Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ-Methodenreport

Methodological aspects of labour market data

05/2014
EN

ReLOC linkage: a new method for linking firm-level data with the establishment-level data of the IAB

Johannes Schäffler

Bundesagentur für Arbeit

ReLOC linkage: a new method for linking firm-level data with the establishment-level data of the 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.

Contents

List of tables

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 applied for the first time in the project “Research on Locational and Organizational 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

I would particularly like to thank Tanja Hethey-Maier and Anja Gruhl, former and current employees of the German Record Linkage Center (GRLC), respectively, who gave me advice at the beginning of my research and provided me with their Stata do-file as a template for preprocessing. Many thanks also to Manfred Antoni for reading this method report and making very helpful comments. The GRLC is sponsored by the German Research Foundation (DFG).

I would also very much like to thank Matthias Dorner from the Research Data Center (FDZ), who provided me with his modifications of the Stata do-file from Tanja Hethey-Maier and Anja Gruhl. Many thanks also to Cerstin Rauscher, Ali Athmani and Alaa Jasim from the IT Services and Information Management (ITM) division for providing the necessary establishment-level data and for their advice in this regard.

1 Introduction

The linkage of firm-level data with the establishment-level data of the Institute for Employment Research (IAB) can form a significant basis for the validity of analyses. As part of the "Research on Locational and Organizational Change" (ReLOC) project, the problem arose that a firm-level database had to be linked to the IAB's establishment-level data for research questions to be adequately processed. This is a database of German companies that founded Czech companies between 1990 and 2009 or that took over shares of their equity. To measure the effects of the investments and connections between German and Czech firms as fully as possible, it was necessary to identify all of the establishments belonging to the German firms investing if possible. As the IAB data does not contain a clear firm identifier, there has been no procedure available as yet to carry out this assignment systematically without the cooperation of the German Federal Statistical Office. For this reason, a new method was added to the linkage of firm-level data with the IAB's establishment-level data in the course of the ReLOC project, and this method will be described in this article. It makes use of the fact that in the course of assigning IDs to establishments, the Federal Employment Agency (BA) records the name of the associated company. This basically enables the linkage of companies and establishments solely on the basis of the company name and without using an address. Meanwhile, this method is also being applied in other IAB projects and represents an expedient addition considering the current situation in regard to data and legislation. However, the success of this approach depends on the company names being recorded correctly and also on their being unique throughout Germany, so before the data was linked, the correctness and structure of the establishment and company names were closely checked based on the legal requirements for naming companies. As links can also arise for other projects the findings and the procedure selected for this data linkage will be explained in detail in the following.

2 Initial situation

The IAB is the research facility of the BA. To fulfill its research tasks, one of the data sources it uses is the registrations of employees which employers pass on to the social insurance agencies and thus to the BA as part of the Data Capture and Transfer Regulation (DEÜV). When these registrations are made, a distinct ID number, the establishment number (Betriebsnummer – BNR), is assigned for the establishment where the respective person is employed. In terms of the registration process, an establishment is a defined regional and economic unit with at least one employee who is liable for social insurance contributions or is in marginal part-time employment.[1] In regard to the regional limitation, it is the municipality which is decisive. Therefore, branches[2] of the same firm which are in different locations can be assigned the same establishment number (Betriebsnummern-Service 2013) if they are in the same municipality and show the same business focus. On the other hand, a firm can have more than one establishment number if it consists of at least two branches, and these are distributed in different municipalities or operate in different industrial sectors within one municipality.[3] The BA Betriebsnummern-Service assigns the establishment numbers.[4] Infor-

1 Marginal part-time employment has only been subject to declaration since April 1, 1999. Prior to that, establishments which only had employees in marginal part-time employment did not require an establishment number.

2 In principle, a firm can consist of one or more local units (branches) which are part of the firm and are not legally separate.

3 This must be distinguished from the term "affiliated group". An affiliated group is an association of firms under one leadership, but where these firms maintain their legal autonomy (Section 18 of the German Corporation Act (AktG)).

4 The Betriebsnummern-Service is located in Saarbrücken and has been responsible for the national assignment and updating of establishment numbers since January 1, 2008. Prior to that the

mation such as the address of the establishment and a name of it is recorded here, but firm identifiers are not assigned. On the contrary the Business Register System (URS) of the Federal Statistical Office contains a distinct firm number as well as the BA establishment number. However, the assignment of establishments to firms contained in the URS cannot be made available to other institutions due to statutory rules, so it is not possible to link the databases of other institutions with those of the Federal Statistical Office without express consent of the observation units affected (Bender et al. 2007). One project in which a data fusion like this occurred for the first time was the "Combined Firm Data for Germany" (KombiFiD) project (see Spengler and Lorek 2010; Biewen et al. 2012, for example). Here, the data from the IAB, the Federal Statistical Office, and the German Federal Bank was linked, where the URS served as a master file with which all IDs of the same firm were identified in the IAB's establishment-level data. For that project, around 55,000 firms were written to asking for their permission, which was then granted by around 30 percent of them. The aim of the project is, on the one hand, to provide a new, comprehensive database which offers new possibilities for analysis through the combination of the different databases. On the other hand, it is pioneering work in the sense of a feasibility study for future cross-institutional linkage projects. With the aid of a legal opinion, it was also clarified whether a data linkage of this kind is possible without prior consent, and if so, under what conditions (Biewen et al. 2012). The resulting suggestion to extend the scope of application of Section 13a of the Federal Statistics Act (BstatG) to databases of other institutions like the BA and the German Federal Bank has unfortunately not yet led to a change in legislation.

Implementing a procedure such as that used for the KombiFiD project may not be suitable for smaller databases or projects, because for one thing the resource costs of cooperation and gaining consent are very high, and for another, the number of firms giving their consent may be too low for the statistical analysis of some questions to be possible. Furthermore, there is the risk of a selective choice of companies due to a heterogeneous response or consent behavior (Biewen et al. 2012). If the selectivity depends on non-observable variables, it can distort the results in the research analyses based on it. Against this background, the IAB's ReLOC project had to rely on an alternative approach. In the scope of the project, a database consisting of German companies with Czech affiliates was created as the basis for a company survey on both sides of the border (Hecht et al. 2013a). In the Czech Commercial Register, which is publicly accessible, there is information available on the date on which the German companies made the investments. By identifying the German establishments whose companies have invested in the Czech Republic, it is possible to follow their employment trends before and after the investment. Here it is important to identify all the establishments involved if possible, as investment decisions are generally made at the firm level and can influence all of a firm's establishments (Hecht et al. 2013b). Furthermore, the establishments within a firm can be affected differently depending on their size, activity and regional position. Using differences in productivity and factor prices, for example, is not particularly attractive if the transport costs are high (Helpman 1984; Markusen 2002). At the same time, the productivity effect which Feenstra and Hanson (1996), and Grossman and Rossi-Hansberg (2008) identified, can be more marked when the distance to the Czech affiliate is smaller. Establishments on the German-Czech border are therefore subject to different effects and incentives than those that are farther away from their Czech locations. The purpose and activity of a German establishment also have an influence on whether this establishment is competing with the Czech affiliate in regard to the performance of its tasks. If certain tasks can be performed more cheaply in the Czech Republic, this makes it more likely that the number of staff in the German establishment will be reduced. Based on comparative advantages, it is also to be expected that German headquarters carrying out central management, marketing and R&D will be confronted with different effects than those facing production plants with a high

assignment was made locally by the employment agencies in the municipality in which the respective establishments were located.

percentage of routine tasks. Therefore, identifying the establishments solely on the basis of the official company headquarters can lead to distortions in the measurement of the effect of foreign investments.

3 Databases

3.1 ReLOC database

The ReLOC database which is to be linked consists of 3406 German companies with a Czech affiliate.[5] The starting point for identifying the German companies is the Czech Commercial Register, which shows the names and addresses of all Czech companies and their owners (including the country). After all German owners were identified in the period from January to August 2010, the current names and addresses of the corresponding German firms were retrieved from commercial company databases like GENIOS and FirmenWissen in particular. These databases offer information such as the company location entered in the German Commercial Register or Register of Cooperatives free of charge, and also refer to the commercial data provider Creditreform (Hecht et al. 2013b).

The multinational firms also formed the basis of a company survey conducted in Germany and the Czech Republic. For this survey, the German firms investing in the Czech Republic were interviewed. Additionally, German firms without a foreign affiliate were interviewed as a reference group. In the Czech Republic, the survey was aimed at the Czech affiliates, and companies without foreign owners were likewise taken as a reference group (Hecht et al. 2013b). After the successful linkage of the German firms with a Czech affiliate, which is explained as an example in this article, the German firms without foreign investments were then linked in the same way.

The 3406 German firms with Czech affiliates are companies under German private law with the following legal forms (see Table 1): *Gesellschaft mit beschränkter Haftung* (*GmbH*) [private limited liability company], *Gesellschaft mit beschränkter Haftung & Compagnie Kommanditgesellschaft* (*GmbH & Co. KG*) [private limited liability company & Co. limited partnership], *Aktiengesellschaft* (*AG*) [stock corporation], *Kommanditgesellschaft* (*KG*) [limited partnership], *eingetragener Kaufmann* or *eingetragene Kauffrau* (*e.K.*) [registered merchant], *offene Handelsgesellschaft* (*OHG*) [general partnership], *Aktiengesellschaft & Compagnie Kommanditgesellschaft* (*AG & Co. KG*) [stock corporation & Co. limited partnership], *eingetragene Genossenschaft* (*eG*) [incorporated cooperative], *Unternehmergesellschaft* (*UG*) (*haftungsbeschränkt*) [limited liability entrepreneurial company], *Aktiengesellschaft & Compagnie Kommanditgesellschaft auf Aktien* (*AG & Co. KGaA*) [stock corporation & Co. limited joint-stock partnership], Limited Company (Ltd.), *Limited Company & Compagnie Kommanditgesellschaft* (*Ltd. & Co. KG*) [limited company & Co. limited partnership], *Kommanditgesellschaft auf Aktien* (*KGaA*) [limited joint-stock partnership], traders not registered in the Commercial Register, and freelancers (including *Partnerschaftsgesellschaften*).

[5] Due to two duplications identified later, there were 3408 companies in number in Hecht et al. (2013a), and Hecht et al. (2013b).

Table 1: Legal forms of the ReLOC firms

Legal form	Number	Percentage
Firms with *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%
Firms without *Kaufmannseigenschaft*		
Traders not registered in the Commercial Register and freelancers	278	8.16%
Total	3406	100.00%

Source: ReLOC database.

Therefore, a clear majority of the firms (91.84%) are registered in the Commercial Register or Register of Cooperatives, and are consequently characterized as merchants in terms of the German Commercial Code (*Kaufmannseigenschaft*).[6] As will be shown later, this played an important role in the procedure and implementation of the data linkage.

3.2 The BA's establishment-level data

All establishments with at least one employee who is liable for social insurance contributions or is in marginal part-time employment are stored in the BA's central historized establishment files which contains data such as their establishment number, address, and name. After the initial registration by the BA Betriebsnummern-Service, changes in particular characteristics of the establishment which the employers are obligated to pass on, such as name or address, are stored as a new entry, so that the old statuses of the establishment files can also be restored. For reasons of data privacy, sensitive information like this is not contained in the IAB data used for research purposes. Instead, the establishment number is replaced with a unique, system-independent establishment number for the purposes of anonymization. However, this system-independent establishment number can be connected to the real one by means of an assignment table. This can be used to identify the respective system-independent establishment numbers by linking a company database with the BA's establishment-level data. They can then be correlated to the IAB's establishment-level data in a further step without using sensitive characteristics.

There are three fields available, each containing space for 30 characters, to record the establishment name (see Table 2). It is possible to enter information in a field before the previous field has been filled completely. For establishments of a company with *Kaufmannseigen-*

[6] See Section 4.1 for more details on the meaning of the *Kaufmannseigenschaft* and the legal forms from Table 1.

schaft[7]*,* the first item to be recorded is the legal company name – that is, the company name which is listed in the Commercial Register or Register of Cooperatives.[8] This is sometimes followed by additional information describing the establishment in more detail. This often gives the location or function of the establishment, and, specifically in the case of registered merchants, the first and last name of the owner if their name is not part of the company name. The company name and additional details are not systematically distributed across different fields. However, as the company name comes first, it can be separated from any additional designations when its ending is identified. As a name of a company with *Kaufmannseigenschaft* must always include a designation indicating its legal form, and as this is generally to be found at the end, the name can be cut off after this designation so that the company name can be fully identified. This is promoted by laws which ensure that the designation indicating the legal form of the company is either written out in full or is abbreviated such that it is generally comprehensible (see Section 19 HGB [German Commercial Code], Section 4 GmbHG [German Limited Liability Companies Act], Section 4 AktG [German Corporation Act], or Section 3 GenG [German Cooperative Societies Act], for example).

A company name is likewise given for establishments of *Gesellschaften bürgerlichen Rechts* (GbR) [partnership under civil law], *Partnerschaftsgesellschaften*, and associations. In the case of *Partnerschaftgesellschaften* and registered associations (*e.V.*), this is the name which is listed in the pertinent register (Register of *Partnerschaftsgesellschaften* or Register of Associations).

In the case of sole traders which are not recorded in the Commercial Register, freelancers, and farmers, the first and last name of the owner, and often designations concerning industry, activities, establishment or other advertising purposes are recorded. The name[9] of the owner does not necessarily appear first, but is entered in a separate field.

Table 2: Fictitious examples of the establishment names recorded at the BA

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

Source: Selected examples from the establishment-level data from the BA statistics department to illustrate the structure and distribution of the name components of establishments appendant to companies under private law.

7 This includes any company that has to be registered in the Commercial Register or Register of Cooperatives, and therefore any *GmbH*, *AG*, *KG*, *OHG*, *eG*, and *e.K.*, as well as any special/mixed form of these legal forms, such as *GmbH & Co. KG*, *KGaA*, *UG (haftungsbeschränkt)*, *SE*, *AG & Co. KG*, *AG & Co. KGaA*, *AG & Co. OHG*, *GmbH & Co. OHG*, *SE & Co. KG*, *UG (haftungsbeschränkt) & Co. KG*, and *Ltd. & Co. KG* (see also Section 4.1 for more details).

8 Under the terms of the German Commercial Code, the legal name of a company with *Kaufmannseigenschaft* is the "Firma" (Section 17 HGB).

9 The first name stands before the last name.

4 Linkage possibilities

In the BA's establishment-level data, the firm name is the only possible identifier for the firm the establishments belong to. Whether this allows them to be combined purely by means of the firm name depends on how unique they are in Germany, and on their having been recorded correctly in the databases to be linked. The basis for these prerequisites and their being fulfilled will be examined in this section.

4.1 Uniqueness of firm names

4.1.1 Firms without *Kaufmannseigenschaft*

In the case of companies under private law, the regulations concerning the naming of firms and their influence on the uniqueness of firm names basically divide these organizations into two groups: firms with *Kaufmannseigenschaft* and firms without *Kaufmannseigenschaft*. Traders not registered in the commercial register, and freelance entrepreneurs are assigned to the latter category.

A trade that does "not require a business concern to be set up for trading due to its size or type" is not characterized as a merchant in terms of the German Commercial Code, and does therefore not obtain the *Kaufmannseigenschaft* except if it is registered voluntarily in the Commercial Register (Section 1 and Section 2 HGB). For this kind of companies the rules concerning the use of a company name are relatively strict. In business dealings, such small traders should use their first and last name as a matter of principle (IHK Köln 2011). This derives from various specific stipulations. Until March 24, 2009, Section 15b of the German Industrial Code (GewO) comprehensively regulated how traders not listed in the Commercial Register were to present themselves in general business dealings. They had to put their last name and at least one first name written in full on all business letters to a particular recipient. Subsequently, the Third SME Relief Act (MEG III) came into effect to reduce bureaucratic constraints for small businesses in particular, which led to the abovementioned section being declared void. However, even after it was abolished, the Chambers of Industry and Commerce recommend, for example, that traders present themselves with their first and last names in business dealings if their company is not listed in the Commercial Register. This is because there are still statutory sources for required data in business dealings in various specific stipulations. This is mostly in regard to the precontractual duty to provide information. There are examples in the Value Added Tax Act (UStG), the Regulation on the Service Providers' Duty to Inform (DL-InfoV), or the Telemedia Act (TMG). Here, first and last names have to be given in full to comply with the statutory duty to provide information, and it is not sufficient to give only part of the name (IHK Dresden 2013). However, it is permissible to use designations concerning industry, activities, establishment or other designations for advertising purposes (letter combinations, made-up terms, etc.) in addition to the name (IFB Nürnberg 2007; IHK Dresden 2013).[10]

Examples: "Max Meyer IT-Service", "Max Meyer Café Alphorn", "Max Meyer IT-MMax".

Another form of an enterprise without *Kaufmannseigenschaft* is freelance activity. Pursuant to Section 18 of the German Income Tax Act (EstG), this particularly includes "a scientific, artistic, writing, or teaching activity performed as a self-employed person". Physicians, architects, lawyers, tax consultants, management consultants, scientists, and artists are typical examples of independent freelance professions.[11] Freelancers only have to include their last

[10] In the case of two or more traders forming a partnership under civil law (*GbR*), the recommendations are basically the same. Besides this, a designation indicating the legal form can be added. *Example: "Max Meyer & Andreas Schmidt IT-Service GbR"*.

[11] In 2011 there were around 1.15 million freelancers in Germany (Brehm et al. 2012).

name in their company name, and they are also allowed to add additional information (IFB Nürnberg 2007). If two or more freelancers form a *Partnerschaftsgesellschaft*, the company name must include the name of at least one person, the professions practiced in the partnership, and the additional words "und Partner [and partner]" or "Partnerschaft [partnership]", pursuant to Section 2 of the *Partnerschaftsgesellschaftsgesetz* (PartnGG).

Examples: "Dr. Joseph Meyer Steuerberater [Tax Consultant]", "Meyer Freie Kulturwissenschaftlerin [Cultural Scientist]", "Meyer & Partner Rechtsanwälte [Attorneys]", "Dr. Peters & Meyer Rechtsanwälte in Partnerschaft [Attorneys in Partnership]".

Trademark protection can play an important role in the choice of the company name. A brand name is protected by a listing in the trademark register (Section 4 no. 1, German Trademark Act (MarkenG)). The protection of a registered brand is valid throughout Germany in the relevant classes of goods and services. If the brand and the company name partly overlap, this increases the likelihood of the company name being unique in Germany. The German Trademark Act protects distinctive business identifiers (Section 5 (1) and (2), (MarkenG)) like the name or parts of the name even if they are not registered as a brand. However, this protection is limited to the region if the purpose and location of the company is only intended to be regional and there is no intent to expand the company outside the region. Hotels, guest houses, pharmacies, driving schools, or hairdressing salons are typical examples of this. Furthermore, the protection is only extended to similar sectors (Ströbele and Paul 2012: pp. 163-165, 1004 and 1011-1024). If only the person's name is used as a company name, it is not protected, as entrepreneurs are presenting their companies themselves if they are not listed in the Commercial Register, and it is not possible to forbid others to use their own name (Section 12 German Civil Code (BGB)).

Thus, the national uniqueness of the company names of traders not listed in the Commercial Register, and of freelancers depends on how common the owners' names are, and on the additional terms used. In regard to these forms of enterprise, with their often very limited regional bias and variety of product, the protection of company names by the German Trademark Act is to be estimated as rather low if there is no listing in the trademark register. In view of the large number of these firms, it is therefore to be assumed that the firm names will frequently not be unique. However, as these kinds of enterprises are very small by definition, it is mostly unlikely that they will have several establishments, so in these cases the address can be taken from the BA's establishment file to identify the company.

4.1.2 Firms with *Kaufmannseigenschaft*

In regard to the naming of companies, the rules for those with *Kaufmannseigenschaft* are not as strict. This includes every business that requires "a business concern set up for trading due to its size or type" (Section 1 HGB). *GmbH*s (Section 13 (3) GmbHG), *AG*s (Section 3 (1) AktG), *KG*s (Section 161 (1) HGB), and *OHG*s (Section 105 (1) HGB) count as commercial companies and are therefore always regarded as merchants in terms of the German Commercial Code (Section 6 (1) HGB). The same applies to companies representing special/mixed types of these legal forms, as for example *GmbH & Co. KG*s, *AG & Co. KG*s, and *Ltd. & Co. KG*s, which are special forms of the *KG*, *KGaA*s, and *AG & Co. KGaA*s[12], mixtures of the *KG* and *AG* (Section 278 *AktG*), and *UG*s, a special form of the *GmbH* (Section 5 *GmbHG*). Pursuant to Section 17 (2) GenG, *eG*s also always receive the *Kaufmannseigenschaft*. Natural persons, on the other hand, only obtain this characteristic if there is a requirement for commercial business operations, or if they opt to be listed in the Commercial Register voluntarily pursuant to Section 2 HGB. In this case they have to add "eingetragener

[12] The *AG & Co. KGaA* is a special form of the *KGaA*.

Kaufmann [registered merchant]" or an abbreviation of this phrase, as for example "e.K.", to their company name (Section 19 (1) HGB).

As these types of companies have to be registered in the Commercial Register or Register of Cooperatives, where the legal circumstances are apparent, such as the owners, they have much greater freedom of choice in regard to their company name. This consists of several components (IHK Köln 2011):

- a designation which can contain a name, a factual term, or a made-up term, and can consist of one or more words,
- the designation indicating the legal form, and
- other additional words where required.

Examples: "Max Meyer GmbH", "Max Meyer IT-Service GmbH", "Max Meyer GmbH IT-Service", "MMax Meyer GmbH", "KTV IT-Service GmbH", "Meyer IT-Service Berlin GmbH".

However, to guarantee that the company is identifiable, and that the contracting partner is assigned unambiguously, there are also limitations here concerning the choice of name. According to Section 18 (1) HGB, the firm name must be distinctive. Thus, in regard to the firm name, anything distinctive would be permitted. It does not have to make sense (a "*YXCVBN GmbH*" would be permitted, for example), and nor does it necessarily have to include the name of the owner, the location, or the activity. However, general names or descriptions of activities are not sufficient to achieve distinctiveness: a "*Handwerker* [Craftsman] *GmbH*", for instance, would not reveal any features distinguishing it from the businesses of other craftsmen. A company name which only consists of a description of an activity and a region ("*IT-Service Berlin GmbH*", for example) would not be sufficiently distinctive either. In cases like these it would be possible to achieve the necessary level of distinction by adding one more word (such as "*Meyer IT-Service Berlin GmbH",* or "*KTV Handwerker GmbH"*) (IHK Köln 2011).

However, a firm name may also not be admissible even if it is basically distinctive, as a new firm name has to be clearly distinct from all firm names in the same municipality which already exist and which are listed in the Commercial Register or Register of Cooperatives (Section 30 (1) HGB). In contrast to companies not listed in the Commercial Register, this also applies if the firm name next to the designation indicating the legal form only consists of the first and last names of the owner (Section 30 (2) HGB). In a case like this, an additional term must be added to the person's name to make it distinctive. Simply adding a different designation to indicate the legal form would not however suffice (IHK Köln 2011). However, the influence of Section 30 (1) HGB on the national uniqueness of a firm name is likely to be rather small due its limitation to the municipality.

The German Trademark Act explained in Section 4.1.1. also applies to these forms of enterprise of course, so there are several regulations and incentives for companies with *Kaufmannseigenschaft* which work toward a name which is unique throughout the country. These include the distinctiveness of a name, its uniqueness within the municipality, its protection through the German Trademark Act, and the fundamental interest in a name with recognition value. Nevertheless, the same firm name can be assigned several times, so the greater the sphere of the company's influence is, and the larger the company itself is, the more likely it is to be nonambiguous. The larger the company, the greater the degree of familiarity and variety of products will tend to be, and therefore the protection of the name and the likelihood of a brand being listed in the trademark register will also tend to increase. As the ReLOC database contains multinational companies, which tend to be larger, there is likely to be a relatively high degree of protection in regard to company names.

4.2 Checking the databases to be linked

4.2.1 Uniqueness of firm names

In the previous section the legal principles for naming companies were clarified. To gain insight into the actual national uniqueness of the company names in the database to be linked, a random sample of 250 companies with *Kaufmannseigenschaft* was pulled from the ReLOC database, and the frequency of these names was tested using the German Commercial Register and Register of Cooperatives.[13] As the linkage of the database should be effected via all establishment numbers used since the reporting process was introduced (January 1, 1973), all names that have ever been recorded in the Commercial Register or Register of Cooperatives were included.

Altogether, 18 company names were identified which appeared at least twice in the Commercial Register or Register of Cooperatives, which means that 7.2 percent of the companies did not have a name that was unique. Here, in 11 cases the name next to the legal form only consisted of one word, and in 10 cases this word consisted of a maximum of six letters. In the remaining seven cases, the name of the company always included a person's last name, and in three of these cases, it also contained the first name in full.

As is to be expected, the uniqueness of the company name increases with its length and complexity. If it consists of more than one word, the designation of the legal form aside, and does not include a person's name, it is unique throughout the country. The risk is greatest when it only consists of one short word aside from the designation of the legal form. If part of a person's name is used, the danger of the company name being assigned several times also increases. This is also influenced by how common the person's name is, and by other possible words added, of course.

If there is only interest in the corporate group level and not in the firm level, the risk of a firm name occurring several times is reduced, as research on the internet and in annual statements showed that in seven of the 18 cases the firms with identical names belonged to the same corporate group.

4.2.2 Mistakes in the name of the establishment/firm

As the ReLOC database was edited before the survey began, the correctness of the establishment names was of foremost importance. Accordingly, in a random sample taken from the ReLOC database, consisting of 100 firms with *Kaufmannseigenschaft*, the names were completely identical to their counterparts in the Commercial Register and Register of Cooperatives in 97 percent of the cases.

A sample of 100 establishments was likewise taken from the BA's establishment-level data. The population was restricted to establishments which had at least one employee who was liable for social insurance contributions or was in marginal part-time employment on June 30, 2009, and which can be classified as belonging to a company with *Kaufmannseigenschaft* in accordance with the designation indicating their legal form. The establishment names were researched in the Commercial Register and Register of Cooperatives, and completely identical company names were found in 85 cases. Here, what is known as preprocessing, the standardizing and editing of the establishment names as applied later in the process of data linkage (see Section 5), had already been taken into account. It also appears that preprocessing, where the name is cut off after the designation of the legal form, is very important for data linkage, as 22 (25.9%) of the 85 establishment names contained additional words after

[13] With the exception of the Register of *Partnerschaftsgesellschaften*, there is no central register for companies without *Kaufmannseigenschaft*, so these were excluded from this test from the start.

the designation of the legal form, and in 19 cases (22.4%) these words were not a component of the actual firm name.

Strategies for data linkage can also be derived from an analysis of the mistakes made in recording names. Therefore, in the case of each of the 15 establishments/firms recorded incorrectly in the BA establishment-level data, in the Commercial Register and Register of Cooperatives the firm name was sought to which they were most likely to be assigned. Thus, most frequently one or more words were left out, added, or put in the wrong place (see Table 3). It was considerably less common for words to be changed, either due to spelling mistakes, or the intentional shortening of a word, or the incorrect designation of the legal form.

Table 3: Mistakes in the BA's establishment/firm names

Type of mistake	Frequency
Word(s) left out	7
Order of words mixed up	4
Spelling mistakes	3
Word(s) added	2
Word(s) shortened	2
Incorrect legal form	1

Source: Establishment-level data from the BA statistics department (Nuremberg 05/2012): Random sample of 100 establishment numbers with at least one employee liable for social security contributions or in marginal part-time employment on June 30, 2009; multiple indications are possible.

The main problem is therefore that there are either too many or too few words in incorrect establishment/firm names, so to calculate the similiarity between the names it would therefore make sense to choose a measure of distance which does not penalize missing or superfluous name components too strongly, and where their precise position only plays a small role. One good alternative here would be to use *n*-grams, where the name is divided up into character groups of *n* length. The number of correlating character groups in the names compared decides their similarity.

Example of bigrams (n=2) using padding[14]: Similarity between "AHAB Hausbau" and "AHAB Bau" when blank spaces are deleted and all letters are capitalized.

"AHABHAUSBAU" is turned into the 12 character groups "_A", "AH", "HA", "AB", "BH", "HA", "AU", "US", "SB", "BA", "AU", and "U_". "AHABBAU" is turned into the eight character groups "_A", "AH", "HA", "AB", "BB", "BA", "AU", and "U_". This results in seven correlating character groups ("_A", "AH", "HA", "AB", "BA", "AU", and "U_"), and a Dice coefficient of 0.7 as the measure of similarity.[15]

The advantage of *n*-grams is that they do not only look for the components of a name in the environment of a particular place when comparing two names. In the example given above, this can be seen clearly in the bigrams "BA", "AU", and "U_". These occur in both names, but because the word "HAUS" is missing from "AHABBAU", these bigrams occur in different positions in the two examples. Nevertheless, they increase the Dice coefficient. Although the similarity is also reduced because the missing word affects the adjacent character groups (here, "BB" instead of "BH" or "SB"), no added weighting is given to the exact position of the *n*-grams in regard to the counting of the correlating character groups. The greater the value

14 In the case of padding, the beginning and ending of the expression are considered by adding a blank space to the first and last characters.

15 Dice coefficient: number of correlations*2/(number of base pairs in the first sequence + number of base pairs in the second sequence) = 7*2/(12+8) = 0.7.

of *n* that is chosen, the more strongly the fact that a word is missing will reduce the Dice coefficient, because when *n* is greater, the missing characters appear more often in the character groups created, while the number of character groups created decreases at the same time. Unigrams (*n*=1) are not suitable here, as their implementation would result in the differentation between correct and incorrect names being relatively low. If one were to use unigrams, even two names consisting of characters which were the same but all in different positions would result in a Dice coefficient of one. For this reason, bigrams are used later to calculate the similarity of establishment/firm names.

5 Data linkage

The previous paragraphs clearly show that it is possible to link a firm-level database with the BA's establishment-level data solely based on the firm names. As the firm name had not been correctly recorded in full for some establishments, and there was the further problem of firm names not being unique throughout Germany, the linkage was performed in two steps:

In the first step, both the names and addresses of the ReLOC firms were compared with those in the establishment-level data. Accordingly, in this step the search was for establishments at the location of the company headquarters.

In the second step, all the establishments whose names correlated to those of firms from the ReLOC database were identified. To do this, the addresses of the firms and establishments were not used. At the same time, the names of firms for which an establishment had been found were tested in regard to their national uniqueness by consulting the Commercial Register and Register of Cooperatives. If this was not the case, the establishment was subsequently excluded.

The record linkage took place between August and December of 2012. The BA's establishment-level data contains all the establishment numbers which had ever been used for the registration process until May 31, 2012, which is around 11.8 million. The names and addresses from the establishment file were updated in keeping with their status on December 31, 2009. In the case of establishment numbers which were not assigned until after December 31, 2009, the first status of the names and addresses recorded was used.

The ReLOC database consists of 3128 companies with *Kaufmannseigenschaft*, and 278 companies without *Kaufmannseigenschaft*. This differentiation is important for the following linkage steps, as the two groups were treated differently in part because of the different regulations concerning the naming of a company (see Section 4.1). Therefore, the second linkage step, which was made solely on the basis of the names of the companies and establishments, was only carried out for companies with *Kaufmannseigenschaft*. The establishments of companies without *Kaufmannseigenschaft*, on the other hand, were always identified by including their names as well as their addresses.

Linking the establishment and company databases only makes sense if preprocessing is implemented. Preprocessing was applied using a Stata do-file. In this case I accessed a Stata do-file created by Tanja Hethey-Maier and Anja Gruhl (Biewen et al. 2012) with additional modifications by Matthias Dorner, and integrated my own changes.

Preprocessing the establishment[16] and firm names involved the following consecutive steps which built on those preceding them:[17]

1. Quotation marks, brackets, and hyphens were replaced by blank spaces.

2. Commonly known abbreviations of or within designations of legal forms were identified and written out correctly in full:

   Examples: "GmbH" was written as "Gesellschaft mit beschränkter Haftung", "KG" as "Kommanditgesellschaft", "AG" as "Aktiengesellschaft", "OHG" and "oHG" as "offene Handelsgesellschaft", "KGaA" as "Kommanditgesellschaft auf Aktien", "StG" as "stille Gesellschaft", "eG", "e.G.", and "e. G." as "eingetragene Genossenschaft", etc.

3. Standardizing *"und"*:

   "u.", "und", "u", "and", and "+" were *replaced by "&".*

4. The following punctuation marks were replaced by blank spaces:

   Periods, exclamation marks, commas, colons, forward slashes, back slashes, and "greater than, less than" signs ("<" and ">").

5. Capitalizing all letters

6. Standardizing umlauts:

   Examples: "Ä" changes to "AE", "Ü" to "UE", "ß" to "SS", etc.

7. Deleting accents:

   Examples: "Á" to "A", "À" to "A", "Â" to "A", etc.

8. Designations of legal forms which are only partly written out in full were written in their full and correct form:

   Examples: "GESELLSCHAFT MIT BESCHR HAFTUNG" was changed to "GESELLSCHAFT MIT BESCHRAENKTER HAFTUNG", "KOMMANDITGESELL" to "KOMMANDITGESELLSCHAFT", "AKTIENGES" to "AKTIEN-GESELLSCHAFT", "EINGETR KAUFM" to "EINGETRAGENER KAUFMANN", and "OFFENE HANDELSG" to "OFFENE HANDELSGESELLSCHAFT", etc.

9. The full designations of legal forms were identified, and saved in a new variable in abbreviated form.

10. All characters appearing after the (identified) designation of the legal form in the establishment/firm name were removed. The resulting name was saved in a new variable.

---

16 The establishment name is comprised of the three consecutive fields available for recording the name (see Section 3.2): Field 1 + blank space + Field 2 + blank space + Field 3.

17 Between the steps, blank spaces at the beginning or end of the establishment/company name were deleted, and multiple consecutive blank spaces replaced by a single one where necessary.

11. All blank spaces were deleted.

In these steps, the following variables were generated for the firm/establishment names from the ReLOC database, and for the BA's establishment-level data (see Table 4):

- *Name_without_legal_form*: edited establishment/firm name up to the designation of the legal form, which is excluded from this variable however.
- *Legal_form*: standardized designation of the legal form in abbreviated form.
- *Name_with_legal_form*: edited establishment/firm name up to the designation of the legal form, which is also contained in this variable.

Table 4: Fictional examples of preprocessing

Establishment/firm name before preprocessing	Variables after preprocessing		
	Name_without_legal_form	***Legal_form***	***Name_with_legal_form***
AKB Fräs-Technik GmbH Niederlassung Berlin	AKBFRAESTECHNIK	GmbH	AKBFRAESTECHNIKGmbH
AKB Fräs-Technik Gesell. mbH	AKBFRAESTECHNIK	GmbH	AKBFRAESTECHNIKGmbH
René Muster IT Service	RENEMUSTERITSERVICE		RENEMUSTERITSERVICE
Rene Muster IT-Service eingetr. Kaufmann	RENEMUSTERITSERVICE	e.K.	RENEMUSTERITSERVICEe.K.
Rene Muster IT-Service e.K. Berlin	RENEMUSTERITSERVICE	e.K.	RENEMUSTERITSERVICEe.K.
"Insigno" Immobilienges. mbh	INSIGNOIMMOBILIEN	GmbH	INSIGNOIMMOBILIENGmbH
Klaus+Peter Meyer KG Technischer Großhandel	KLAUS&PETERMEYER	KG	KLAUS&PETERMEYERKG
Klaus & Peter Meyer Kommanditges. Zentrale Berlin	KLAUS&PETERMEYER	KG	KLAUS&PETERMEYERKG
Klaus u. Peter Meyer KG	KLAUS&PETERMEYER	KG	KLAUS&PETERMEYERKG

Different ways of writing the following legal forms were taken into account: *GmbH, AG, KG, OHG, e.K., eG, GmbH & Co. KG, KGaA, UG (haftungsbeschränkt), Ltd., GbR, e.V., StG, SE, AG & Co. KG, AG & Co. KGaA, AG & Co. OHG, GmbH & Co. OHG, UG (haftungsbeschränkt) & Co. KG, Ltd. & Co. KG,* and *SE & Co. KG.*

Based on the abovementioned background in regard to the establishment and firm names, variables were formed for them where all characters coming after the designation of the legal form were removed (Step 10), because otherwise the additional words or phrases coming after the firm name, some of which were to be found in the establishment-level data (see Section 3.2), would have distorted the similarites calculated, and reduced the number of correct linkages. If the name of a firm ends with the designation of the legal form – this is the case for 96 percent of the firms with *Kaufmannseigenschaft* in the ReLOC database –, the firm name is fully identified in the BA's establishment names, as long as it had been recorded correctly.

The identification of designations of the legal form which were completely abbreviated (Step 2), or only partly written out in full (Step 8) was done separately. The information content of periods and the use of upper or lower case letters could therefore be used for the abbreviations, while in the case of the legal forms that were only partly written out in full, the danger of their designation not being recognized was reduced by the preprocessing steps made previously.

Street and city names were also standardized. Here also, all punctuation, accents, and blank spaces were deleted, letters capitalized, and umlauts standardized. In regard to street names, the numbers were also separated from the streets. This gave rise to the *Street* variable, which contained the edited street without the number, and the *City* variable, with the edited information in regard to the city. Here I used Tanja Hethey-Maier's Stata do-file without own changes (Biewen et al. 2012).

5.1 Use of names and addresses

The data was first linked on the basis of names and addresses, where a gradual loosening of the name and address criteria was permitted. Loosening the criteria in regard to the address makes sense due to possible mistakes in the address information for one thing, and for another, there may also be other establishments belonging to the firm located in the vicinity of its headquarter.

Without preprocessing, there were very few companies in the ReLOC database with the same name, the same street, and the same three-digit ZIP code as an establishment from the establishment file (see Table 5). Establishments with identical features were only found in the cases of 460 (14.7%) companies with *Kaufmannseigenschaft*, and eight (2.9%) without it. Thus, it was urgently necessary to preprocess the establishment and firm names, and the address data. This increased the number of companies with *Kaufmannseigenschaft* for which at least one establishment number was identified to 1904, which corresponds to a share of 60.9 percent. For firms without *Kaufmannseigenschaft*, the number with at least one linkage remained low. This can be explained by the fact that these companies were very small by definition. There were fewer of them with an employee liable for social security contributions or in marginal part-time employment, and therefore fewer with an establishment number.

Table 5: Result of the exact comparison of names and addresses

Model	Matching criteria	Number of valid linkages			
		with *Kaufmannseigenschaft*		without *Kaufmannseigenschaft*	
		establishment no.	firms	establishment no.	firms
1	**Without preprocessing** Name, street, and 3-digit ZIP identical	489	460	8	8
2	**With preprocessing** *Name_with_legal_form*, *Street,* and 3-digit ZIP identical	2305	1904	17	16

Source: ReLOC database and establishment-level data of the BA statistics department (Nuremberg 05/2012).

As a large percentage of firms remained for which no valid linkage with the establishment-level data was found in Models 1 and 2, in the next models the exact comparison was replaced by an error-tolerant similarity function (see Table 6). The similarities in regard to establishment/firm names and street names were calculated using bigrams, whereby the legal form was not included in the calculation. The index consulted to evaluate the similarity was a result of the aggregated Dice coefficients of the establishment/firm names and street names. As the Dice coefficient for the variables to be compared showed a maximum of one, if there was a complete correlation between a firm and an establishment in regard to their name and streetname, this gave a value of two. The calculations were done using the Record Linkage Software Merge Toolbox (MTB) V0.73, which was developed at the University of Duisburg

(Schnell et al. 2005).[18] Blocking was also implemented here. This meant that the only observations from the two databases that were compared with each other were those that also showed the same value in the corresponding blocking variable that had been determined.

In Model 3, the companies are only compared to establishments with either the same three-digit ZIP code or the same city. All pairs with a similarity index of at least 1.4 were checked manually, and classified as a match or non-match. Although the legal form was excluded from the calculation of the similarity index, it was included in the evaluation of whether or not it was a correct match. In this way 2908 establishment numbers were assigned to 2422 companies with *Kaufmannseigenschaft*, and 174 establishment numbers to 161 companies without *Kaufmannseigenschaft*. This increased the linkage quota of the companies to 77.4 percent and 57.9 percent, respectively. The increase in the quota was particularly marked for companies without *Kaufmannseigenschaft*, as the names in the two databases differed more strongly with these types of companies, particularly in regard to the recording of industry, activity, establishment or other advertising identifiers. There is no obligation for these to be included in the company name, and they are therefore not always given or used consistently. Accordingly, in Model 3 the average similarity index for companies without *Kaufmannseigenschaft* was 1.71, while it was 1.95 for companies with *Kaufmannseigenschaft*.[19]

The street names were not taken into account in Models 4 and 5. Thus, the similarity index was consistent with the Dice coefficient for the establishment/firm names and was therefore able to reach a maximum value of one. As the number of possible linkages increased as a result, Model 4 only includes linkages with a similarity value of at least 0.7 in the classification as a match or non-match. In Model 5, the ZIP code districts were extended, and therefore the number of linkages to be tested increased, so the threshold value was set at 0.8. A further difference in regard to the previous models was that Models 4 and 5 were only used for firms with *Kaufmannseigenschaft*, as the fact that the street was left out of the similarity criteria meant that there was an increased probability that firms whose names were not unique in Germany would be linked incorrectly. As this particularly affects companies without *Kaufmannseigenschaft*, and there is no central register for this types of enterprises which makes it possible to test whether a name is used multiple times, these were excluded from Models 4 and 5.

In Model 4, the number of valid linkages increased by 175 to 3083, and in Model 5, this number was increased by a further 813 to 3896. The number of firms for which at least one establishment number was identified was increased by 94 to 2516 by Model 4, and by a further 47 to 2563 by Model 5. The stronger increase in valid linkages was therefore a result of Model 5, while the number of firms assigned increased more strongly in Model 4. This is explained by the fact that the establishments of several firms of above-average size which were located in metropolitan areas were included in Model 5 due to the expansion of the ZIP code districts.

[18] The software can be downloaded free of charge at: http://record-linkage.de.

[19] Although the Dice coefficients for the street names are also included in the similarity index, this does not explain this large difference.

Table 6: Result of the error-tolerant matching of names and addresses

Model	Blocking variable	Variables	Similarity function	Number of valid linkages			
				with *Kaufmannseigenschaft*		without *Kaufmannseigenschaft*	
				establishment no.	firms	establishment no.	firms
3	3-digit ZIP **or** *City*	*Name_without_legal_form* and *Street*	Bigrams	2908	2422	174	161
4	3-digit ZIP **or** *City*	*Name_without_legal_form*	Bigrams	3083	2516		
5	2-digit ZIP **or** *City*	*Name_without_legal_form*	Bigrams	3896	2563		

Source: ReLOC database and establishment-level data of the BA statistics department (Nuremberg 05/2012).

Thus, altogether, at least one valid linkage was found for 81.9 percent of the companies with *Kaufmannseigenschaft*. As expected, the quota for companies without *Kaufmannseigenschaft* was considerably lower, at 57.9 percent.

5.2 Only using the names

In the second step, only the establishment and firm names were used to identify valid linkages. The address of a firm or establishment therefore had no influence at all. Thus, all the establishments from the BA's establishment-level data whose names correlated with those of firms from the ReLOC database were identified. This name matching was performed using the *Name_with_legal_form* variable. Only those linkages where an establishment and a firm showed an exact match in regard to this variable, hence a Dice coefficient of one, were recognized as valid.

A fictitious example: The establishments assigned to the firm "KTV IT-Service GmbH" ("KTVITSERVICEGmbH") are "KTV IT-Service GmbH Berlin" ("KTVITSERVICEGmbH") and "KTV IT Service Ges. mbH" ("KTVITSERVICEGmbH"), but not "KTG IT-Service GmbH" ("KTGITSERVICEGmbH") or "KTV IT-Service e.K." ("KTVITSERVICEe.K.").

Hence, the preprocessing explained in Section 5 is particularly important here, as no error tolerance exceeding this can be permitted. Although an error-tolerant procedure was tested by calculating the similarities between all establishment and firm names using bigrams, it crystallized that when the similarity of the names decreased, there was a much stronger increase in the number of false linkages than in the number of correct ones, and that it would be much too labor-intensive to test these linkages due to the size of the establishment file. Thus, in this stage of linkage, errors in the recording of names which were not included in preprocessing led to valid linkages not being identified.

The linkages resulting from the exact name matching were tested for the uniqueness of the respective firm name behind it by means of the Commercial Register and Register of Cooperatives.[20] As the linkage was made using the entire set of establishment numbers, names which had already been deleted from these registers again counted as well. The uniqueness of the firm name was not tested for all firms however, but only if the (unedited) name before the designation of the legal form consisted of only one word, or contained a person's last name. This procedure was derived from the insights described in Section 4.2.1, as from the sample presented, it was only those firm names showing these characteristics that were not unique. If a firm name was used multiple times throughout Germany, the linkage was marked as not valid, and was not used further. This served to avoid incorrect assignments of estab-

[20] These registers are also used by the Chambers of Commerce to find out whether the names desired by new companies are already in use.

lishments to firms. On the other hand, it is to be assumed that larger firms are not affected by this problem as often (see Section 4.1), which means that they and their establishments may be slightly over-represented in the final version of the database.

As in Models 4 and 5, for the reasons mentioned, the names were only matched for firms with *Kaufmannseigenschaft*. A total of 53779 establishment numbers were found for 2502 of the 3128 (80.0%) firms (see Table 7). However, 3017, or 5.6 percent of the establishment numbers assigned, and 185, or 7.4 percent of the firms with at least one linkage did not fulfill the criterion of name uniqueness. This left 50762 establishment numbers which were distributed over 2317 firms, and were also still in use.

Table 7: Name uniqueness of potential linkages

Firm name unique in Germany	Establishment no.	Firms
Yes	50762	2317
No	3017	185
Total	53779	2502

Source: ReLOC database and establishment-level data of the BA statistics department (Nuremberg 05/2012); name uniqueness was checked with the aid of the German Commercial Register and Register of Cooperatives (accessed from: www.handelsregister.de).

5.3 Result of the two linkage steps

Table 8 shows the final number of valid linkages. The firms from the ReLOC database were assigned a total of 51539 establishment numbers. Of these, 3293 were identified both in the procedure using the names and addresses of the establishments/firms and in the procedure which only used the names of the establishments/firms for matching purposes. In contrast, 777 establishment numbers were only linked in the first step, and 47469 only in the second. Therefore, the first step, which takes mistakes in the establishment/firm names into account due to the similarity function, led to a considerable gain. What could not be clarified with this procedure was the number of establishment numbers which were not found at all. To do this, it would have been necessary to have a key like that of the Business Register System (URS) of the Federal Statistical Office, which enables establishments to be fully assigned to companies correctly.

Table 8: Source of valid linkages

		Name only		
		not identified	identified	total
Name and address	not identified	not known	47469	47469
	identified	777	3293	4070
	total	777	50762	51539

Source: ReLOC database and establishment-level data of the BA statistics department (Nuremberg 05/2012).

The 51539 establishment numbers are distributed across 2903 companies. Therefore, at least one establishment number exists for 85.2 percent of the companies. Differences arise if a distinction is made again between companies with and without *Kaufmannseigenschaft*. Thus, at least one valid linkage was found for 2742 of the 3128, and thus 87.7 percent of the firms with *Kaufmannseigenschaft*. For the 278 firms without *Kaufmannseigenschaft*, at least one valid linkage was found for 161, and therefore 57.9 percent of them.

The ReLOC database was linked with the BA's complete base of establishment numbers. However, the number of a firm's active establishments can change over the course of time, as new establishments can be opened or old ones closed down. An establishment's registrations of employees can be used to evaluate its actual activity. Table 9 shows the number of establishment numbers per firm, limited to establishments with at least one employee liable for social security contributions or in marginal part-time employment on June 30, 2009. Companies that showed no employees are therefore not represented here. 69.04 percent of the companies without *Kaufmannseigenschaft* had an "active" establishment number. 12.55 percent possessed two "active" establishment numbers, and 5.63 percent even had more than nine. Thus, in this database the share of companies with only one establishment (single-site companies) is relatively low. In comparison, according to the German Federal Statistical Office, around 98 percent of all firms with at least one employee liable for social security contributions or with taxable turnover on December 31, 2002, were single-site companies (Nahm und Phillip 2005). Likewise, in the BA's establishment-level data, around 97 percent of the edited establishment names (*Name_with_legal_form*) limited to establishments with at least one employee liable for social security contributions or in marginal part-time employment on June 30, 2009 only appeared once. The share of single-site companies, on the other hand, is more like that of the KombiFiD project. Here, the percentage lay between 71.7 percent and 77.5 percent, depending on the year of observation and the version, and in this project the database likewise consisted of companies of above-average size (Biewen et al. 2012).

Table 9: Establishment numbers per firm

Establishment no. per firm	Firms			
	with *Kaufmannseigenschaft*		without *Kaufmannseigenschaft*	
	number	share	number	share
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%		
Total	2445	100.00%	99	100.00%

Source: ReLOC database and establishment-level data of the BA statistics department (Nuremberg 05/2012): establishments with at least one employee liable for social security contributions or in marginal part-time employment on June 30, 2009.

6 Summary and outlook

Linking the ReLOC database with the IAB's establishment-level data enables an in-depth analysis of the effects of German investments on the companies investing and their establishments, and also shows a new method of linking firm-level data with establishment-level data. This does not require the key of the German Federal Statistical Office and could therefore offer an expedient alternative for many projects. This procedure was very profitable for the linkage of the ReLOC database with its multinational companies, as, unsurprisingly, its share of single-site companies is very small. Using the establishment and firm names alone multiplies the number of establishment numbers assigned by about 13. As not all firm names are unique in Germany and are not always recorded correctly, it can also make sense to include the addresses in the data linkage in a separate step, even though this means accepting that company headquarters will be over-represented. However, as has been shown, due to different regulations concerning the naming of firms, companies with *Kaufmannseigenschaft* and companies without *Kaufmannseigenschaft* should be treated differently.

The procedure presented here also offers starting points for further development, such as the possible use of further variables to assign establishments to firms. This could mean that an error-tolerant assignment procedure which goes beyond simple preprocessing could be applied more strongly. A promising variable would be the number or percentage of employees who change from one establishment to another over the course of time. In some ways this is similar to the procedure developed by Hethey and Schmieder (2013) to identify establishments that are started up or closed down. Establishments with names that are similar but not identical, and a higher than average exchange of employees could then also be assembeled to a firm. In the case of firms with names that are not unique, this could also be used to connect establishments identified with the help of the firm address with other establishments which share not only the same name but also a significant number of employees. However, the best alternative would still be a change in the legislation to make the combination of administrative data beyond the institutional limits more viable, as it was suggested in the course of the KombiFiD project (Biewen et al. 2012). A permanent linkage of the URS with the establishment-level data of the IAB would enable future projects to account for the firm level with relatively little effort. However, as long as the legislation remains as it is, new approaches whereby the assignment is made on the basis of the establishment/firm name constitute a worthwhile improvement.

References

Bender, Stefan; Wagner, Joachim; Zwick, Markus (2007): KombiFiD – Kombinierte Firmendaten 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. As of January 2013, http://www.arbeitsagentur.de/web/wcm/idc/groups/public/documents/webdatei/mdaw/mdk5/~edisp/l6019022dstbai391187.pdf (last accessed on 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 - Journal of Applied Social Science Studies, 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 Research Report, 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 Research Report, 04/2013, Nürnberg.

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

Hethey-Maier, Tanja; Schmieder, Johannes (2013): Does the use of worker flows improve the analysis of establishment turnover? Evidence from German administrative data. Schmollers Jahrbuch – Journal of Applied Social Science Studies, 133(4): 477–510.

IHK – Industrie- und Handelskammer Dresden (2013): Geschäftsbezeichnung und Namensangaben von nicht im Handelsregister eingetragenen Einzelunternehme(r)n. As of January 2013, http://www.dresden.ihk.de/servlet/link_file?link_id=27167&ref_knoten_id=48764&ref_detail=portal&ref_sprache=deu (last accessed on July 24, 2013).

IHK – Industrie- und Handelskammer Köln (2011): Wie finde ich die richtige Firmierung für mein Unternehmen? As of January 2011, http://www.ihk-koeln.de/upload/Voraussetzung_Firmierung_8901.pdf (last accessed on July 22, 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 (updated August 13, 2010), Nürnberg.

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

Imprint

FDZ-Methodenreport 5/2014 (EN)

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Agency
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Stefan Bender, Dagmar Theune

Technical production
Dagmar Theune

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

Internet
http://fdz.iab.de/

Corresponding author:
Johannes Schäffler
Institute for Employment Research (IAB)
Regensburger Str. 104
D-90478 Nürnberg
Phone: +49-911-179 2126
E-Mail: Johannes.Schaeffler@iab.de