschaft angewendet werden. Denn durch das Wegfallen der Straße als Ähnlichkeitskriterium erhöht sich die Wahrscheinlichkeit einer falschen Verknüpfung für Unternehmen, die keinen bundesweit einzigartigen Namen tragen. Da dies besonders Unternehmen ohne Kaufmannseigenschaft betrifft und für diese Unternehmensform kein zentrales Register existiert, das die mehrfache Existenz eines Namens überprüfbar macht, werden diese aus Modell 4 und 5 ausgeschlossen.

In Modell 4 erhöht sich die Zahl der gültigen Verknüpfungen um 175 auf 3083, durch Modell 5 um weitere 813 auf 3896. Die Zahl der Unternehmen, für die mindestens eine Betriebsnummer identifiziert wurde, erhöht sich durch Modell 4 um 94 auf 2516 und durch Modell 5 um weitere 47 auf 2563. Der stärkere Zuwachs an gültigen Verknüpfungen ergibt sich damit durch Modell 5, während die Zahl der hierbei zugeordneten Unternehmen in Modell 4 stärker ansteigt. Dies erklärt sich dadurch, dass in Modell 5 die Betriebe einiger überdurchschnittlich großer Unternehmen, die in Ballungsgebieten angesiedelt sind, durch die Ausweitung der Postleitzahlbezirke miteinbezogen werden.

Tabelle 6: Ergebnis des fehlertoleranten Namens- und Adressabgleichs

Modell	Blocking-Variable	Variablen	Ähnlichkeitsfunktion	Anzahl gültiger Verknüpfungen			
				mit Kaufmannseigenschaft		ohne Kaufmannseigenschaft	
				BNR	Unternehmen	BNR	Unternehmen
3	3-stellige PLZ oder Ort	Name_ohne_Rechtsform und Strasse_ohne_Hnr	Bi-Gramme	2908	2422	174	161
4	3-stellige PLZ oder Ort	Name_ohne_Rechtsform	Bi-Gramme	3083	2516		
5	2-stellige PLZ oder Ort	Name_ohne_Rechtsform	Bi-Gramme	3896	2563		

Quelle: ReLOC-Datenbank und Betriebsdaten der Statistik der BA (Nürnberg 05/2012).

Damit wird insgesamt für 81,9% der Unternehmen mit Kaufmannseigenschaft mindestens eine gültige Verknüpfung gefunden. Für Unternehmen ohne Kaufmannseigenschaft liegt die Quote erwartungsgemäß deutlich darunter und beträgt 57,9%.

5.2 Nur Verwendung der Namen

Im zweiten Schritt werden zur Identifikation gültiger Verknüpfungen nur die Betriebs- und Unternehmensnamen verwendet. Die Adresse eines Unternehmens oder Betriebs hat folglich keinerlei Einfluss. Damit werden alle Betriebe in den Betriebsdaten der BA identifiziert, deren Namen mit dem eines Unternehmens aus der ReLOC-Datenbank übereinstimmen. Dieser Namensabgleich erfolgt über die Variable *Name_mit_Rechtsform*. Es werden nur Verknüpfungen als gültig anerkannt, bei denen ein Betrieb und ein Unternehmen hinsichtlich dieser Variable eine exakte Übereinstimmung, also einen Dice-Koeffizienten von 1, aufweisen.

Fiktives Beispiel: Dem Unternehmen „KTV Hausbau GmbH“ („KTVHAUSBAUGmbH“) werden die Betriebe „KTV Hausbau GmbH Niederlassung Berlin“ („KTVHAUSBAUGmbH“) und „KTV Haus-Bau Ges. mbH“ („KTVHAUSBAUGmbH“) zugeordnet, nicht aber Betriebe wie „KTG Hausbau GmbH“ („KTGHAUSBAUGmbH“) oder „KTV Hausbau GmbH & Co. KG“ („KTVHAUSBAUGmbH&Co.KG“).

Daher ist hier das in Kapitel 5 erläuterte Preprocessing von besonderer Bedeutung. Denn es kann keine darüber hinausgehende Fehlertoleranz zugelassen werden. Es wurde zwar ein fehlertolerantes Verfahren getestet, indem mittels Bi-Grammen die Ähnlichkeiten zwischen allen Betriebs- und Unternehmensnamen berechnet wurden. Dabei hat sich aber gezeigt, dass mit sinkender Ähnlichkeit der Namen die Anzahl falscher Verknüpfungen sehr viel stärker als die Anzahl richtiger Verknüpfungen steigt und aufgrund der Größe der Betriebsdatei die Überprüfung dieser Verknüpfungen zu arbeitsaufwendig ist. Damit führen bei diesem Verlinkungsschritt Fehler in der Namenserfassung, die nicht beim Preprocessing berücksichtigt werden, zu einer Nicht-Identifikation einer gültigen Verknüpfung.

Die aus dem exakten Namensabgleich resultierenden Verknüpfungen werden auf die Einzigartigkeit des jeweils dahinterstehenden Unternehmensnamens überprüft. Dies geschieht über das Handels- und Genossenschaftsregister.¹⁹ Da die Verknüpfung über den kompletten Betriebsnummernbestand erfolgt, werden auch Namen, die aus diesen Registern bereits wieder gelöscht wurden, mitgezählt. Die Einzigartigkeit des Unternehmensnamens wird aber nicht für alle Unternehmen getestet, sondern nur, wenn der (unaufbereitete) Name bis zum Rechtsform-Zusatz aus nur einem Wort besteht oder den Nachnamen einer Person enthält. Dieses Vorgehen leitet sich aus den Erkenntnissen aus Kapitel 4.2.1 ab, da aus der dort vorgestellten Stichprobe nur Unternehmensnamen, die diese Charakteristika aufweisen, nicht einzigartig sind. Ist ein Unternehmensname bundesweit mehrfach vergeben, so wird die Verknüpfung als ungültig markiert und nicht weiter verwendet. Dies dient der Vermeidung falscher Zuordnungen von Betrieben zu Unternehmen. Andererseits ist davon auszugehen, dass größere Unternehmen von dieser Problematik seltener betroffen sind (siehe Kapitel 4.1), wodurch sie bzw. ihre Betriebe in der endgültigen Datenbank möglicherweise leicht überrepräsentiert sind.

Der Namensabgleich wird wie schon in Modell 4 und 5 aus den besagten Gründen nur für Unternehmen mit Kaufmannseigenschaft durchgeführt. Insgesamt werden 53779 Betriebsnummern zu 2502 der 3128 (80,0%) Unternehmen gefunden (siehe Tabelle 7). Dabei ist das Kriterium der Namenseinzigartigkeit für 3017, 5,6% der zugeordneten Betriebsnummern, und 185 Unternehmen, 7,4% der Unternehmen mit mindestens einer Verknüpfung, verletzt. Damit verbleiben 50762 Betriebsnummern, die sich auf 2317 Unternehmen verteilen und die auch weiter verwendet werden.

Tabelle 7: Namenseinzigartigkeit potentieller Verknüpfungen

Unternehmensname bundesweit einzigartig	Betriebsnummern	Unternehmen
Ja	50762	2317
Nein	3017	185
Summe	53779	2502

Quelle: ReLOC-Datenbank und Betriebsdaten der Statistik der BA (Nürnberg 05/2012); Namenseinzigartigkeit wurde mithilfe des deutschen Handels- und Genossenschaftsregisters (Zugang über: www.handelsregister.de) überprüft.

5.3 Ergebnis aus beiden Verknüpfungsschritten

Tabelle 8 zeigt die endgültige Anzahl der gültigen Verknüpfungen. Insgesamt wurden den Unternehmen aus der ReLOC-Datenbank 51539 Betriebsnummern zugeordnet. Dabei sind 3293 Betriebsnummern sowohl im Verfahren, das die Betriebs-/Unternehmensnamen und Adressen nutzt, als auch im Verfahren, das nur die Betriebs-/Unternehmensnamen zum Abgleich verwendet, identifiziert worden. 777 Betriebsnummern werden hingegen nur im ersten

¹⁹ Diese Register werden auch von den Handelskammern genutzt, um für neu gegründete Unternehmen zu recherchieren, ob der von ihnen gewünschte Name bereits verwendet wird.

und 47469 nur im zweiten Schritt verknüpft. Damit ergibt sich auch aus dem ersten Schritt, der durch die Ähnlichkeitsfunktion Fehler in den Betriebs-/Unternehmensnamen berücksichtigt, ein nicht unerheblicher Zugewinn. Was mit diesem Verfahren nicht geklärt werden kann, ist die Anzahl der Betriebsnummern, die grundsätzlich nicht gefunden wurden. Hierzu bräuchte es einen Schlüssel wie den des Unternehmensregisters (URS), der eine fehlerfreie und vollständige Zuordnung von Betrieben zu Unternehmen ermöglicht.

Tabelle 8: Quelle der gültigen Verknüpfungen

		Nur Name		Summe
		nicht identifiziert	identifiziert	
Name und Adresse	nicht identifiziert	k.A.	47469	47469
	identifiziert	777	3293	4070
Summe		777	50762	51539

Quelle: ReLOC-Datenbank und Betriebsdaten der Statistik der BA (Nürnberg 05/2012).

Die 51539 Betriebsnummern verteilen sich auf 2903 Unternehmen. Damit liegt für 85,2% der Unternehmen mindestens eine Betriebsnummer vor. Unterscheidet man wieder nach Unternehmen mit und ohne Kaufmannseigenschaft, ergeben sich Unterschiede. So wurde für 2742 der 3128 und damit für 87,7% der Unternehmen mit Kaufmannseigenschaft mindestens eine gültige Verknüpfung gefunden. Für die 278 Unternehmen ohne Kaufmannseigenschaft findet sich für 161 und damit 57,9% mindestens eine gültige Verknüpfung.

Die ReLOC-Datenbank wurde mit dem kompletten Betriebsnummernbestand der BA verknüpft. Die Zahl der aktiven Betriebe eines Unternehmens kann sich aber über den Zeitverlauf ändern. So können neue Betriebsstätten eröffnet oder alte geschlossen werden. Zur Bewertung der tatsächlichen Aktivität eines Betriebs können seine Beschäftigtenmeldungen herangezogen werden. Tabelle 9 zeigt die Anzahl der Betriebsnummern pro Unternehmen, beschränkt auf Betriebe mit mindestens einem sozialversicherungspflichtig oder geringfügig Beschäftigten zum 30.06.2009. Unternehmen, die keinen Beschäftigten aufweisen, werden hier folglich nicht abgebildet. Unter den Unternehmen ohne Kaufmannseigenschaft haben 69,04% eine „aktive“ Betriebsnummer. 12,55% besitzen 2 „aktive“ Betriebsnummern und 5,63% sogar mehr als 9. Damit ist in dieser Datenbank der Anteil der Unternehmen mit nur einem Betrieb (Einbetriebsunternehmen) relativ gering. Zum Vergleich, laut Statistischem Bundesamt waren rund 98% aller Unternehmen, die zum 31.12.2002 mindestens einen sozialversicherungspflichtig Beschäftigten oder einen steuerbaren Umsatz aufgewiesen haben, Einbetriebsunternehmen (Nahm und Phillip 2005). Auch in den Betriebsdaten der BA treten rund 97% der aufbereiteten Betriebsnamen (*Name_mit_Rechtsform*), beschränkt auf Betriebe mit mindestens einem sozialversicherungspflichtig oder geringfügig Beschäftigten zum 30.06.2009, ausschließlich einmal auf. Der Anteil der Einbetriebsunternehmen ähnelt stattdessen eher dem des KombiFiD-Projekts. Bei diesem liegt der Anteil, abhängig vom Beobachtungsjahr und der Version, zwischen 71,7% und 77,5%. Auch in diesem Projekt besteht die Datenbank aus überdurchschnittlich großen Unternehmen (Biewen et al. 2012).

Tabelle 9: Anzahl der Betriebsnummern pro Unternehmen

BNR pro Unternehmen	Unternehmen			
	mit Kaufmannseigenschaft		ohne Kaufmannseigenschaft	
	Anzahl	Anteil	Anzahl	Anteil
1	1687	69,04%	97	97,98%
2	307	12,55%	2	2,02%
3	118	4,82%		
4	62	2,53%		
5	42	1,72%		
6	35	1,43%		
7	24	0,98%		
8	22	0,90%		
9	10	0,41%		
>=10	138	5,63%		
Summe	2445	100,00%	99	100,00%

Quelle: ReLOC-Datenbank und Betriebsdaten der Statistik der BA (05/2012 Nürnberg): Betriebe mit mindestens einem sozialversicherungspflichtig oder geringfügig Beschäftigten zum 30.06.2009.

6 Zusammenfassung und Ausblick

Die Verbindung der ReLOC-Datenbank mit den Betriebsdaten des IAB ermöglicht zum einen eine tiefgehende Analyse der Auswirkungen der deutschen Investitionen auf die investierenden Unternehmen und deren Betriebe. Zum anderen zeigt sie auch eine neue Methode der Verknüpfung von Unternehmens- mit Betriebsdaten. Diese benötigt nicht den Schlüssel des Statistischen Bundesamtes und kann damit für viele Projekte eine sinnvolle Alternative bieten. Für die Verknüpfung der ReLOC-Datenbank mit ihren multinationalen Unternehmen war dieses Vorgehen sehr gewinnbringend, da ihr Anteil von Einbetriebsunternehmen erwartungsgemäß sehr gering ist. Mit der alleinigen Verwendung der Betriebs- und Unternehmensnamen erhöht sich die Zahl der identifizierten Betriebsnummern in etwa auf das 13-fache. Da nicht alle Unternehmensnamen bundesweit einzigartig sind und nicht immer korrekt erfasst werden, kann es auch sinnvoll sein, in einem gesonderten Schritt die Adressen in die Datenverknüpfung miteinzubeziehen, wenngleich dadurch eine Überrepräsentanz von Unternehmenszentralen in Kauf genommen wird. Wie gezeigt wurde, sollten aber aufgrund unterschiedlicher Regelungen zur Namensgebung Unternehmen mit Kaufmannseigenschaft und Unternehmen ohne Kaufmannseigenschaft unterschiedlich behandelt werden.

Das hier vorgestellte Verfahren bietet auch Ansatzpunkte für eine Weiterentwicklung. So wäre eine Nutzung weiterer Variablen für die Zuordnung von Betrieben zu Unternehmen denkbar. Damit könnte ein fehlertolerantes Zuordnungsverfahren, das über das reine Pre-processing hinausgeht, noch stärker Anwendung finden. Eine vielversprechende Variable wäre beispielsweise die Anzahl oder der Anteil der Arbeitnehmer, die zwischen zwei Betrieben im Zeitverlauf wechseln. Das ähnelt in gewisser Weise dem Verfahren von Hethy und Schmieder (2012), das zur Identifikation von Betriebsschließungen und -gründungen entwickelt wurde. So könnte man auch Betriebe mit nicht identischem aber ähnlichem Namen und einem überdurchschnittlich hohen Austausch an Arbeitnehmern zu einem Unternehmen zusammenfassen. Bei Unternehmen mit nicht bundesweit einzigartigem Namen könnten damit einem Betrieb, der unter Zuhilfenahme der Unternehmensadresse identifiziert wurde, ebenfalls weitere Betriebe desselben Unternehmens zugeordnet werden. Die beste Alternative ist aber immer noch eine Änderung der Gesetzeslage hin zu einer praktikableren Zusammenführung administrativer Daten über institutionelle Grenzen hinweg, wie es im Zuge des KombiFiD-Projekts vorgeschlagen wurde (Biewen et al. 2012). Eine dauerhafte Verknüpfung des URS mit den Betriebsdaten des IAB würde für zukünftige Projekte mit verhältnismäßig geringem Aufwand die Mitberücksichtigung der Unternehmensebene ermöglichen. Solange sich die Gesetzeslage aber nicht ändert, stellen neue Ansätze, bei denen die Zuordnung auf Basis des Betriebs-/Unternehmensnamens erfolgt, eine sinnvolle Verbesserung dar.

Literatur

Bender, Stefan; Wagner, Joachim; Zwick, Markus (2007): KombiFiD – Kombinierte Firmen-daten für Deutschland. Konzeption der Machbarkeitsstudie für eine Zusammenführung von Unternehmensdaten der Statistischen Ämter, des Instituts für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit und weiterer Datenproduzenten. FDZ-Methodenreport, 05/2007, Nürnberg.

Betriebsnummern-Service (2013): Betriebsnummernvergabe. Stand Januar 2013, <http://www.arbeitsagentur.de/web/wcm/idc/groups/public/documents/webdatei/mdaw/mdk5/~edisp/16019022dstbai391187.pdf> (letzter Abruf am 09.05.2014).

Biewen, Elena; Gruhl, Anja; Gürke, Christopher; Hethey-Maier, Tanja; Weiß, Emanuel (2012): "Combined firm data for Germany" – possibilities and consequences of merging firm data from different data producers. Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, 132(2): 361-377.

Brehm, Thorsten; Eggert, Kerstin; Oberlander, Willi (2012): Die Lage der Freien Berufe. Institut für Freie Berufe (IFB), Nürnberg.

Feenstra, Robert C.; Hanson, Gordon H. (1996): Globalization, outsourcing and wage inequality. American Economic Review, 86(2): 240-245.

Grossman, Gene M.; Rossi-Hansberg, Esteban (2008): Trading tasks: a simple theory of offshoring. American Economic Review, 98(5): 1978-1997.

Hecht, Veronika; Hohmeyer, Katrin; Litzel, Nicole; Moritz, Michael; Müller, Jo-Ann; Phan thi Hong, Van; Schäffler, Johannes (2013a): Motive, Strukturen und Auswirkungen deutscher Direktinvestitionen in Tschechien – erste Untersuchungsergebnisse aus dem IAB-Projekt ReLOC - Research on Locational and Organisational Change. IAB-Forschungsbericht, 01/2013, Nürnberg.

Hecht, Veronika; Litzel, Nicole; Schäffler, Johannes (2013b): The ReLOC project – method report for implementing a cross-border company survey in Germany and the Czech Republic. IAB-Forschungsbericht, 04/2013, Nürnberg.

Helpman, Elhanan (1984): A simple theory of international trade with multinational corporations. Journal of Political Economy, 92(3): 451-471.

Hethey, Tanja; Schmieder, Johannes (2010): Using worker flows in the analysis of establishment turnover – evidence from German administrative data. FDZ-Methodenreport, 06/2010, Nürnberg.

IHK – Industrie- und Handelskammer Dresden (2013): Geschäftsbezeichnung und Namensangaben von nicht im Handelsregister eingetragenen Einzelunternehmer(r)n. Stand Januar 2013, http://www.dresden.ihk.de/servlet/link_file?link_id=27167&ref_knoten_id=48764&ref_detail=portal&ref_sprache=deu (letzter Abruf am 24.07.2013).

IHK – Industrie- und Handelskammer Köln (2011): Wie finde ich die richtige Firmierung für mein Unternehmen? Stand Januar 2011, http://www.ihk-koeln.de/upload/Voraussetzung_Firmierung_8901.pdf (letzter Abruf am 22.07.2013).

IFB – Institut für freie Berufe Nürnberg (2007): Wie darf man sein Unternehmen nennen. Gründungsinformation 32, Nürnberg.

Markusen, James R. (2002): Multinational firms and the theory of international trade, Cambridge: MIT Press.

Nahm, Matthias; Phillip, Katja (2005): Strukturdaten aus dem Unternehmensregister und Aspekte der Unternehmensdemografie. Statistisches Bundesamt, Wiesbaden.

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

Spengler, Anja; Lorek, Kerstin (2010): Verknüpfung und Abgleiche von Unternehmensregisterdaten des Statistischen Bundesamtes mit Betriebsdaten des Instituts für Arbeitsmarkt- und Berufsforschung. FDZ-Methodenreport, 01/2010 (aktualisiert am 13.08.2010), Nürnberg.

Ströbele, Paul; Hacker, Franz (2012): Markengesetz, Kommentar. 10. Auflage, Köln: Carl Heymanns Verlag.

Impressum

FDZ-Methodenreport 5/2014

Herausgeber

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

Redaktion

Stefan Bender, Johannes Schäffler, Axel Bülow

Technische Herstellung

Axel Bülow, Heiner Frank

Rechte

Nachdruck - auch auszugsweise - nur mit
Genehmigung des FDZ gestattet

Bezugsmöglichkeit

http://doku.iab.de/fdz/reporte/2014/MR_05-14.pdf

Internet

<http://fdz.iab.de/>

Rückfragen zum Inhalt an:

Johannes Schäffler,
IAB, Weddigenstr. 20-22, 90478 Nürnberg
Telefon: +49(0)911 / 179-2126
[mailto: Johannes.Schaeffler@iab.de](mailto:Johannes.Schaeffler@iab.de)